On the use of meta-analysis techniques for multi-lab experiments

by

Ramlah Albayyat

M.S., Minnesota State University, Minnesota, 2016

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2023

Abstract

There has been a growing concern in different branches of science regarding the replicability of scientific findings. Recent research has shown that results from a large fraction of experimental studies are not able to be replicated by follow-up studies—a phenomenon known as the *replicability crisis*. In response, multi-lab studies have grown substantially in popularity, both to confirm that an original scientific claim can withstand rigorous followup testing and to test the validity of published, controversial scientific findings. As a recent example of the latter kind, Many Labs 4 (Klein et al., 2019) was a large-scale, pre-registered study designed to replicate an experiment by Greenberg et al. (1994) testing a claim regarding *terror management theory*—a theory in Psychology that a person's awareness of their own mortality will alter their behavior in a myriad of ways. In this replication study, the original experiment was repeated across 21 labs, with some labs even involving the original authors from Greenberg et al. (1994) in their experiment. Despite strict protocols being followed, the results from the original study were not able to be replicated. Chatard et al. (2020) rebutted against this finding, and claims that by imposing a strong restriction on the studies to be included in the analysis, and by performing a meta-analysis on this subset of studies, the original result is replicated.

This rebuttal leads to several key questions. First, are meta-analytic techniques, which estimate an overall effect size by aggregating effect sizes across studies, as-or-more effective than, linear mixed models, which use individual responses rather than effect sizes, when analyzing data from multi-lab studies? Second, since both mixed models and meta-analysis approaches rely on estimating the variability in effect sizes across labs, and since estimation of this parameter is typically unreliable, are sensitivity analysis approaches—which consider a range of potential values of this between-lab variance—preferable when analyzing multi-lab data? Finally, is the finding from Many Labs 4 accurate, or does the rebuttal by Chatard

et al. (2020) hold significant merit? We aim to find answers to all of these questions in this dissertation.

First, we perform an extensive literature review on random-effects meta-analysis methods, and perform an extensive simulation to evaluate the best practices for performing meta-analysis when using individual participant data (IPD) instead of aggregate data from multi-lab studies. We then compare the best performing meta-analysis estimators to those obtained from a mixed model. Overall, we consider a total of 5,760 combinations of estimators and multi-lab experimental settings for meta-analysis, and another 80 for mixed models. We find that the meta-analysis and linear mixed model approaches yield similar results, with meta-analysis methods performing slightly better. Additionally, we find that both methods for estimation suffer from the same significant pitfall—estimation of across-lab variability in treatment effects is often inconsistent and unreliable when the number of labs included in the study are small.

Second, to overcome issues with estimation of across−lab variability, we develop a sensitivity analysis approach for determining the significance of effect size estimates for both meta-analysis and mixed-model estimators. These methods allow for researchers to consider a range of different values for the across−lab variance, and helps them determine for what values the estimate of the effect size is statistically significant. While effective, we find that these approaches can be misleading when the claimed or estimated across−lab variance is much lower than the actual across−lab variance, which can be possible when sample sizes and/or the number of labs under study is small.

Finally, we apply our methods to re-analyze the data from Many Labs 4 (Klein et al., 2019). Our analysis corroborates the original finding of (Klein et al., 2019) that the original experiment is unable to be replicated with any kind of consistency. We also identify issues with the analysis in the rebuttal by Chatard et al. (2020) that would lead them to obtain inaccurate results.

On the use of meta-analysis techniques for multi-lab experiments

by

Ramlah Albayyat

M.S., Minnesota State University, Minnesota, 2016

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2023

Approved by:

Major Professor
Michael Higgins

Copyright

© Ramlah Albayyat 2023.

Abstract

There has been a growing concern in different branches of science regarding the replicability of scientific findings. Recent research has shown that results from a large fraction of experimental studies are not able to be replicated by follow-up studies—a phenomenon known as the *replicability crisis*. In response, multi-lab studies have grown substantially in popularity, both to confirm that an original scientific claim can withstand rigorous followup testing and to test the validity of published, controversial scientific findings. As a recent example of the latter kind, Many Labs 4 (Klein et al., 2019) was a large-scale, pre-registered study designed to replicate an experiment by Greenberg et al. (1994) testing a claim regarding *terror management theory*—a theory in Psychology that a person's awareness of their own mortality will alter their behavior in a myriad of ways. In this replication study, the original experiment was repeated across 21 labs, with some labs even involving the original authors from Greenberg et al. (1994) in their experiment. Despite strict protocols being followed, the results from the original study were not able to be replicated. Chatard et al. (2020) rebutted against this finding, and claims that by imposing a strong restriction on the studies to be included in the analysis, and by performing a meta-analysis on this subset of studies, the original result is replicated.

This rebuttal leads to several key questions. First, are meta-analytic techniques, which estimate an overall effect size by aggregating effect sizes across studies, as-or-more effective than, linear mixed models, which use individual responses rather than effect sizes, when analyzing data from multi-lab studies? Second, since both mixed models and meta-analysis approaches rely on estimating the variability in effect sizes across labs, and since estimation of this parameter is typically unreliable, are sensitivity analysis approaches—which consider a range of potential values of this between-lab variance—preferable when analyzing multi-lab data? Finally, is the finding from Many Labs 4 accurate, or does the rebuttal by Chatard

et al. (2020) hold significant merit? We aim to find answers to all of these questions in this dissertation.

First, we perform an extensive literature review on random-effects meta-analysis methods, and perform an extensive simulation to evaluate the best practices for performing meta-analysis when using individual participant data (IPD) instead of aggregate data from multi-lab studies. We then compare the best performing meta-analysis estimators to those obtained from a mixed model. Overall, we consider a total of 5,760 combinations of estimators and multi-lab experimental settings for meta-analysis, and another 80 for mixed models. We find that the meta-analysis and linear mixed model approaches yield similar results, with meta-analysis methods performing slightly better. Additionally, we find that both methods for estimation suffer from the same significant pitfall—estimation of across-lab variability in treatment effects is often inconsistent and unreliable when the number of labs included in the study are small.

Second, to overcome issues with estimation of across−lab variability, we develop a sensitivity analysis approach for determining the significance of effect size estimates for both meta-analysis and mixed-model estimators. These methods allow for researchers to consider a range of different values for the across−lab variance, and helps them determine for what values the estimate of the effect size is statistically significant. While effective, we find that these approaches can be misleading when the claimed or estimated across−lab variance is much lower than the actual across−lab variance, which can be possible when sample sizes and/or the number of labs under study is small.

Finally, we apply our methods to re-analyze the data from Many Labs 4 (Klein et al., 2019). Our analysis corroborates the original finding of (Klein et al., 2019) that the original experiment is unable to be replicated with any kind of consistency. We also identify issues with the analysis in the rebuttal by Chatard et al. (2020) that would lead them to obtain inaccurate results.

Table of Contents

List of Diagrams

List of Figures

List of Tables

Acknowledgments

First, I would like to thank the Almighty Allah, “All praise is to Allah, the Lord of the Creation,” for the blessings of patience and motivation while completing this dissertation. Without His grace, I would not have been able to accomplish this feat.

I also give profound thanks to my advisor, Dr. Michael Higgins, for his support, motivation, and guidance throughout this research. Without his advice and help, I could not have accomplished this dissertation. My sincere appreciation also goes to Dr. Christopher Vahl, Dr. James Higgins, and Dr. Pietro Poggi-Corradini forserving on my dissertation committee.

Most importantly, my heartfelt gratitude goes to my mother, Fawzia Alhalal, and my father, Hameed Albayyat, for encouraging me to follow my dreams. I could not have achieved this significant milestone without them, for they believed in me and taught me the value of education. It is hard quantify and express the gratitude I have for them.

Next, I would like to thank my family. To my beloved husband, Ali Alkurtass, who offered never-ending patience, support, and confidence in me. Your sacrifices and encouragement were key as I traveled this journey. Additional thanks to my daughter, Zainab, and my son, Hassan, for their unconditional love. I cannot put into words how grateful I am. Gratitude also goes to my siblings Fatimah, Hussain, Mawahib, Hisham, Rameen, and Mansour, whose support and encouragement were essential to making this work possible.

I would be remiss if I did not thank the government of my country, Saudi Arabia, and the Saudi ministry of education for their financial support throughout my studies. I could not have pursued my doctoral degree without their help.

I also want to thank the faculty and staff of the Kansas State University Statistics Department for their support and kindness.

Dedication

This doctoral dissertation is dedicated to my parents, Fawzia Alhalal and Hameed Albayyat, who encouraged me to pursue higher education and taught me to believe in myself.

I also dedicate this dissertation to those who support me more than anyone else: my husband, Ali Alkurtass, and my children Zainab and Hassan.

Chapter 1

Introduction

1.1 Introduction

There has been a growing concern in different branches of science regarding the replicability of scientific findings. Recent research has shown that results from a large fraction of experimental studies cannot be replicated by follow-up studies—a phenomenon known as the *replicability crisis*. For example, the journal Nature conducted a survey of 1,500 scientists and found that 70% of those polled had been unable to replicate others' results, and 50% could not replicate even their own results (Baker, 2016)

One contributing cause to the replication crisis is that each experiment is performed within its own unique research environment. Such differences in environments may lead to differing treatment effects within each experiment—otherwise known as treatment-by environment interaction or heterogeneous treatment effects. This may make it difficult for another researcher to replicate an experimental finding, even if the follow-up study design and protocol is as identical as possible to the original study. Recent work by Higgins et al. (2021) has formalized this framework and derived adjusted p-values and confidence intervals for inferences on the difference in means that consider this potential environment-by treatment interaction

From a design perspective, multi-lab or multi-facility studies—in which the same experimental study is performed simultaneously across several locations with each location closely following the same pre-specified protocol—provide an additional pathway for identifying and mitigating issues with experiments that may be susceptible to large treatment-by-environment interaction effects. These types of experiments have become increasingly popular, both for verifying a scientific result is valid prior to publication (Jaljuli et al., 2021; Kafkafi et al., 2005) and for confirming or questioning the validity of a scientific claim. For an example of the latter, the Many Labs series of experiments (Buttrick et al., 2020; Ebersole et al., 2016, 2020; Klein et al., 2014a,b, 2018, 2022) perform multi-facility experiments to test the validity of commonly held theories in the psychological sciences.

Despite the emergence of multi-lab experiments, there is no commonly accepted way to analyze data from a multi-lab experiment. Often, it is assumed that the treatment effect—or effect size—within each lab is a random variable. When individual-participant data is available, linear mixed models seem quite appropriate data analysis. However, random effect meta-analytic techniques—in which overall effect size estimates are obtained using summaries from multiple studies, rather than the individual data—are also widely used for these data.

However, when the number of labs and/or the number of units within each lab is small, estimates from both mixed model and meta-analysis approaches can be quite variable. In particular, both mixed model and meta-analytic approaches require accurate estimates of across-lab variability in order to be effective. However, current best practices for estimating this across-lab variability—for example, maximum likelihood (ML) by Hardy and Thompson (1996), restricted maximum likelihood (REML) by Viechtbauer (2005) , and the Paule-Mandel estimator (PM) by Paule and Mandel (1982)—are often imprecise, especially when the number of studies is small.

In this dissertation, we focus on determining the best approach to estimate treatment effects from multilab experiments. First, we review the literature on random effects meta-analysis methods, and we perform an extensive simulation to evaluate the best practices for performing meta-analysis.We then compare the best performing meta-analysis estimators to those obtained from a mixed model. Next, we develop a sensitivity-analysis approach for determining the significance of effect size estimates for both meta-analysis and mixed-model estimators. Instead of estimating the across-experiment variability directly, we consider a range of different values of the environmental effect ratio (EER)—a ratio of the standard deviations of the across-experiment variability and the within-experiment variability—and determine under which values of the EER lead to statistically significant estimates of the treatment effect. Finally, we apply what we learn from the first two projects to re-evaluate the Many Labs 4 experiment (Chatard et al., 2020; Klein et al., 2022), a well-known multi-lab replication study in which a controversial finding involving terror management theory was unable to be replicated.

1.2 History of Replication Crisis

1.2.1 The File-Drawer Problem and Scientific Misconduct

Since 1959, the researchers faced "the file drawer problem:" journals are filled with significant results and 5% of these studies had Type I errors. However, the file drawers are filled with 95% of the non-significant studies (Rosenthal, 1979; Sterling, 1959). Later, Bauchau (1997) discussed the so-called file drawer problem; for each published paper telling a significant result, a considerable number of similar studies stayed secreted in file drawers because they did not achieve a significant result. He began putting the file drawer problem in a statistical framework to discuss it and highlight other related problems. According to an article in the Indian Journal of Anaesthesia, Nagarajan et al. (2017) reported "It is speculated that every significant result in the published world has 19 non-significant counterparts in file drawers."

Because of the file drawer problem, called publication bias, the rate of fabrication, falsification, and plagiarism in study research has increased. Harvey (2020) said that although the reason for fraud is complicated, there is good evidence that the structure of higher education perpetuates research fraud because of the grants received for published research. Many more examples of well-known research fraud cases exist across all fields. A renowned social psychologist in the Netherlands, Diederik Stapel, admitted that he faked data and made-up entire experiments that were published. Marc Hauser, a once-popular professor at Harvard, and Karen Ruggiero, a former professor at the University of Texas, both allegedly fabricated data for published studies. More often, data is modified slightly to achieve the desired results rather than completely fraudulent. For example, plant biologist Oliver Voinnet did multiple cases of data manipulation. He was issued a 2-year suspension from the Centre National de la Recherche Scientifique in 2015. Another form of fraud is replicating experiments enough times until eventually, by chance, a significant result is achieved and is published even if the null hypothesis were correct. (Bauchau, 1997; Carpenter, 2012; Diener and Biswas-Diener, 2016; Harvey, 2020; Sterling, 1959). Additionally, in a survey of almost 6,000 American psychologists, 67% admitted to selecting studies that would work, 74% failed to report all dependent measures, 71% continued to collect data until getting a significant result, 54% reported unexpected findings as expected, 35% had doubts about the integrity of at least one part of their own research, and 1.7% admitted to having faked their data (Laws, 2013).

Ioannidis (2005) published an article "Why Most Published Research Findings Are False" that caused serious concerns about the integrity of research findings. It has become the turning point in the years since. This realization became widespread by early 2010, and it was called the replication crisis Resnick (2018). In 2016, a group of Belgian psychology researchers wanted to see if there was a file drawer problem, so they conducted a follow-up study of their own study that was performed between 2009 and 2014 on 453 subjects. They found that the results obtained were not like they expected. Therefore, they had decided to open their own file drawer and make a bold stand by publishing their unpublished studies anyway. They wanted to encourage other laboratories to do the same (Lane et al., 2016). To highlight the nature of this crisis, the journal Nature conducted a survey of 1,500 scientists and found that 70% of those polled had been unable to replicate others' results, and 50% could not replicate even their own results (Baker, 2016).

1.2.2 Scientists Moved Forward

Alarm about a replication crisis launched a wave of large-sample replication projects to replicate published findings. The largest project was the Open Science Collaboration, which reported in papers published in 2008 in three high-ranking psychology journals. Collaboration (2015) found that about a third to half of the 100 experiments were replicated. Likewise, Camerer et al. (2018) designed a study to replicate 21 social science experiments. These studies had been published in Nature and Science between 2010 and 2015.These attempts were conducted by four different countries: the United States, Sweden, Singapore, and Austria. Only 14 were successfully replicated. A registered replication reports project was conducted by Simons et al. (2014) that consisted of many laboratories with high-powered studies. Here, 36 of 97 findings failed. In addition to these projects, several Many Labs studies have been designed, and many researchers attempted to replicate either one specific study or a whole set of studies. In the Many Labs 1 project (Klein et al., 2014b), there were 36 research groups that attempted to replicate the 13 effects with a total of 6,344 participants; 10 of 13 effects were replicated successfully. In the Many Labs 2 project (Klein et al., 2018), a team of 186 researchers conducted replications of the 28 findings with more than 15,000 participants from 36 countries and territories. Fourteen of the 28 findings failed to replicate, although the sample size of more than 60 laboratories is quite large. The Many Labs 3 project was conducted by Ebersole et al. (2016). Ten known effects in 20 participant pools were replicated, only three successfully (Ebersole et al., 2020; McShane et al., 2019; Stroebe, 2019).

The inability to replicate the initial findings builds doubt that the result truly exists and presents a crisis for scientific progress. Therefore, researchers should carefully consider which factors could affect the replication procedures and avoid them (Diener and Biswas-Diener, 2016; Yaffe, 2019).

1.3 Causes of Replicability Crisis

The failures of replicability do not necessarily mean the initial findings are a false positive, the original study authors were frauds, the experimental design and research practices are poor, or the original report was incorrect (Ioannidis, 2015; Stroebe et al., 2012). We should carefully think about the barriers preventing replication before carrying out a replication study and do our best to avoid them, or there will be too many inconclusive replication experiments (Hedges and Schauer, 2019a; Maxwell et al., 2015). There is no consensus among researchers about the causes of replication failure. It could be due to lack of experience and expertise, substandard research practices, studies designed with small sample sizes, statistically significant results resulting from chance, that the original results of the studies are valid for only some people in some circumstances and not universal or lasting, different research environments, publication bias, p-hacking, inappropriate analyses given the data, variability of p-values, overstated evidence, and underpowered studies (Bello and Renter, 2018; Boos and Stefanski, 2011; Chen et al., 2021; Diener and Biswas-Diener, 2016; Francis, 2012; Gibson, 2021; Higgins et al., 2021; Ioannidis et al., 2009; Lewandowsky and Oberauer, 2020; of Sciences et al., 2019; Shiffrin et al., 2018)

Regarding psychology, some have argued that psychology uses a nebulous and unclear theory or prefers to work with under-specified conceptualizations of the interest variables (Feest, 2019; Hicks, 2023; Klein, 2014; McPhetres et al., 2021; Muthukrishna and Henrich, 2019; Stroebe and Strack, 2014). On the other hand, psychological research is afflicted with low statistical power, according to a 2018 analysis of 200 meta-analyses. Replication studies that have low power are liable to false negatives (Stanley et al., 2018).

1.3.1 Multi-labs Replications

Until now, there is no unanimity on how replication studies should be conducted and no significant understanding of how replication design should be. Some think the idea of replication is obvious and straightforward; simply repeat the study with the same or larger sample size than the initial experiment, then check whether the results of both studies agree or not. Some evaluate the replication study by only a single study, while others assess by multiple laboratories. There is no unified approach. Hence, some studies that claimed to have failed to replicate may not have failed at all, but there may be low statistical power in single replication studies. We know that a single study is not as expensive and time-consuming as multiple studies. And the need for multiple replication studies may disappoint some researchers who want immediate results. Yet, researchers should carefully think about the barriers preventing replication before carrying out a replication study and do their best to avoid them, or there will be too many inconclusive replication experiments (Gustafsson, 2017; Hedges and Schauer, 2019a; Maxwell et al., 2015).

Replication in multi-study designs has many advantages compared to traditional single-study designs. Individual studies may be subject to publication bias; studies run in single labs may not be able to accumulate the sample sizes needed to provide sufficient precision, and it may be difficult to differentiate whether significant results are widely applicable or confined to the lab in which the study was conducted. In contrast, preregistered multi-labs studies with large sample sizes provide more objective and reliably estimated effect sizes that do not suffer from publication bias or selective reporting and ensure high power. The power usually rises more quickly if the number of labs is increased (at least 50) instead of the participants per lab (Bello and Renter, 2018; Bonett, 2012; Hedges and Schauer, 2019b; Kvarven et al., 2020; Lewis et al., 2022).

1.3.2 Replication Issue in Multi-environments

The different environments can be a potent impediment to a successful replication, yet the researchers do not consider it when measuring applicability. The various experimental environments, including but not limited to time, location, weather, and laboratory environment (i.e., personnel, equipment, measurement techniques), can affect the responses either in the same way or in a unique way to each treatment in a follow-up study. These differences contribute to a unique research environment for each experiment, and it can be difficult for another researcher to repeat a study in the same way, even if both studies are conducted with high accuracy (Filazzola and Cahill Jr, 2021; Fraser et al., 2020; Higgins et al., 2021; Powers and Hampton, 2019). In other words, if researchers can replicate the original study in vastly different environments, it is all the more impressive (Loken and Gelman, 2017).

We consider the initial experiment with two treatments (a treatment group and a control group) as follows

$$Y_{ik} = \mu_i + \varepsilon_{ik} \qquad i = 1, 2 \ \ , \ k = 1, .., n_i \tag{1.1}$$

where μ_i is the mean value for the i^{th} treatment condition, and ε_{ik}'s are independent and identically distributed random variables (abbreviated as i.i.d.) as $N(0, {\sigma^2}_e)$. When the same experiment is performed in a different environment selected randomly, the study results will differ by more than the sampling error. In this manner, two sources of variation will be considered: within-study variability and variability arising from differences between studies. Thus, the following model for replication studies with two treatments: a treatment group and a control group will be

$$Y_{ik} = \mu_i + \theta + \zeta_i + \varepsilon_{ik} \qquad i = 1, 2 \ \ , \ k = 1, .., n_{ij} \tag{1.2}$$

where

$$Y_{ik} \sim N(\mu_i, \sigma_e^2 \ ) \ , \quad \mu_i \sim N(\mu, \sigma_\theta^2 + \sigma_\zeta^2)$$

and μ_i is actual mean value for the i^{th} treatment in the environment. The term θ represents a common random source of variability to all observations. The ζ_i's (interaction terms) represent random sources of variability unique to each treatment. The ε_{ik}'s are distributed as ind $N(0, \sigma_e^2)$, common variance. Similarly, we assume that ζ_i's are distributed as iid $N(0, {\sigma^2}_\zeta)$, and θ is distributed as $N(0, {\sigma^2}_\theta)$. We further assume that these random terms are mutually independent (Higgins et al., 2021).

Because there is environment-by-treatment interaction, which appears to be common in research (Berliner, 2002; Kafkafi et al., 2005, 2017, 2018), the interaction with laboratories should be considered (Higgins et al., 2021; Jaljuli et al., 2021). The measures of replicability should account for a changing environment when making statistical inferences and the standard deviations of environment-by-treatment interaction divided by the experimental error (Higgins et al., 2021), which is defined as

$$EER^2 = \frac{\sigma_\zeta^2}{\sigma_e^2} \tag{1.3}$$

1.4 Replication and Meta-Analysis

For almost three decades, there has existed a necessary and symbiotic relationship between meta-analysis and replication (Allen and Preiss, 1993; Sharpe and Poets, 2020). Meta-analysis provides a technique for reviewing existing research findings and determining to what degree the finding has been clearly replicated (Allen and Preiss, 1993; Eden, 2002). Sharpe and Poets (2020) considered meta-analysis as the organization of the scientific chaos and as the response to the replication crisis. Fletcher (2022) considered the meta-analytic measures as the best measures of the replication that align functionally with its target. From statisticians' opinion, the results of meta-analytic methods have more formidable statistical power than those obtained from singular studies, raising the probability of discovering the associations of interest (Jain et al., 2019; Konstantopoulos, 2006; Sartal et al., 2021). Meta-analysis has been offered as the solution for the replication crisis (Braver et al., 2014; Collaboration, 2015; Fabrigar and Wegener, 2016; McShane et al., 2016; Nelson et al., 2018; Sharpe and Poets, 2020).

In contrast, some authors criticized classifying it as the replication crisis savior. Surprisingly, some of the criticism is due to personal reasons, not scientific and logical reasons. For example, when the original scientists become discontented and frustrated because their results did not replicate, they tried to offer explanations or excuses for non-replication by blaming the meta-analysis method. Yet, the most surprising is that some criticisms have arisen not from its failings but from its extraordinary success (Tebala, 2015). Or the criticism can be due to a misunderstanding of the usage of the meta-analysis method. Field (2003) wrote that "despite the obvious faith psychologists and other scientists and practitioners have placed in meta-analysis, there is a growing body of evidence to suggest that it is often used incorrectly, and may not be the answer to our literature review prayers after all." However, meta-analysis' tremendous popularity has muted many critics to the extent. Hall and Hall and Rosenthal (1991) observed, "The weight of consensus is clearly on the side of acceptance: One rarely finds condemnations of the whole [meta-analysis] enterprise anymore". Considering this, Sharpe and Poets advised that anyone who wanted to do meta-analysis accurately must implement it thoughtfully, prudently, and publicly (Sharpe and Poets, 2020).

1.4.1 Meta-Analysis

Meta-analyses started to gain popularity mid-20th century and is currently widely used in many fields such as statistics, psychology, epidemiology, medicine, and social science. It is a statistical procedure that combines estimates of effect sizes from several independent related studies to estimate an overall effect size (Chung et al., 2013). In statisticians' opinion, the results of meta-analytic methods have more formidable statistical power than those obtained from singular studies, raising the probability of discovering the associations of interest (Jain et al., 2019; Konstantopoulos, 2006; Sartal et al., 2021).

There are two common statistical models that can be used to perform metaanalyses: the fixed-effect model and the random-effects model. It is essential to select the appropriate model to guarantee that the different statistics are estimated correctly. Typically, studies have differences in their design, conduct, and participants, making the results from individual studies vary noticeably, which points to treatment effect heterogeneity (Stanley et al., 2018; van Erp et al., 2017). In the presence of heterogeneity, a random-effects meta-analysis is considered where both within-study and between-study variability are estimated (DerSimonian and Laird, 1986; Higgins et al., 2009). Otherwise, seriously underestimated statistical errors can result in misleading conclusions (Langan et al., 2019; Sugasawa and Noma, 2021; Veroniki et al., 2016).

Meta-analysis has been applied to more complicated models and complicated statistical distributions to describe the issue at hand better (Mengersen and Schmid, 2013). In most cases, we assume with this model a normal distribution of the underlying effects across studies (Langan et al., 2019). Numerous methods have been put forward to measure out the amount of between-study variance that varies in popularity and complexity. Therefore, the choice of heterogeneity variance method is essential (as shown in Appendix A.1; (Veroniki et al., 2016)). The most commonly implemented approach is the DerSimonian and Laird (DL) method and is the default approach in many software routines (Langan et al., 2019; Veroniki et al., 2016). The R package `metafor` is useful for conducting meta-analyses (Viechtbauer and Viechtbauer, 2015).

1.4.2 Model of Replication Studies

There are two families models in meta-analysis: the fixed-effect model and the random effects model. We consider the random-effects model, where the true effects are assumed to be drawn at random from a population of effects. Thus, the actual effect sizes vary across studies for any observed effect δ_j. For a meta-analysis of m studies, a random-effects model is expressed as

$$\hat{\delta}_j \sim N(\delta_j, \sigma_j^2) \quad , \quad j = 1, ..., m \tag{1.4}$$

$$\delta_j \sim N(\Delta, \tau^2) \tag{1.5}$$

where $\hat{\delta}_j$ are the observed standardized mean difference (SMD) effect sizes and Δ is the overall effect size.

These sets of studies have identical heterogeneity τ^2, but unique sampling error σ_j^2. However, sometimes to simplify things or to derive measures of heterogeneity, we consider the special case in which the within-study variances are equal in all studies (Hedges and Vevea, 1998; Higgins and Thompson, 2002; Rubio-Aparicio et al., 2020; Siegel et al., 2021).

1.4.2.1 A Standardized Mean Difference (SMD) Effect Size

The standardized mean difference (SMD) is one of the most widely used (Sánchez-Meca and Marin-Martinez, 1998), and it is usually estimated by Cohen's d or Hedges' g estimators (Lin and Aloe, 2021).

$$\hat{\delta}_{C.j} = \frac{\bar{Y}_{1j.} - \bar{Y}_{2j.}}{S_{p.j}} \tag{1.6}$$

$\hat{\delta}_{C.j}$ is called Cohen's d, referring to (Cohen, 2013), and $S_{p.j}$ is the pooled standard deviation for the effect size in study j is defined by

$$S_{p.j} = \sqrt{\frac{(n_{1j} - 1)S_{1j}^2 + (n_{2j} - 1)S_{2j}^2}{n_{1j} + n_{2j} - 2}} \tag{1.7}$$

where n_{2j} and n_{2j} are group sample sizes, and S_{1j}^2 and S_{2j}^2 are the corresponding group variances. However, Hedges (1981) showed that $\hat{\delta}_{C.j}$ are biased in studies with small sample sizes , $E(\hat{\delta}_{C.j}) = \delta_j / J_j$. Therefore, the unbiased estimate should be given by

$$\hat{\delta}_{H.j} = J_j \left(\frac{\bar{Y}_{1j.} - \bar{Y}_{2j.}}{S_{p.j}} \right) \tag{1.8}$$

where J_j is given by

$$J_j = \frac{\Gamma(\upsilon_j/2)}{\sqrt{\upsilon_j/2}\ \Gamma[(\upsilon_j - 1)/2]} \tag{1.9}$$

where $\upsilon_j = n_{1j} + n_{2j} - 2$. The commonly used approximation of J is given by (Hedges, 1981; Hedges and Olkin, 2014), which is

$$J_{Hedges.j} \approx \left(1 - \frac{3}{4\upsilon_j - 1} \right) \tag{1.10}$$

The absolute values of the error is less than 0.0005 and the error does not exceed 1.5×10^{-5} if $\upsilon_j > 50$. Xue (2020) improved the accuracy of Hedges's approximation by using higher order roots of Wallis ratio $\Gamma(x+1)/\Gamma(x+\frac{1}{2})$. There are six Mortici's approximations for Wallis ratio (Dumitrescu, 2015; Mortici, 2010). Xue (2020) compared them by measuring the absolute values of their errors to the real value after putting $x = \frac{\upsilon}{2} - 1$ in all approximations. The more accurate and efficient approximation is given bellow

$$J_{xue.j} \approx \sqrt[12]{1 - \frac{9}{\upsilon} + \frac{69}{2\upsilon_j^2} - \frac{21}{\upsilon_j^3} + \frac{687}{8\upsilon_j^4} - \frac{441}{8\upsilon_j^5} + \frac{247}{16\upsilon_j^6}} \tag{1.11}$$

Table 1.1: *Accuracy of different approximations of J_j*

υ_j	$\lvert J_j - J_{Hedges.j}\rvert$	$\lvert J_j - J_{xue.j}\rvert$
10	0.000331315	8.474759 e-09
50	1.268082 e-05	2.120526 e-14
100	3.148026 e-06	3.164136 e-14
300	3.480854 e-07	1.173506e-13
500	1.25187 e-07	1.2923 e-13

Hedges' g can be an unbiased estimator within each study but not in the synthesized Hedges' g (Lin and Aloe, 2021). Thus, averaging effect size by Hedges'g estimator ($\hat{\Delta}_H$) does not guarantee unbiasedness in meta-analyses as researchers expect. More troublesome, Hedges'g might lead to a larger bias in meta-analysis results (Lin, 2018; Lin and Aloe, 2021). For example, if we consider a particular case where m studies have the same sample size n under the fixed effect model ($J_i = J$) (Lin, 2018) then

$$\hat{\Delta}_H = \frac{\sum_{j=1}^{m} \hat{\delta}_{H.j}/s_{H.j}^2}{\sum_{j=1}^{m} 1/s_{H.j}^2} = \frac{\sum_{j=1}^{m} J\ \hat{\delta}_{C.j}/J^2\ s_{C.j}^2}{\sum_{j=1}^{m} 1/J^2\ s_{C.j}^2} = J\hat{\Delta}_C \tag{1.12}$$

If we assume $\hat{\Delta}_C > 0$ ($\hat{\Delta}_C$ indicates Cohen d estimator), then we have $\hat{\Delta}_H < \hat{\Delta}_C$ since J is always less than 1. Now, let assume that the overall Cohen's d underestimates and the true overall SMD Δ is positive, then we have that $\hat{\Delta}_H < \hat{\Delta}_C < \Delta$. Thus, in this situation we see that Hedges' g is more biased then Cohen's d. On the other hand, Hedges' g can be less biased if we assume that $\hat{\Delta}_C$ overestimates, which leads to $\hat{\Delta}_C > \Delta$. Due to that, Lin (2018) and Hamman et al. (2018) suggested that researchers should stop saying that Hedges' g is less biased than Cohen's d.

1.4.3 Evaluation the Performance of Meta-Analysis

Meta-analysis often includes a small number of studies (Seide et al., 2019). Half or more of the meta-analyses performed in the literature contain only two or three studies (Hedges and Vevea, 1998; Henmi and Copas, 2010; Poole and Greenland, 1999). For example, from 3,263 selected meta-analyses, 31% were based on two studies, 38% were based on three to five studies, 18% were based on six to 10 studies, and 13% were based on more than 10 studies (IntHout et al., 2015). Commonly, these studies have small sample sizes (Davey et al., 2011). For instance, from 20,185 studies, 45% were very small, 29% were small, 14% were medium, and 11% were large (IntHout et al., 2015). With these conditions, sampling error might be influential with small sample sizes, which substantially causes bias to increase (Doncaster and Spake, 2018; Hamman et al., 2018; Lin, 2018), and few studies pose additional problems for inference (Günhan et al., 2020; Hamman et al., 2018).

When there are a small number of studies, the between-study variance is estimated imprecisely (Chung et al., 2013; IntHout et al., 2015). When heterogeneity is likely to be present, but there is no evidence, standard practice assumes homogeneity and applies a simpler fixed-effect meta-analysis. This is likely to yield false results. Thus, we should be concerned about zero or even low heterogeneity, which can be undetected or underestimated (Kontopantelis et al., 2013), because high heterogeneity is more regularly present in meta-analysis (Alba et al., 2016). Because the publications under few studies and small sizes present poor performance (Hardy and Thompson, 1998; Harwell, 1997; Sánchez-Meca and Marín-Martínez, 1997), some researchers argue for excluding small studies from meta-analyses (Lin, 2018; Turner et al., 2013).

1.4.3.1 A Review and Recommendation of heterogeneity variance estimators

Between-study estimators received significant attention in meta-analysis publications (Novianti et al., 2014). There are 20 methods for estimating the between-study variance τ^2 (Langan, 2015) and the majority of these estimators are based on the method of moments (van Aert and Jackson, 2018). A list of these estimators, along with the abbreviations used in this dissertation, are given in Table A.1.

Nevertheless, estimating the heterogeneity variances are usually inaccurate (Higgins et al., 2009; Turner et al., 2012). The estimators may produce divergent, or even contradictory, results (Novianti et al., 2014). For example, Langan (2015) showed the conflicting results of the heterogeneity variance from 18 different estimators that compare hawthorn extract with a placebo for treatment of chronic heart failure (see (Guo et al., 2008)); $\hat{\tau}^2$ ranged from 0 (by using Cochran's ANOVA method) up to 24.56 (by using Rukhin's simple estimator BP).

Table 1.2: *Conflicting Results of Estimating the Heterogeneity Variance*

Estimators	$\hat{\tau}^2$	**Estimators**	$\hat{\tau}^2$
DL	6.56	ML	1.27
DL_P	6.56	$REML$	9.31
CA	0	$ARML$	7.16
PM	5.88	SJ	13.93
SJ_{CA}	0.01	HM	11.14
BP	24.56	HS	0.80

of heterogeneity (Langan, 2015). Many simulation studies have been conducted to establish which methods for estimating heterogeneity have the most reasonable properties.

In terms of bias, Panityakul et al. (2013) summarized the results between six estimators (namely HE, DL, ML, REML, SJ, and MP, full names are shown in Table A.1) when the number of studies is between $m = 10$ and $m = 30$ with small and large sample sizes. It has shown that with small sample sizes, MP implements the best estimate when m=10, and

HE seems to be best, followed by MP when m= 30. While with large sample size, REML appears to perform best followed by MP when m=10, and with 30studies, MP appears to be performing best followed by REML. All recommended estimators have positively biased close to zero. Panityakul et al. (2013)) noted that REML is approximately unbiased and performing better than others with large sample sizes (between 100 and 300 per group); REML was also recommended in the other independent study by Viechtbauer (2005) but PM was not included in the study. REML estimates require an iteration that fewer than 0.02% failed to converge to a heterogeneity variance estimate with a small number of studies and significant differences in study sizes (Kontopantelis et al., 2013; Langan, 2015).

Novianti et al. (2014) recommended two estimators, DL_2 and PM over DL, HE, REML, SJ,SJ_{HE} with different numbers of studies (k=10, 15, 20, 30, and 50.). On the other hand, Kontopantelis et al. (2013) and Viechtbauer (2005) noted that when within-study variances are known and when the number of studies are large, DL becomes asymptotically unbiased.

Sidik and Jonkman (2007) compared seven estimators (namely HE, MM, ML, REML, EB, MV, MV_{vc}) in terms of bias and mean squared error by using Monte Carlo simulation. The results show that when the number of studies is large (i.e. $k > 30$), the HE provided accurate estimation; when the heterogeneity variance is moderate to large, both MV_{vc} and EB are the best overall.

Langan (2015) did further research because the comparisons above were based on a small subset of heterogeneity estimation methods and it cannot be considered as conclusive. For each simulated meta-analysis, the author calculated heterogeneity variance estimates from the 14 following methods:

$$CA,\ DL,\ PM,\ PM_{DL},\ PM_{CA},\ HM,\ HS,\ SJ,\ SJ_{CA},\ ML,\ REML,\ B0,\ BP,\ and\ MBH$$

These methods were generated in five different ways based on the sample size: small-to-medium study sizes, small equally sized studies, medium equally sized studies, small and large studies, and large studies only. Langan (2015) showed that estimates calculated from $B0$, BP, MBH, SJ, SJ_{CA}, HS and ML have poor properties and should not be used.

Moreover, the author recommended to exclude CA, PM_{CA}, and HM from the consideration because there are alternative methods that have equal or better properties (namely DL, PM_{DL}, PM and REML).With large differences in study size, DL and PM perform best. PM_{DL}L and PM are more robust with inaccurate within-study variances; both methods have similar properties. REML commonly has a low mean squared error, but only with few studies, and significant differences in study size. The iterative method failed to yield a REML estimate.

Langan et al. (2019) recommended using REML, although the author believed that the two-step DL estimator (PM_{DL}) and REML have the same properties. Because of that, REML is already widely known and available in most statistical software packages. Additionally, Hönekopp and Linden (2022) suggested using the REML estimator because it remains a best estimator across varied circumstances.

From several previous simulation studies, we see that no method is perfect but some methods work better than others. Most researchers recommended using the restricted maximum likelihood (REML) by Viechtbauer (2005) and Paule-Mandel (PM) by Paule and Mandel (1982) because they perform better than others.

1.5 Organization of the Dissertation

The purpose of this dissertation is multifaceted. Chapter 2 focuses on finding the best approach (meta-analysis or mixed model) that researchers should use when analyzing multi-laboratory replication studies. However, because there are conflicting recommendations about the accuracy estimators on meta-analysis, we have first to find the optimal estimation of summary effect size by conducting around 5,760 meta-analysis simulations, and second, we must compare the two approaches after completing another simulation of the mixed model, with 80 scenarios. In Chapter 3, we develop a sensitivity-analysis approach for determining the significance of effect size estimates for both meta-analysis and mixed-model estimators. These methods allow researchers to consider a range of different values for the across-lab variance and help them determine for what values the estimate of the effect size is statistically significant. In Chapter 4, we reexamine the last failure replication study that was based on great collaborative efforts involving 37 researchers from 18 universities with original authors and 2,281 participants. Then, we apply what we learn in Chapter 2 and Chapter 3 to the Many Labs 4 projects. Chapter 5 concludes the entire work and gives brief information about future work.

Chapter 2

On the Use of Meta-Analysis Techniques for Multi-Lab Experiments

2.1 Introduction

There has been a growing concern in almost every scientific field about the trustworthiness of statistically significant findings. Recent research has shown that results from a large fraction of experimental studies are not able to be replicated by follow-up studies—a phenomenon known as the *replicability crisis*. For example, the journal Nature conducted a survey of 1,500 scientists and found that 70% of those polled had been unable to replicate others' results, and 50% could not replicate even their own results (Baker, 2016).

To help combat the replicability crisis, *multi-lab experiments*—those that are conducted at multiple facilities simultaneously—have substantially grown in popularity. Some multi-lab experiments are explicitly conducted to verify or contradict controversial scientific findings (Ebersole et al., 2016, 2020; Jaljuli et al., 2023; Klein et al., 2014b, 2018). Others are conducted explicitly to ensure that potential findings are replicable prior to publication. Overall, multi-lab experiments help ensure the viability of a scientific result if multiple facilities obtain the same result independently.

On the other hand, meta-analysis has long been regarded as a potential solution to the replicability crisis (Allen and Preiss, 1993; Braver et al., 2014; Collaboration, 2015; Fabrigar and Wegener, 2016; McShane et al., 2016; Nelson et al., 2018; Sharpe and Poets, 2020). Rather than relying on the result of one study, meta-analysis collates results from many studies and aims to determine to what degree the finding has been clearly replicated (Allen and Preiss, 1993; Eden, 2002). Meta-analytic methods tend to provide more statistical power than those obtained from singular studies, raising the probability of discovering associations of interest (Jain et al., 2019; Konstantopoulos, 2006; Sartal et al., 2021). However, meta-analysis is not a perfect solution to the replicability crisis—estimates will only be reliable if the studies used to construct a meta-analysis are selected carefully, without introducing selection bias, and if the original selected studies are of sufficiently high quality (Eisend and Tarrahi, 2014). If the original studies suffer from issues including p-hacking, reporting errors, and fraud, a meta-analysis of these studies may exacerbate these issues, furthering the replicability crisis (Egger et al., 1998; Ioannidis, 2010, 2016; Nelson et al., 2018; Tebala, 2015).

There is currently some ambiguity about best practices about how to estimate effect sizes from multi-lab experiments. Some studies collate results from multi-lab experiments in a similar fashion to meta-analysis—despite the availability of individual-level data—and use meta-analysis estimators to find estimates of effect sizes (Chatard et al., 2020; Klein et al., 2022). Another approach is to use methods that incorporate individual-level data, for example, linear regression, to obtain estimates for treatment effects and residual standard errors, which can then be combined to obtain an estimate of an effect size. However, little work has been performed on comparing the performance of these two approaches.

In this study, we investigate the efficacy of meta-analysis techniques in estimating effect sizes for multi-lab studies. We begin by allowing treatment effects to vary randomly across labs. We then use the random-effects meta-analysis framework to estimate effect sizes, and we identify the best performing estimators from a large collection of candidates. We compare our best-performing meta-analysis estimators to those obtained from a linear mixed model (Milliken and Johnson, 2009). We show that the meta-analysis approach and the linear mixed model approach yield similar results. Additionally, we find that both methods for estimation suffer from the same significant pitfall—estimation of across-lab variability in treatment effects is often inconsistent and unreliable when the number of labs included in the study are small. Finally, we find that Hartung-Knapp standard errors (Hartung and Makambi, 2003) for random-effects meta-analysis tend to be robust to inaccurate estimates of across-lab variability, and provide an overall reasonable approach for estimating effect sizes from multi-lab experiments.

This study is organized as follows. Section 2.2 gives a summary of notation and summarizes the random effects meta-analysis. Section 2.3 introduces the estimating variance components. In section 2.4 we analyze multi-lab studies with mixed linear Models. In sections 2.5 and 2.6, we conduct simulation studies based on both meta analysis and mixed model respectively. Section 2.7 shows the comparison between meta analysis and mixed models. Section 2.8 provides the conclusion.

2.2 Random-Effects Meta-Analysis

Because there are a lot of different notations in meta-analysis and mixed models, which confuse the reader, it is important to define them.

Table 2.1: *Notations*

Abbreviations	**Definitions**
μ_i	The population mean for treatment i
θ_j	A lab effect that impacts all responses in study j
ζ_{ij}	A lab-by-treatment effect that impacts responses given treatment i in study j
ε_{ijk}	The experimental error

(To be continued)

Abbreviations	**Definitions**
σ_{e_j}	Standard deviations for experimental errors
σ_e	$\sigma_{e_j} = \sigma_e$ across labs
σ_ζ/σ_e	Referred as the *environmental effect ratio*, a measure of the magnitude of the treatment-by-lab variability relative to the experimental error.
σ_j^2	Within study variance
τ^2	Between study variance
n_{ij}	The sample sizes in groups i in study j
$\tilde{n}_j$	The harmonic mean of n_{1j} and n_{1j}, $(1/n_{1j} + 1/n_{2j})$
υ_j	Degrees of freedom, $\upsilon_j = n_{1j} + n_{2j} - 2$
n_i	$\sum_{j=1}^{m} n_{ij}$
N_j	$N_j = n_{1j} + n_{2j}$
δ_j	The population standardized mean difference (SMD) effect sizes
$\hat{\delta}_j$	Standard estimators for the study-level effect sizes
$\hat{\delta}_{C.j}$	Effect size estimation by Cohen's d
$\hat{\delta}_{H.j}$	Effect size estimation by Hedges' g
$S_{p.j}$	Pooled standard deviation for the effect size in study j
$s_{H.j}^2$	Within-study variances using individual effect sizes, by Hedges' g
$\hat{w}_j$	The inverse variance of each effect size
Δ	The true overall Standardized Mean Difference (SMD)
$\hat{\Delta}_C$	Averaging a set of independent effect sizes by Cohen's d
$\hat{\Delta}_H$	Averaging a set of independent effect sizes by Hedges' g
$\hat{\Delta}_{HS}$	Averaging a set of independent effect sizes by the sample size of each study by Hunter and Schmidt
J_j	Bias correction term of Hedge's estimator

(To be continued)

Abbreviations	Definitions
γ_j	The term of within study variance, $\approx \frac{1}{2}$ with large sample sizes
m	Number of studies
$\mathbf{j}_{n_i}$	A vector of n_i ones

Consider two treatment conditions—treatment 1 and treatment 2—and a set of m independent studies, numbered 1 through m, that are designed to estimate the effect size of the difference between these two treatments. Meta-analysis is a statistical tool used to combine these estimates into one overall estimate that is designed to be more accurate than any estimate from any single study (Hedges, 1981, 1982).

Recent work has considered the use of meta-analysis techniques to estimate effect sizes from multi-lab experiments. In these applications, each of the m studies is conducted within its own lab, and hence, meta-analysis methods need to account for differing environments across labs when estimating effect sizes. The lab environment encompasses many factors, including lab protocol and personnel, the facilities at the lab, the weather at the time of the experiment, and the time of year that the experiment is conducted. Differences in these environments may lead to variable treatment effects across labs (Higgins et al., 2021). In practice, these differences in treatment effects are often treated as random effects. Thus, meta-analytic techniques to analyze data from multi-lab studies often use a *random-effects meta-analysis* (REMA) model.

To make this rigorous, we begin with the following model of response for multi-lab studies, which is often inherently (though not explicitly) assumed by REMA models:

$$Y_{ijk} = \mu_i + \theta_j + \zeta_{ij} + \varepsilon_{ijk},\ i = 1, 2,\ j = 1, \dots, m,\ k = 1, .., n_{ij}. \tag{2.1}$$

Here, Y_{ijk} is the response for the k^{th} observation given treatment i in study j; μ_i is the population mean for treatment i; θ_j is a lab effect that impacts all responses in study j; ζ_{ij} is a lab-by-treatment effect that impacts responses given treatment i in study j; and ε_{ijk} is the experimental error.

It is often assumed that the experimental errors are independent random variables with distribution $\varepsilon_{ijk} \sim N(0, \sigma_{e_j}^2)$. While some meta-analysis methods permit the standard deviations for experimental errors σ_{e_j} to vary across labs, we proceed in our discussion under the assumption of *homogeneous within-study standard deviations*— that $\sigma_{e_j} = \sigma_e$ is the same across labs. Under this assumption, the overall effect size can be unambiguously defined as

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma_e}. \tag{2.2}$$

Random differences in treatment effects assumed by a REMA model can be implemented in (2.1) through distributional assumptions on the θ_j and ζ_{ij} terms:

$$\theta_j \sim N(0, \sigma_\theta^2),\ \zeta_{ij} \sim N(0, \sigma_\zeta^2), \tag{2.3}$$

where the θ_j, ζ_{ij} and ε_{ijk} random variables are assumed to be mutually independent. Under these assumptions, the true *study-level effect size* for lab j is

$$\delta_j = \frac{\mu_{1j} - \mu_{2j}}{\sigma_e} = \frac{\mu_1 + \zeta_{1j} - (\mu_2 + \zeta_{2j})}{\sigma_e} = \Delta + \frac{\zeta_{1j} - \zeta_{2j}}{\sigma_e}. \tag{2.4}$$

It follows from (2.3) that

$$\delta_j \sim N\left(\Delta, 2\left(\frac{\sigma_\zeta}{\sigma_e}\right)^2\right). \tag{2.5}$$

In the replicability literature, the σ_ζ/σ_e term may be referred to as the *environmental effect ratio* (EER) (Higgins et al., 2021), a measure of the magnitude of the treatment-by-lab variability relative to the experimental error. Note that, under the model of response in (2.1), standard estimators $\hat{\delta}_j$ for the study-level effect sizes will follow a non-central t-distribution (see Section 2.2.1 for details).

Since meta-analytic methods aggregate effect sizes within each study, REMA models may be posited by imposing assumptions about the distribution of effect sizes rather than on the responses themselves. Since the t-distribution converges asymptotically to a normal distribution, these assumptions often include normally distributed estimated effect sizes:

$$\hat{\delta}_j \sim N(\delta_j, \sigma_j^2), \ j = 1, \ldots, m, \tag{2.6}$$

$$\delta_j \sim N(\Delta, \tau^2). \tag{2.7}$$

In this model, the τ^2 term plays the same role as the EER in allowing heterogeneity in effect sizes across labs: $\tau^2 = 2(\sigma_\zeta/\sigma_e)$. The *sampling error* σ_j^2 is unique to each study—however, under the assumption of homogeneous within-study standard deviations, differences in σ_j^2 are due solely to differences in sample sizes (Hedges and Vevea, 1998; Higgins and Thompson, 2002; Rubio-Aparicio et al., 2020; Siegel et al., 2021).

2.2.1 Estimation of Study-Level Effect Sizes

The study-level effect sizes δ_j are commonly estimated by Cohen's d, denoted by $\hat{\delta}_{C.j}$, or Hedges' g, denoted by $\hat{\delta}_{H.j}$ (Lin and Aloe, 2021). We now describe in detail these estimators.

Suppose that study j has n_{kj} units given Treatment k and let S_{kj} denote the sample standard deviation of their responses, $k \in \{1, 2\}$. The Cohen's d estimator is given by

$$\hat{\delta}_{C.j} = \frac{\bar{Y}_{1j\cdot} - \bar{Y}_{2j\cdot}}{S_{p.j}} \tag{2.8}$$

where $S_{p.j}$ pooled standard deviation for the effect size in study j:

$$S_{p.j} = \sqrt{\frac{(n_{1j} - 1)S_{1j}^2 + (n_{2j} - 1)S_{2j}^2}{n_{1j} + n_{2j} - 2}}. \tag{2.9}$$

The exact distribution of $\sqrt{n_{1j}n_{2j}/(n_{1j} + n_{2j})}\ \hat{\delta}_{C.j}$ is a non-central t-distribution with $n_{1j} + n_{2j} - 2$ degrees of freedom, and non-centrality parameter $\sqrt{n_{1j}n_{2j}/(n_{1j} + n_{2j})}\ \delta_j$. In practice, however, it is often approximated using a normal distribution if the number of units in study

j is sufficiently large:

$$\hat{\delta}_{C.j} \dot{\sim} N(\Delta, \tau^2 + \sigma_j^2). \tag{2.10}$$

Hedges' g was introduced as a way to correct for the bias in Cohen's d (Hedges, 1981). The Hedges' g estimator is given as

$$\hat{\delta}_{H.j} = J_j \left(\frac{\bar{Y}_{1j.} - \bar{Y}_{2j.}}{S_{p.j}} \right) \tag{2.11}$$

where

$$J_j = \frac{\Gamma((n_{1j} + n_{2j} - 2)/2)}{\sqrt{(n_{1j} + n_{2j} - 2)/2}\ \Gamma[((n_{1j} + n_{2j} - 2) - 1)/2]} \tag{2.12}$$

Note that, J_j is always less than 1 since meta-analyses often contain studies with small sample sizes (Lin, 2018). However, as n_{1j} and n_{2j} grow large, $J_j \rightarrow 1$ (see appendix A.2.1 for further details), and so, Hedges' g converges to Cohen's d.

Several studies have criticized the use of Hedges' g in meta-analysis. While Hedges' g is an unbiased estimator for the study-level effect size, the unbiasedness may not carry over for meta-analysis (Lin and Aloe, 2021), and in fact, several studies assessing the performance of meta-analysis estimators have indicated that estimates using Hedges' g may be more biased than those that use Cohen's d (Lin, 2018; Lin and Aloe, 2021). However, under our multi-lab setting, we find that Hedges g tends to outperform Cohen's d when estimating the overall effect size Δ.

2.2.2 Combining Study-Level Effect Size Estimates

Prior to discussing how meta-analysis techniques estimate within and between-study variances, we first focus on common methods on how study-level effect size estimates $\hat{\delta}_j$ are combined to create an overall estimate $\hat{\Delta}$ of the effect size Δ. Although many of the most common meta-analysis methods rely on estimates of the variance components, the methods for combining study-level effect sizes to form an estimate $\hat{\Delta}$ are well-established and are largely without controversy. On the other hand, methods for estimating variance components have many subtle intricacies that can dramatically effect the reliability of their estimates, thereby affecting the overall estimator $\hat{\Delta}$. As such, we prefer to focus on this discussion in a separate section (see Section 2.3).

A primary goal of meta-analysis is to combine information from a set of studies to produce an estimate of an unknown parameter (Langan, 2015; Marin-Martinez and Sánchez-Meca, 2010). In the case of estimating an effect size, when variances of study-level effect sizes are known, combining estimates using a weighted sum, where the weight of an effect size is equal to the inverse of its variance, is optimal (Hedges, 1983; Shahar et al., 2017). Thus, the most commonly-used meta analysis estimators combine studies in this way, where study-level variances are estimated using data:

$$\hat{\Delta}_H = \frac{\sum_{j=1}^{m} \hat{w}_j \; \hat{\delta}_{H.j}}{\sum_{j=1}^{m} \hat{w}_j}. \tag{2.13}$$

Under the random-effects meta-analysis model, the weights are given by $\hat{w}_j = 1/(s^2_{H.j} + \hat{\tau}^2)$, where $s^2_{H.j}$ is an estimate of the within-study variance σ^2_j and $\hat{\tau}$ is an estimate of the between-study variance τ. Thus, the accuracy of the within-study and between-study variance estimates is critical for obtaining a precise overall effect estimate.

The variance of $\hat{\Delta}_H$ is commonly estimated using (Borenstein et al., 2021)

$$\hat{V}(\hat{\Delta}_H) = \frac{1}{\sum_{j=1}^{m} \hat{w}_j}, \tag{2.14}$$

and Wald confidence intervals are often used to provide interval estimates for the effect sizes:

$$\hat{\Delta}_H \pm z_{(1-\alpha/2)} \sqrt{\hat{V}(\hat{\Delta}_H)}, \tag{2.15}$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the normal distribution.

Since both the within-study and the between-study variances have to be estimated, this procedure only presents reliable results when there is a large number of studies and very little or zero heterogeneity variance (Biggerstaff and Tweedie, 1997; Follmann and Proschan, 1999; Hartung and Makambi, 2003). One proposed way to improve confidence interval estimates is to use quantiles from a t-distribution with $m-1$ degrees of freedom, where m is the number of studies used in the meta-analysis (Follmann and Proschan, 1999; Hartung and Makambi, 2002; Knapp and Hartung, 2003; Langan, 2015; Rukhin, 2013; Sánchez-Meca and Marín-Martínez, 2008):

$$\hat{\Delta}_H \pm t_{(1-\alpha/2,m-1)}\sqrt{\hat{V}(\hat{\Delta}_H)}. \tag{2.16}$$

However, this adjustment for the formation of confidence intervals does not overcome other issues with using the estimate $\sqrt{\hat{V}(\hat{\Delta}_H)}$ to perform inference and form confidence intervals, especially when the number of studies and/or the number of observations within each study is small (Böhning et al., 2002; Jackson and White, 2018; Lin and Aloe, 2021). This variance estimate is quite sensitive to the estimates of the variance components $s^2_{H.j}$ and $\hat{\tau}^2$, which can exhibit significant bias when sample sizes are small (see Section 2.3 for details) (Doncaster and Spake, 2018; Hamman et al., 2018; Lin, 2018; Marin-Martinez and Sánchez-Meca, 2010; Sánchez-Meca and Marin-Martinez, 1998; Sidik and Jonkman, 2006).

As an alternative that is more robust to incorrectly estimated variance components, Hartung (1999) proposes to use the overall effect-size variance estimate

$$\hat{V}_{HK}(\hat{\Delta}_H) = \frac{\sum_{j=1}^{m} \hat{w}_j (\hat{\delta}_{H.j} - \hat{\Delta}_H)^2}{(m-1)\sum_{j=1}^{m} \hat{w}_j}. \tag{2.17}$$

Using the Hartung-Knapp (HK) variance estimate, a confidence interval can then obtained using

$$\hat{\Delta}_H \pm t_{(m-1,1-\alpha/2)}\sqrt{\hat{V}_{HK}(\hat{\Delta}_H)}. \tag{2.18}$$

Many studies have concluded that HK variance estimates lead to statistical inferences and confidence intervals that are much closer to their nominal significant level α than most other meta-analysis approaches (Knapp and Hartung, 2003; Langan, 2015; Langan et al., 2019; Rukhin, 2013; Sánchez-Meca and Marín-Martínez, 2008). In our simulation results (see Section 2.5.4), we also find this to be true.

Finally, some estimators attempt to overcome the effect of variance component estimates $s_{H.j}^2$ and $\hat{\tau}^2$ from impacting overall effect sizes estimates by using estimators for the effect size that do not incorporate these component estimates—for example, taking an unweighted mean of effect sizes or weighting by the size of the sample within the study (Hamman et al., 2018; Sánchez-Meca and Marin-Martinez, 1998). However, often these methods do not improve the precision of overall effect size estimates; estimated variances for these estimators will still depend on these components.

2.3 Estimating variance components

We now turn our attention to the problem of estimating the within-study variances σ_j^2 and the between study variance τ. As we have shown, estimates of the overall effect size depend on the estimates of the variance components $s_{H.j}^2$ and $\hat{\tau}^2$. These variance component estimates may appear in the formula for the estimator of the effect size, the standard error of the estimator, or both. Thus, as we will confirm in Section 2.5 , accurate estimation of these quantities–especially τ—is critical for getting accurate results from meta-analysis.

2.3.1 Estimating Within-Study Variance

Using properties of the non-central t-distribution, the exact variance of the within-study variance of $\hat{\delta}_{H.j}$ conditional on the study-level effect size δ_j is

$$\sigma_j^2 = Var(\hat{\delta}_{H.j}|\delta_j) = \frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\left(J_j^2\ \tilde{n_j} + \frac{\gamma_j\ J_j^2\ \delta_j^2}{(n_{1j}+n_{2j}-2)}\right) \tag{2.19}$$

where $\tilde{n_j} = 1/n_{1j} + 1/n_{2j}$ and

$$\gamma_j = (n_{1j}+n_{2j}-2) - \frac{(n_{1j}+n_{2j}-4)}{J_j^2}. \tag{2.20}$$

Note that this is a function of the study-level effect sizes δ_j. Estimates of this variance depend largely on how these effect sizes are estimated and whether to include terms in the estimate that vanish asymptotically (as sample sizes within each study become large).

A natural estimator for σ_j^2 can be obtained simply by replacing δ_j with an estimate of this quantity:

$$s_{H.j}^2 = \frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\left(J_j^2\ \tilde{n_j} + \frac{\gamma_j\hat{\delta}_{H.j}^2}{(n_{1j}+n_{2j}-2)}\right). \tag{2.21}$$

However, there is some debate whether the inclusion of such nuance in the estimation of σ^2 is worthwhile, or whether its better to use a simpler, more parsimonious form for the estimator. Additionally, it is not clear, under the assumption of homogeneous within-study standard deviations, whether to use the estimates of the within-study effect sizes or an estimate of the overall effect size. Hence, several alternatives of this variance have been proposed.

In practice, the most commonly-used estimator of σ_j^2 is

$$s_{H.j}^2 = \tilde{n_j} + \frac{\hat{\delta}_{H.j}^2}{2(n_{1j}+n_{2j})}. \tag{2.22}$$

This estimator can be directly obtained from (2.19) after applying the approximations $J_j^2 \approx 1$, $(n_{1j}+n_{2j}-2)/(n_{1j}+n_{2j}-4) \approx 1$ and $\gamma \approx 0.5$, see Sections A.2 for farther details. This estimator is recommended by Hedges and Vevea (1998), Hedges and Olkin (2014), Koricheva et al. (2013) , Pigott (2012), and Lipsey and Wilson (2001).

On the other hand, Cooper et al. (2019) , Borenstein et al. (2021), and Stangl and Berry (2000) suggested to use Cohen's d's estimator instead of Hedges' g:

$$s_j^2 = \tilde{n_j} + \frac{\hat{\delta}_{C.j}^2}{2(n_{1j}+n_{2j})}. \tag{2.23}$$

Similarly, Sinha et al. (2011) and Egger et al. (2008) suggested to use Cohen's d's estimator but with different denominator:

$$s_j^2 = \tilde{n_j} + \frac{\hat{\delta}_{C.j}^2}{2(n_{1j}+n_{2j}-2)}. \tag{2.24}$$

while Lin (2018) suggested to use this same estimator, but substituting Cohen's d with Hedges' g:

$$s_j^2 = \tilde{n_j} + \frac{\hat{\delta}_{H.j}^2}{2(n_{1j} + n_{2j} - 2)} \tag{2.25}$$

These estimators are true on the situation where the sample size is very large. However, given that meta-analyses commonly include studies with small sample sizes (Davey et al., 2011; Lin, 2018), this omission can affect the accuracy of variance. We should be caution to perform within study variances if we have small sample sizes.

Since there is correlation between effect sizes $\hat{\delta}_{H.j}^2$ and their within-study variances $s_{H.j}^2$, which can inject bias into the variance estimate (Hamman et al., 2018; Lin, 2018; Lin and Aloe, 2021), Lin and Aloe (2021) developed a method that uses the sample-size-weighted SMD estimate (Hunter and Schmidt, 2004) ($\hat{\Delta}_{HS}$) to calculate adjusted within-study variances as follows:

$$s_j^2 = \tilde{n_j} + \frac{\hat{\Delta}_{HS}^2}{2(n_{1j} + n_{2j})}. \tag{2.26}$$

When there are few studies included in a meta-analysis, Doncaster and Spake (2018) recommend using a method that weights each study by the inverse of the mean-adjusted error variance (Hedges and Olkin, 2014) to eliminate or substantially reduce this bias. This method has the same accuracy on sample size but more accuracy in estimation of the effect sizes' significance:

$$s_j^2 = \tilde{n_j} + \frac{\left(\sum_{j=1}^m \hat{\delta}_{H.j}/m\right)^2}{2(n_{1j} + n_{2j})}. \tag{2.27}$$

On the other hand, Hedges (1982) recommended to use

$$s_j^2 = \tilde{n_j} + \frac{\left(\sum_{j=1}^m \hat{\delta}_{H.j}/s_{H.j}^2 \Big/ \sum_{j=1}^m 1/s_{H.j}^2\right)^2}{2(n_{1j} + n_{2j})}. \tag{2.28}$$

and Doncaster and Spake (2018) suggests to use

$$s_j^2 = J_j^2\left(\tilde{n_j} + \frac{\left(\sum_{j=1}^{m} \hat{\delta}_{C.j}/m\right)^2}{2(n_{1j}+n_{2j})}\right). \tag{2.29}$$

2.3.2 Methods for Estimating the Between-Study Variance

In random-effects meta-analysis, accurate estimation of the between-study variance τ^2 can be quite challenging, especially when the number of studies and the number of units within study are small (Higgins et al., 2009; Turner et al., 2012). There are currently over 20 methods available for estimating the between study variance τ^2 (Langan, 2015). These methods incorporate a variety of techniques, including method-of moments, maximum likelihood, and Bayesian estimation (van Aert and Jackson, 2018). The best estimator for a statistical analysis may vary across applications, and even for a given dataset, different estimators may produce divergent, or even contradictory, results (Novianti et al., 2014). Additionally, estimators commonly provide estimates of zero between-study variance, even when it is strongly present in the data (see section 1.4.3).

Note that difficulties with estimating between-study variances are not isolated to meta-analysis. Rather, this challenge persists for almost every setting that exhibits between-study variability (Jackson et al., 2017)—for example, mixed model approaches when individual data is available.

For random-effects meta-analysis, we find that most of the researchers recommend using either the restricted maximum likelihood (REML) (Viechtbauer, 2005) method or the Paule-Mandel (PM) (Paule and Mandel, 1982) method for estimating τ^2. Maximum Likelihood (ML) (Hardy and Thompson, 1996) is also a commonly-used approach. We now review these three methods for estimating the between-study variance.

2.3.2.1 Maximum Likelihood (*ML*) Estimator

The maximum likelihood (ML) estimator is an iterative method, so $\hat{\tau}^2_{ML}$ cannot be given in a simple formula.

$$\hat{\tau}^2_{ML} = \frac{\sum_{j=1}^{m} \hat{w}_j{}^2 \left(\left(\hat{\delta}_{H.j} - \hat{\Delta}_{ML} \right)^2 - s^2_{H.j} \right)}{\sum_{j=1}^{m} \hat{w}_j{}^2} \tag{2.30}$$

where

$$\hat{w}_j = \frac{1}{\hat{\tau}^2_{ML} + s^2_{H.j}} \tag{2.31}$$

$$\hat{\Delta}_{ML} = \frac{\sum_{j=1}^{m} \hat{\delta}_{H.j}/(\hat{\tau}^2_{ML} + s^2_{H.j})}{\sum_{j=1}^{m} 1/(\hat{\tau}^2_{ML} + s^2_{H.j})} \tag{2.32}$$

Given that $\hat{\Delta}_{ML}$, and $\hat{\tau}^2_{ML}$ have no closed-form solutions, a numerical approach algorithms by Eliason (1993) is needed, which is as follow

1. Choose an initial estimate of $\hat{\tau}^2_0$, say $\hat{\tau}^2_0 = 0$.
2. Apply $\hat{\tau}^2_0$ to $\hat{\Delta}_0$.

The iterative procedure is run until there is no significant difference ϵ between the current step s and the previous step s-1, $\mid \hat{\tau}^2_s - \hat{\tau}^2_{s-1} \mid < \epsilon$, and $\mid \hat{\Delta}_s - \hat{\Delta}_{s-1} \mid < \epsilon$.

If at any step the estimate is negative, then the process of iteration stops and we set $\hat{\tau}^2_{ML} = 0$ (Langan, 2015; Sangnawakij et al., 2019). The convergence is not guaranteed for any iteration method (Kontopantelis et al., 2013); a fewer than 0.02% of meta-analyses failed to converge to a heterogeneity variance estimate (Chung et al., 2013; Langan, 2015).

2.3.2.2 Restricted maximum likelihood (*REML*) Estimator

The restricted maximum likelihood (REML) estimator is an iterative method, so $\hat{\tau}^2_{REML}$ cannot be given in a simple formula.

$$\hat{\tau}^2_{REML} = \frac{\sum_{j=1}^{m} \hat{w}_j{}^2 \left(\left(\hat{\delta}_{H.j} - \hat{\Delta}_{REML} \right)^2 - s^2_{H.j} \right)}{\sum_{j=1}^{m} \hat{w}_j{}^2} + \frac{1}{\sum_{j=1}^{m} \hat{w}_j} \tag{2.33}$$

where

$$\hat{w}_j = \frac{1}{\hat{\tau}^2_{REML} + s^2_{H.j}} \tag{2.34}$$

$$\hat{\Delta}_{REML} = \frac{\sum_{j=1}^{m} \hat{\delta}_{H.j}/(\hat{\tau}^2_{REML} + s^2_{H.j})}{\sum_{j=1}^{m} 1/(\hat{\tau}^2_{REML} + s^2_{H.j})} \tag{2.35}$$

Given that $\hat{\Delta}_{REML}$, and $\hat{\tau}^2_{REML}$ have no closed-form solutions, a numerical approach algorithms by Eliason (1993) is needed, which is as follow

1. Choose an initial estimate of $\hat{\tau}^2_0$, say $\hat{\tau}^2_0 = 0$.
2. Apply $\hat{\tau}^2_0$ to $\hat{\Delta}_0$.

The iterative procedure is run until there is no significant difference ϵ between the current step s and the previous step s-1, $\mid \hat{\tau}^2_s - \hat{\tau}^2_{s-1} \mid < \epsilon$, and $\mid \hat{\Delta}_s - \hat{\Delta}_{s-1} \mid < \epsilon$

If at any step the estimate is negative, then the process of iteration stops and we set $\hat{\tau}^2_{REML} = 0$ (Langan, 2015; Sangnawakij et al., 2019). The convergence is not guaranteed for any iteration method (Kontopantelis et al., 2013); a fewer than 0.02% of meta-analyses failed to converge to a heterogeneity variance estimate (Chung et al., 2013; Langan, 2015).

2.3.2.3 Paule-Mandel (*PM*) Estimator

The *PM* estimator is an estimator that incorporates the method-of-moments technique.

It is given as follows:

$$\hat{\tau}^2_{PM} = \frac{Q - \sum_{j=1}^{m} \hat{w}_j s^2_{H.j} + \left(\sum_{j=1}^{m} \hat{w}_j^2 s^2_{H.j} / \sum_{j=1}^{m} \hat{w}_j \right)}{\sum_{j=1}^{m} \hat{w}_j - \left(\sum_{j=1}^{m} \hat{w}_j^2 / \sum_{j=1}^{m} \hat{w}_j \right)} \tag{2.36}$$

where

$$Q = \sum_{j=1}^{m} \hat{w}_j (\hat{\delta}_{H.j} - \hat{\Delta}_{PM})^2, \tag{2.37}$$

$$\hat{w}_j = 1/(s^2_{H.j} + \hat{\tau}^2_{PM}), \tag{2.38}$$

and

$$\hat{\Delta}_{PM} = \frac{\sum_{j=1}^{m} \hat{w}_j \hat{\delta}_{H.j}}{\sum_{j=1}^{m} \hat{w}_j} \tag{2.39}$$

Given that $\hat{\Delta}_{PM}$, and $\hat{\tau}^2_{PM}$ have no closed-form solutions, iteration is required.

1. Choose an initial estimate of $\hat{\tau}^2_0$, say $\hat{\tau}^2_0 = 0$.
2. Apply $\hat{\tau}^2_0$ to $\hat{\Delta}_0$.
3. The iterative procedure is run until there is no significant difference ϵ between the current step s and the previous step s-1, $\mid \hat{\tau}^2_s - \hat{\tau}^2_{s-1} \mid < \epsilon$, and $\mid \hat{\Delta}_s - \hat{\Delta}_{s-1} \mid < \epsilon$

If at any step the estimate is negative, then the process of iteration stops and we apply $\hat{\tau}^2_{PM} = 0$ (Langan, 2015; Sangnawakij et al., 2019). The process always converges regardless of the initial estimate (Rukhin et al., 2000).

2.4 Analyzing Multi-lab Studies with Mixed Linear Models

The model in equation 2.1 can be expressed in matrix notation as defined

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e} \ , \qquad \boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.40}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{j}_{n_1} & 0 \\ 0 & \mathbf{j}_{n_2} \end{bmatrix}. \tag{2.41}$$

Here, $n_i = \sum_{j=1}^{m} n_{ij}$, $\beta = (\mu_1, \mu_2)'$, and $\mathbf{j}_{n_i}$ is a vector of n_i ones.

The covariance of the error term $\mathbf{e}$ can be decomposed as follows:

$$\boldsymbol{\Sigma} = \sigma^2_\theta \mathbf{M}_\theta + \sigma^2_\zeta \mathbf{M}_\zeta + \sigma^2_e \mathbf{I}_n, \tag{2.42}$$

where

$$\mathbf{M}_\theta = \begin{bmatrix} \mathbf{J}_{n_{11}} & 0 & \cdots & 0 & \mathbf{J}_{n_{11}n_{21}} & 0 & \cdots & 0 \\ 0 & \mathbf{J}_{n_{12}} & \cdots & 0 & 0 & \mathbf{J}_{n_{12}n_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{J}_{n_{1m}} & 0 & 0 & \cdots & \mathbf{J}_{n_{1m}n_{2m}} \\ \mathbf{J}_{n_{21}n_{11}} & 0 & \cdots & 0 & \mathbf{J}_{n_{21}} & 0 & \cdots & 0 \\ 0 & \mathbf{J}_{n_{22}n_{12}} & \cdots & 0 & 0 & \mathbf{J}_{n_{22}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{J}_{n_{2m}n_{1m}} & 0 & 0 & \cdots & \mathbf{J}_{n_{2m}} \end{bmatrix}, \tag{2.43}$$

and

$$\mathbf{M}_\zeta = \begin{bmatrix} \mathbf{J}_{n_{11}} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{J}_{n_{12}} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{J}_{n_{1m}} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \mathbf{J}_{n_{21}} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \mathbf{J}_{n_{22}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \mathbf{J}_{n_{2m}} \end{bmatrix}. \tag{2.44}$$

Here, the matrix $\mathbf{I}_k$ is the $k \times k$ identity matrix, $\mathbf{J}_k$ is the $k \times k$ ones matrix, and $\mathbf{J}_{k\ell}$ is the $k \times \ell$ ones matrix.

There are several ways to estimate $\hat{\sigma}^2_\theta$, $\hat{\sigma}^2_\zeta$, and $\hat{\sigma}^2_e$: method of moments, maximum likelihood (ML), and restricted maximum likelihood (REML). When the design is balanced, all estimators are identical when the solutions are all positive. With unbalanced data, the method of moment estimators are the most straightforward to compute since the others require iterative algorithms (Beyer, 2019; Milliken and Johnson, 2009).

2.4.1 Confidence Interval

The $(1-\alpha)100\%$ confidence interval about $\boldsymbol{a'\hat{\beta}}$ is

$$\mathbf{a}'\hat{\beta} \pm t_{(\alpha/2,\hat{df})} \sqrt{\mathbf{a}' \left(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1} \mathbf{a}} \tag{2.45}$$

where the approximation of degree of freedom is obtained by

$$\hat{df} = \frac{2\left(\mathbf{a}' \left(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1} \mathbf{a}\right)^2}{\hat{Var}\left(\mathbf{a}' \left(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1} \mathbf{a}\right)} \tag{2.46}$$

The denominator degree of freedom is approximated by the Kenward-Roger (Kenward and Roger, 1997) or Satterthwaite approximation (Satterthwaite, 1941) approaches. However, since the Kenward-Roger approximation can be applied only to REML model Luke (2017), Satterthwaite approximation will be used.

2.5 Simulation Study with Multi-Lab Data

We now perform a large simulation study to determine the best meta-analysis estimators for analyzing multi-lab data. The simulations and analyses for all methods (ML, REML, and PM) are carried out in R using our own code. To perform a comprehensive comparison, 5,760 meta-analysis scenarios are conducted, but only a subset is presented. We use 10,000 repetitions for each combination

2.5.1 Simulation Scenarios

In our simulations, we consider meta-analyses that range between five and 30 studies. Each study compares a treatment group with a control group. In each meta-analysis, we generate sample sizes for each group, n_{1j} observations were generated by using $\mu_1 + \zeta_{1j} + \theta_j + e_{1jk}$ and n_{2j} observations by using $\mu_{2j} + \zeta_{2j} + \theta_j + e_{2jk}$. The data of ζ_{ij}, θ_j and e_{ijk} are generated as independent and identically distributed with $\zeta_{ij} \sim N(0, \sigma^2_\zeta)$, $\theta_j \sim N(0, \sigma^2_\theta)$, and $e_{ijk} \sim$

$N(0,\sigma_e^2)$, respectively, $i = 1, 2$ and j denotes the number of studies. The value of σ_e^2 was set to 1.

For each study j, we calculate the sample means $\bar{Y}_{1j.}$ and standard deviations S^2_{ij} of these observations. Then the standardized mean difference $\hat{\delta}_{C.j}$ and $\hat{\delta}_{H.j}$ and variance $s^2_{H.j}$ are calculated. The categorized values of Δ are defined as zero, low, medium, and high magnitude (0, 0.2, 0.5, 0.8 respectively; (Cohen, 2013)). However, because the between study variance is not great among the values of Δ we will consider only 0 and 0.5 when estimating τ^2.

The values of $\tau^2 (= 2\sigma^2_\zeta/\sigma^2_e)$ are selected to be 0.05 (considered minimum), 0.1, 0.3, 0.6, and 1.5 (considered maximum). The study sample sizes n_{1j} and n_{2j} are generated in two ways: equal and in the same range but not necessary equal. For equal sample size $n_{1j} = n_{2j}$, 10 is considered small and 50 considered large. For sample sizes that are not necessarily equal but in the same range, 10 to 20 is considered small and 40 to 60 is considered large. In this case, we select the sample sizes completely at random from this range.

2.5.2 Constructing Various Forms for Estimation for Heterogeneity Variance and the Overall Effect Size

Since researchers usually use simpler forms to approximate the within-study variance, and the exact estimate is not used, we will consider all changes of within-study variance.

Diagram 2.1: *Constructing 24 Within-Study Variance Estimators*

$$s^2_{H.j} = \overset{\text{Keep/Remove}}{\boxed{\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}}} \left(\underset{\text{Keep/Remove}}{\boxed{J_j^2}}\ \tilde{n}_j + \frac{\overset{\text{Keep/Change it to 1/2}}{\boxed{\gamma_j}}\ \overset{\text{Keep/Change it to } \hat{\Delta}^2_H}{\boxed{\hat{\delta}^2_{H.j}}}}{n_{1j}+n_{2j}\underset{\text{Keep/Remove}}{\boxed{-2}}} \right)$$

The within-study variance is mathematically estimated as given in the diagram above, using Hedges' g estimator. We will consider every change that can be made to this estimator. Each time, one term will be selected to keep, remove, or change. To reduce the number of estimators, we consider $\gamma_j/(n_{1j} + n_{2j} - 2)$ one term, which means that we keep this term or change it to $1/2(n_{n1j} + n_{2j})$. Doing that will provide 16 different estimators. Using similar strategies for estimating within-study variance by Cohen's d estimator, we will get eight different estimators. The $\hat{\Delta}_H^2$ in the diagram indicates the overall effect size by the unweighted method. After considering all changes to the form of within-study variance, we will have 24 estimators shown in Table 2.2.

Regarding between-study variance (τ^2), in each method (REML, ML, and PM), there will be 48 of τ^2; the τ^2 depends on both within-study variance and overall effect size (we have 24 estimators of within study variance $\big((1) - (24)\big)$ as defined above, and two different methods for estimating overall effect size—an unweighted method that is defined by (b) and an inverse variance method that is defined by (c)). The 48 estimators come from the combination of 24 estimators of s_j^2 × two methods $\big((b) - (c)\big)$ of the overall effect size, see Diagram 2.2 for further explanation.

Diagram 2.2: *Constructing 48 Between-Study Variance Estimators: REML Estimator*

$$\hat{\tau}^2_{REML} = \frac{\sum_{j=1}^{m} \hat{w}_j^{\,2}\Big((\hat{\delta}_j - \overset{\big((b),\ \boxed{(c)}\big)\overset{(*)}{\nearrow}}{\overset{\nearrow}{\boxed{\hat{\Delta}_{REML}}}})^2 - \overset{\big((1)-(24)\big)}{\overset{\nearrow}{\boxed{s_j^2}}}\Big)}{\sum_{j=1}^{m} \hat{w}_j^{\,2}} + \frac{1}{\sum_{j=1}^{m} \hat{w}_j}\ , \quad \hat{w}_j = 1/\Big(\hat{\tau}^2_{REML} + \overset{(*)}{\overset{\nwarrow}{\boxed{s_j^2}}}\Big)$$

Here, $\hat{\Delta}_{REML}$ is estimated by either the unweighted method (defined as (b)) or the inverse variance method (defined as (c)). $*$ indicates that any number that s_j^2 is selected from $\big((1) - (24)\big)$ will be used as well in both overall effect size (in (c) method) and $\hat{w}_j$. Any time Hedges' g estimator (or Cohen's d estimator) of s_j^2 is selected, the overall effect size will be averaged by Hedges' g estimator (or Cohen's d estimator). Similarly, with the estimator of

$\hat{\delta}_j$, if we select $s^2_{H.j}$ ($s^2_{C.j}$), then we will use $\hat{\delta}_{H.j}$ ($\hat{\delta}_{C.j}$).

Table 2.2: *24 Estimators of Within-Study Variance*

(1) $s^2_{C.j} = \tilde{n}_j + \frac{\hat{\delta}^2_{C.j}}{2(n_{1j}+n_{2j})}$	(13) $s^2_{C.j} = \tilde{n}_j + \frac{\gamma_j \hat{\delta}^2_{C.j}}{(n_{1j}+n_{2j}-2)}$
(2) $s^2_{H.j} = J^2_j \tilde{n}_j + \frac{\hat{\delta}^2_{H.j}}{2(n_{1j}+n_{2j})}$	(14) $s^2_{H.j} = J^2_j \tilde{n}_j + \frac{\gamma_j \hat{\delta}^2_{H.j}}{(n_{1j}+n_{2j}-2)}$
(3) $s^2_{H.j} = \tilde{n}_j + \frac{\hat{\delta}^2_{H.j}}{2(n_{1j}+n_{2j})}$	(15) $s^2_{H.j} = \tilde{n}_j + \frac{\gamma_j \hat{\delta}^2_{H.j}}{(n_{1j}+n_{2j}-2)}$
(4) $s^2_{C.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(\tilde{n}_j + \frac{\hat{\delta}^2_{C.j}}{2(n_{1j}+n_{2j})} \right)$	(16) $s^2_{C.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(\tilde{n}_j + \frac{\gamma_j \hat{\delta}^2_{C.j}}{(n_{1j}+n_{2j}-2)} \right)$
(5) $s^2_{H.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(J^2_j \tilde{n}_j + \frac{\hat{\delta}^2_{H.j}}{2(n_{1j}+n_{2j})} \right)$	(17) $s^2_{H.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(J^2_j \tilde{n}_j + \frac{\gamma_j \hat{\delta}^2_{H.j}}{(n_{1j}+n_{2j}-2)} \right)$
(6) $s^2_{H.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(\tilde{n}_j + \frac{\hat{\delta}^2_{H.j}}{2(n_{1j}+n_{2j})} \right)$	(18) $s^2_{H.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(\tilde{n}_j + \frac{\gamma_j \hat{\delta}^2_{H.j}}{(n_{1j}+n_{2j}-2)} \right)$
(7) $s^2_{C.j} = \tilde{n}_j + \frac{\hat{\Delta}^2_{C}}{2(n_{1j}+n_{2j})}$	(19) $s^2_{C.j} = \tilde{n}_j + \frac{\gamma_j \hat{\Delta}^2_{C}}{(n_{1j}+n_{2j}-2)}$
(8) $s^2_{H.j} = J^2_j \tilde{n}_j + \frac{\hat{\Delta}^2_{H}}{2(n_{1j}+n_{2j})}$	(20) $s^2_{H.j} = J^2_j \tilde{n}_j + \frac{\gamma_j \hat{\Delta}^2_{H}}{(n_{1j}+n_{2j}-2)}$
(9) $s^2_{H.j} = \tilde{n}_j + \frac{\hat{\Delta}^2_{H}}{2(n_{1j}+n_{2j})}$	(21) $s^2_{H.j} = \tilde{n}_j + \frac{\gamma_j \hat{\Delta}^2_{H}}{(n_{1j}+n_{2j}-2)}$
(10) $s^2_{C.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(\tilde{n}_j + \frac{\hat{\Delta}^2_{C}}{2(n_{1j}+n_{2j})} \right)$	(22) $s^2_{C.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(\tilde{n}_j + \frac{\gamma_j \hat{\Delta}^2_{C}}{(n_{1j}+n_{2j}-2)} \right)$
(11) $s^2_{H.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(J^2_j \tilde{n}_j + \frac{\hat{\Delta}^2_{H}}{2(n_{1j}+n_{2j})} \right)$	(23) $s^2_{H.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(J^2_j \tilde{n}_j + \frac{\gamma_j \hat{\Delta}^2_{H}}{(n_{1j}+n_{2j}-2)} \right)$
(12) $s^2_{H.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(\tilde{n}_j + \frac{\hat{\Delta}^2_{H}}{2(n_{1j}+n_{2j})} \right)$	(24) $s^2_{H.j} = \frac{(n_{1j}+n_{2j}-2)}{(n_{1j}+n_{2j}-4)} \left(\tilde{n}_j + \frac{\gamma_j \hat{\Delta}^2_{H}}{(n_{1j}+n_{2j}-2)} \right)$

The 24 within-lab variance estimators considered. The difference between (1)-(12) and (13)-(24) is the term $\gamma_j/(n_{1j}+n_{2j}-2)$*. In the first column,* $\gamma_j/(n_{1j}+n_{2j}-2)$ *is replaced to* $1/2(n_{1j}+n_{2j})$*. However, in the second column, we use the original term.*

The same strategy will follow for both $\hat{\tau}^2_{PM}$ and $\hat{\tau}^2_{ML}$, see Diagrams 2.3, and Diagram 2.4.

Diagram 2.3: *Constructing 48 Between-Study Variance Estimators: PM Estimator*

$$\hat{\tau}^2_{PM} = \frac{\sum_{j=1}^{m} \hat{w}_j \big(\hat{\delta}_j - \overset{\left((b),\ \overset{(*)}{\nearrow}\enclose{circle}{(c)}\right)}{\overset{\nearrow}{\enclose{circle}{\hat{\Delta}_{PM}}}}\big)^2 - \sum_{j=1}^{m} \hat{w}_j \overset{(*)}{\overset{\nearrow}{\enclose{circle}{s_j^2}}} + \Big(\sum_{j=1}^{m} \hat{w}_j^2 \overset{\left((1)-(24)\right)}{\overset{\nearrow}{\enclose{circle}{s_j^2}}} \Big/ \sum_{j=1}^{m} \hat{w}_j\Big)}{\sum_{j=1}^{m} \hat{w}_j - \Big(\sum_{j=1}^{m} \hat{w}_j^2 \Big/ \sum_{j=1}^{m} \hat{w}_j\Big)}\ , \quad \hat{w}_j = 1\Big/\Big(\hat{\tau}^2_{REML} + \overset{(*)}{\overset{\nwarrow}{\enclose{circle}{s_j^2}}}\Big)$$

Diagram 2.4: *Constructing 48 Between-Study Variance Estimators: ML Estimator*

$$\hat{\tau}^2_{ML} = \frac{\sum_{j=1}^{m} {\hat{w}_j}^2 \Big(\big(\hat{\delta}_j - \overset{\left((b),\ \overset{(*)}{\nearrow}\enclose{circle}{(c)}\right)}{\overset{\nearrow}{\enclose{circle}{\hat{\Delta}_{ML}}}}\big)^2 - \overset{\left((1)-(24)\right)}{\overset{\nearrow}{\enclose{circle}{s_j^2}}}\Big)}{\sum_{j=1}^{m} {\hat{w}_j}^2}\ , \quad \hat{w}_j = 1\Big/\Big(\hat{\tau}^2_{REML} + \overset{(*)}{\overset{\nearrow}{\enclose{circle}{s_j^2}}}\Big)$$

Regarding overall effect size, two methods are used, unweighted method (b) and inverse variance method (c). The unweighted method has two estimators: one by Hedges' g estimator and the other by Cohen's d estimator. The inverse variance method creates 24 estimators because of the 24 within-study variances. Thus, in total, we have 26 estimators of the overall effect size.

Diagram 2.5: *Constructing 24 Estimators for Overall Effect Size by using Inverse Variance Method*

$$\hat{\Delta} = \frac{\sum_{j=1}^{m} \hat{\delta}_j \Big/ \Big(\overset{\overset{(*)}{\nearrow}\enclose{circle}{(c)}}{\overset{\nwarrow}{\enclose{circle}{\hat{\tau}^2}}} + \overset{\left((1)-(24)\right)}{\overset{\nearrow}{\enclose{circle}{s_j^2}}}\Big)}{\sum_{j=1}^{m} 1\Big/\Big(\hat{\tau}^2 + s_j^2\Big)}$$

The $\hat{\tau}^2$ that is used in the overall effect size by inverse variance method may be estimated by ML, REML, and PM. It is estimated only by method (c) because there are no differences between (b) and (c) for estimating between study variance. As mentioned, $\ast$ indicates that any number that s_j^2 is selected from $\big((1)-(24)\big)$ will be used as well in $\hat{\tau}^2$; the estimator of $\hat{\delta}_j$ follows s_j^2; if we select $s_{H.j}^2$ ($s_{C.j}^2$), than we will use $\hat{\delta}_{H.j}$ ($\hat{\delta}_{C.j}$).

In relation to Type I Error, 48 test statistics T will be constructed (24 test statistic for both (b) and (c) methods). We explain the T test statistic by (c) method in the Diagram 2.6, and T test statistic by (b) method in Diagram 2.7.

Diagram 2.6: *Constructing 24 Test Statistic T by using Inverse Variance Method (c) (t-distribution)*

$$\frac{\hat{\Delta}}{\sqrt{\hat{V}(\hat{\Delta})}} = \frac{\sum_{j=1}^{m} \hat{\delta}_j \Big/ \Big(\overset{\overset{\overset{(\ast)}{\nearrow}}{\boxed{(c)}}\ \nwarrow}{\boxed{\hat{\tau}^2}} + \overset{\overset{\big((1)-(24)\big)}{\nearrow}}{\boxed{s_j^2}}\Big) \Big/ \sum_{j=1}^{m} 1\Big/\big(\hat{\tau}^2 + s_j^2\big)}{\sqrt{\Big(\sum_{j=1}^{m}\big(\hat{\tau}^2 + s_j^2\ \big)^{-1}\Big)^{-1}}}$$

As mentioned, the estimator of $\hat{\delta}_j$ follows s_j^2; if we select $s_{H.j}^2$ ($s_{C.j}^2$), then we will use $\hat{\delta}_{H.j}$ ($\hat{\delta}_{C.j}$). Additionally, $\ast$ indicates that any number that s_j^2 is selected from $\big((1)-(24)\big)$ will be used as well. In $\hat{\tau}^2$; the estimator of $\hat{\delta}_j$ follows s_j^2; if we select $s_{H.j}^2$ ($s_{C.j}^2$), than we will use $\hat{\delta}_{H.j}$ ($\hat{\delta}_{C.j}$).

Diagram 2.7: *Constructing 24 Test Statistic T By Using Unweighted Method (b)*

$$\frac{\hat{\Delta}}{\sqrt{\hat{V}(\hat{\Delta})}} = \frac{\sum_{j=1}^{m} \hat{\delta}_j / m}{\sqrt{\sum_{j=1}^{m}\Big(\underset{\underset{(b)}{\searrow}}{\boxed{\hat{\tau}^2}} + \underset{\underset{\big((1)-(24)\big)}{\searrow}}{\boxed{s_j^2}}\Big)/m^2}}$$

Likewise, the $\hat{\tau}^2$ that is used to find the variance of the overall effect size can be estimated by using (b) method.

We also want to consider the Type I Error that was derived from the Hartung-Knapp method because it generally has better coverage compared to the standard t-distribution when the number of studies are less than 20, based on previous literature, (see Diagram 2.8).

Diagram 2.8: *Constructing 24 Test Statistic T By Using Inverse Variance Method (Hartung-Knapp t-distribution)*

$$\frac{\hat{\Delta}}{\sqrt{\hat{V}(\hat{\Delta})}} = \frac{\sum_{j=1}^{m} \hat{\delta}_j \Big/ \Big(\overset{\overset{\overset{(*)}{\nearrow}}{\enclose{circle}{(c)}}}{\overset{\nwarrow}{\enclose{circle}{\hat{\tau}^2}}} + \overset{\overset{\big((1)-(24)\big)}{\nearrow}}{\enclose{circle}{s_j^2}} \Big) \Big/ \sum_{j=1}^{m} 1\big/\big(\hat{\tau}^2 + s_j^2\big)}{\sqrt{\sum_{j=1}^{m}(\hat{\delta}_j - \hat{\Delta})^2\big/\big(\hat{\tau}^2+s_j^2\big) \Big/ (m-1)\sum_{j=1}^{m} 1\big/\big(\hat{\tau}^2+s_j^2\big)}}$$

2.5.3 Structure of a Data Analysis

The goals of meta-analysis includes combining individual effect sizes in different studies, showing how these effect sizes are dispersed about the overall effect size, and showing whether they vary reasonably or exceedingly from study to study (Bakbergenuly et al., 2019; Borenstein et al., 2017; Higgins et al., 2009; Huedo-Medina et al., 2006; Mittlböck and Heinzl, 2006). Our simulations focus on both between-study variance and the overall effect size at the same time. Because they are correlated, we will not start with between-study variance and then move to the overall effect as previous researchers have done.

To perform a comprehensive comparison of heterogeneity variance estimators ($\hat{\tau}^2$) and overall effect size ($\hat{\Delta}$), we will compare 48 estimators $\big($combinations $\big((1)-(24)\big)\times\ \big((b),(c)\big)\big)$ and 26 the overall effect size $\big(2\ \times (b) +\ 24 \times\ (c)\big)$, where (1)-(24) defines the 24 estimators of within-study variance, (b) indicates the unweighted method and (c) indicates inverse variance method. This simulation will be conducted in three methods (REML, PM, and

ML). Results are produced from a total of 5,760 meta-analysis scenarios.

$$3\ (Methods) \times 24\ \big((1)-(24)\big) \times 2\ \big((b),(c)\big) \times 5\ (\tau^2) \times 4\ (N) \times 2\ (\Delta) = 5760$$

It can be challenging to compare this many simulations. Thus, as data analysis begins, it is important to determine the best strategy for organizing and analyzing it to aid in understanding the process as a whole.

In all simulations, the performance of the method (REML, PM, and ML) is assessed through the heterogeneity variance (in terms of bias, MSE, and the proportion of zero) and also their effect on the overall effect estimate, power, and Type I Error. Because both overall effect size and between-study variance rely on the within-study variance estimator, and because there are 24 within-study variance estimators $\big((1)-(24)\big)$, we decide to make the comparison based on s_j^2. Therefore, for each simulation, we compare 48 estimators of $\hat{\tau}^2$ (or 26 estimators of $\hat{\Delta}$) based on s_j^2. There are 12 estimators of s_j^2 that have the individual effect size $\hat{\delta}_j^2$, and there are 12 estimators that have overall effect size $\hat{\Delta}^2$; see Table 2.2 for details. In other words, there are 12 different forms of within-study variance with the term either $\hat{\delta}^2_{H.j}$ or $\hat{\delta}^2_{C.j}$; the other 12 estimators only replace the term ($\hat{\delta}^2_{H.j}$ and $\hat{\delta}^2_{C.j}$) to ($\hat{\Delta}^2_H$ and $\hat{\Delta}^2_C$), respectively. See the example below for further explanation.

$$(1)\quad s^2_{C.j} = \tilde{n_j} + \frac{\hat{\delta}^2_{C.j}}{2(n_{1j}+n_{2j})} \qquad v.s \qquad (7)\quad s^2_{C.j} = \tilde{n_j} + \frac{\hat{\Delta}^2_C}{2(n_{1j}+n_{2j})}$$

$$(2)\quad s^2_{H.j} = J_j^2\tilde{n_j} + \frac{\hat{\delta}^2_{H.j}}{2(n_{1j}+n_{2j})} \qquad v.s \qquad (8)\quad s^2_{H.j} = J_j^2\tilde{n_j} + \frac{\hat{\Delta}^2_H}{2(n_{1j}+n_{2j})}$$

Similarly, we make the comparisons (3) $v.s$ (9), (4) $v.s$ (10), (5) $v.s$ (11), (6) $v.s$ (12), and (13) $v.s$ (19) and so on.

In terms of the $\hat{\tau}^2$, we estimate with three different methods: REML, ML, and PM. All follow the same strategy, but we will explain $\hat{\tau}^2_{ML}$. Our comparison is based on s_j^2, and our first comparison of $\hat{\tau}^2_{ML}$ is based on estimators (1) and (7) of s_j^2. We have two different method — (b) and (c) — to average effect size; therefore, we will use estimators (1) and (7) of s_j^2 for both.

$$\frac{\sum_{j=1}^m \hat{w}_j{}^2\big((\hat{\delta}_{C.j} - \overset{(b)\nearrow}{\enclose{circle}{\hat{\Delta}_{ML}}})^2 - \overset{(1)\nearrow}{\enclose{circle}{s^2_{C.j}}}\big)}{\sum_{j=1}^m \hat{w}_j{}^2} \quad v.s \quad \frac{\sum_{j=1}^m \hat{w}_j{}^2\big((\hat{\delta}_{C.j} - \overset{(b)\nearrow}{\enclose{circle}{\hat{\Delta}_{ML}}})^2 - \overset{(7)\nearrow}{\enclose{circle}{s^2_{C.j}}}\big)}{\sum_{j=1}^m \hat{w}_j{}^2}$$

, and

$$\frac{\sum_{j=1}^m \hat{w}_j{}^2\big((\hat{\delta}_{C.j} - \overset{(c)\nearrow}{\enclose{circle}{\hat{\Delta}_{ML}}})^2 - \overset{(1)\nearrow}{\enclose{circle}{s^2_{C.j}}}\big)}{\sum_{j=1}^m \hat{w}_j{}^2} \quad v.s \quad \frac{\sum_{j=1}^m \hat{w}_j{}^2\big((\hat{\delta}_{C.j} - \overset{(c)\nearrow}{\enclose{circle}{\hat{\Delta}_{ML}}})^2 - \overset{(7)\nearrow}{\enclose{circle}{s^2_{C.j}}}\big)}{\sum_{j=1}^m \hat{w}_j{}^2}$$

In the next simulations, the s_j^2 will be selected as following for both methods $\big($using either (b) or (c)$\big)$, $(2)(b,c)$ $v.s$ $(8)(b,c)$, $(3)(b,c)$ $v.s$ $(9)(b,c)$, and $(4)(b,c)$ $v.s$ $(10)(b,c)$ and so on. If Cohen's d estimator (or Hedges' g estimator) is used in (c), then Cohen's d estimator (or Hedges' g estimator) also will be used in (b).

Regarding the the overall effect size, the comparison under ML will be:

$$\frac{\sum_{j=1}^m \hat{\delta}_{C.j}/\big(\overset{(c)\nearrow}{\enclose{circle}{\hat{\tau}^2_{ML}}} + \overset{(1)\nearrow}{\enclose{circle}{s^2_{C.j}}}\big)}{\sum_{j=1}^m 1/\big(\hat{\tau}^2_{ML} + s^2_{C.j}\big)} \quad v.s \quad \frac{\sum_{j=1}^m \hat{\delta}_{C.j}/\big(\overset{(c)\nearrow}{\enclose{circle}{\hat{\tau}^2_{ML}}} + \overset{(7)\nearrow}{\enclose{circle}{s^2_{C.j}}}\big)}{\sum_{j=1}^m 1/\big(\hat{\tau}^2_{ML} + s^2_{C.j}\big)} \quad v.s \quad \frac{1}{m}\sum_{j=1}^m \hat{\delta}_{C.j}$$

, and

$$\frac{\sum_{j=1}^m \hat{\delta}_{H.j}/\big(\overset{(c)\nearrow}{\enclose{circle}{\hat{\tau}^2_{ML}}} + \overset{(2)\nearrow}{\enclose{circle}{s^2_{H.j}}}\big)}{\sum_{j=1}^m 1/\big(\hat{\tau}^2_{ML} + s^2_{H.j}\big)} \quad v.s \quad \frac{\sum_{j=1}^m \hat{\delta}_{H.j}/\big(\overset{(c)\nearrow}{\enclose{circle}{\hat{\tau}^2_{ML}}} + \overset{(8)\nearrow}{\enclose{circle}{s^2_{H.j}}}\big)}{\sum_{j=1}^m 1/\big(\hat{\tau}^2_{ML} + s^2_{H.j}\big)} \quad v.s \quad \frac{1}{m}\sum_{j=1}^m \hat{\delta}_{H.j}$$

Similarly, we make the comparisons $(2)(c)$ $v.s$ $(8)(c)$ $v.s$ (b), $(3)(c)$ $v.s$ $(9)(c)$ $v.s$ (b), $(4)(c)$ $v.s$ $(10)(c)$ $v.s$ (b) and so on. Finally, for each parameter combination, the 48 Type I Errors and power by using t-distribution are compared using the 24 estimators of s_j^2, and two methods of averaging effect size.

$$\frac{\sum_{j=1}^{m}\hat{\delta}_{C.j}/m}{\sqrt{\sum_{j=1}^{m}\Big(\underset{(b)}{\boxed{\hat{\tau}^2_{ML}}}+\underset{(1)}{\boxed{s^2_{C.j}}}\Big)/m^2}}\quad v.s\quad \frac{\sum_{j=1}^{m}\hat{\delta}_{C.j}/(\hat{\tau}^2_{ML}+s^2_{C.j})\Big/\sum_{j=1}^{m}1/(\hat{\tau}^2_{ML}+s^2_{C.j})}{\sqrt{\Big(\sum_{j=1}^{m}(\underset{(c)}{\boxed{\hat{\tau}^2_{ML}}}+\underset{(1)}{\boxed{s^2_{C.j}}})^{-1}\Big)^{-1}}}\quad v.s$$

$$\frac{\sum_{j=1}^{m}\hat{\delta}_{C.j}/m}{\sqrt{\sum_{j=1}^{m}\Big(\underset{(b)}{\boxed{\hat{\tau}^2_{ML}}}+\underset{(7)}{\boxed{s^2_{C.j}}}\Big)/m^2}}\quad v.s\quad \frac{\sum_{j=1}^{m}\hat{\delta}_{C.j}/(\hat{\tau}^2_{ML}+s^2_{C.j})\Big/\sum_{j=1}^{m}1/(\hat{\tau}^2+s^2_{C.j})}{\sqrt{\Big(\sum_{j=1}^{m}(\underset{(c)}{\boxed{\hat{\tau}^2_{ML}}}+\underset{(7)}{\boxed{s^2_{C.j}}})^{-1}\Big)^{-1}}}$$

In preliminary simulations, the Hartung-Knapp standard errors yields vastly superior confidence intervals—the actual coverage is very close to the nominal significance level—than when using the estimated standard error from 2.14. Thus, we proceed using only Hartung-Knapp standard errors.

$$\frac{\sum_{j=1}^{m}\hat{\delta}_{C.j}/(\overset{(c)}{\boxed{\hat{\tau}^2_{ML}}}+\overset{(1)}{\boxed{s^2_{C.j}}})\Big/\sum_{j=1}^{m}1/(\hat{\tau}^2_{ML}+s^2_{C.j})}{\sqrt{\sum_{j=1}^{m}(\hat{\delta}_{C.j}-\hat{\Delta}_C)^2/(\hat{\tau}^2_{ML}+s^2_{C.j})\Big/(m-1)\sum_{j=1}^{m}1/(\hat{\tau}^2_{ML}+s^2_{C.j})}}\quad v.s$$

$$\frac{\sum_{j=1}^{m}\hat{\delta}_{C.j}/(\overset{(c)}{\boxed{\hat{\tau}^2_{ML}}}+\overset{(7)}{\boxed{s^2_{C.j}}})\Big/\sum_{j=1}^{m}1/(\hat{\tau}^2_{ML}+s^2_{C.j})}{\sqrt{\sum_{j=1}^{m}(\hat{\delta}_{C.j}-\hat{\Delta}_C)^2/(\hat{\tau}^2_{ML}+s^2_{C.j})\Big/(m-1)\sum_{j=1}^{m}1/(\hat{\tau}^2_{ML}+s^2_{C.j})}}$$

None of the 48 estimators of $\hat{\tau}^2$, or 26 overall effect sizes give optimal results in all scenarios. Thus, for each scenario, we will select the best quarter of the total estimators. Because there are 48 estimators of $\hat{\tau}^2$, we will select 12, and because there are 26 overall effect sizes, we will select 6 (see Figure 2.1). If more than a quarter of the estimators have very similar results, we will select them all. Our selection for each parameter combination is based on the overall behavior of the estimator through a different number of studies.

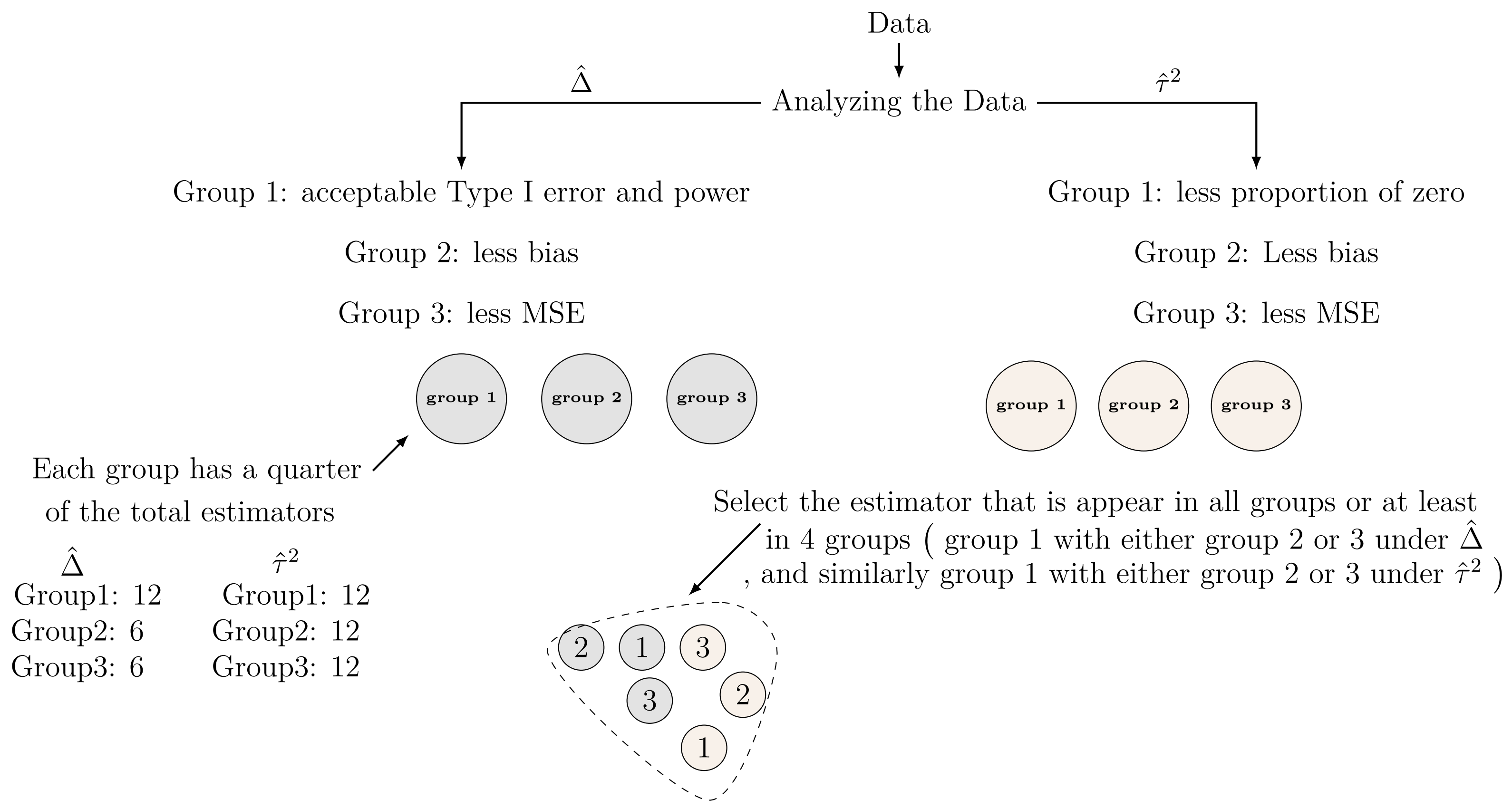


Figure 2.1: *Structure of a data analysis (Primary plan)*

If the results of our simulations give a disappointing picture — each column has different estimators, and there is no commonality between them — we structure our simulation-based first on overall effect size because the primary usage of meta-analysis is to determine whether an effect exists. The strategy is to select the best method instead of the best estimators. We have four different methods: two methods of estimating the overall effect size ((b) method indicates an unweighted method and (c) method indicates inverse variance method). We have two different types of estimations of within-study variance: individual effect size of 12 estimators and the overall effect size of 12 estimators, see the table below.

	(b) Method	(c) Method
s^2_{group1}	Group 1	Group 2
s^2_{group2}	Group 3	Group 4

s^2_{group1} *indicates within-study variance that has individual effect size term*

s^2_{group2} *indicates within-study variance that has the overall effect size term*

To determine which of these four methods should be chosen, we will start with the method that gives acceptable Type I Error and power. Next, we will select the estimators in that method with less bias and less MSE of $\hat{\Delta}$. In the next step, our focus is based on proportion of zero. When there is no evidence of heterogeneity, which is likely, standard practice assumes homogeneity and applies a simpler fixed-effect meta-analysis. That is likely to yield false results (Kontopantelis et al., 2013). Finally, we focus on less bias, and less MSE of $\hat{\tau}^2$ will be selected. See Figure 2.2 for further explanation.

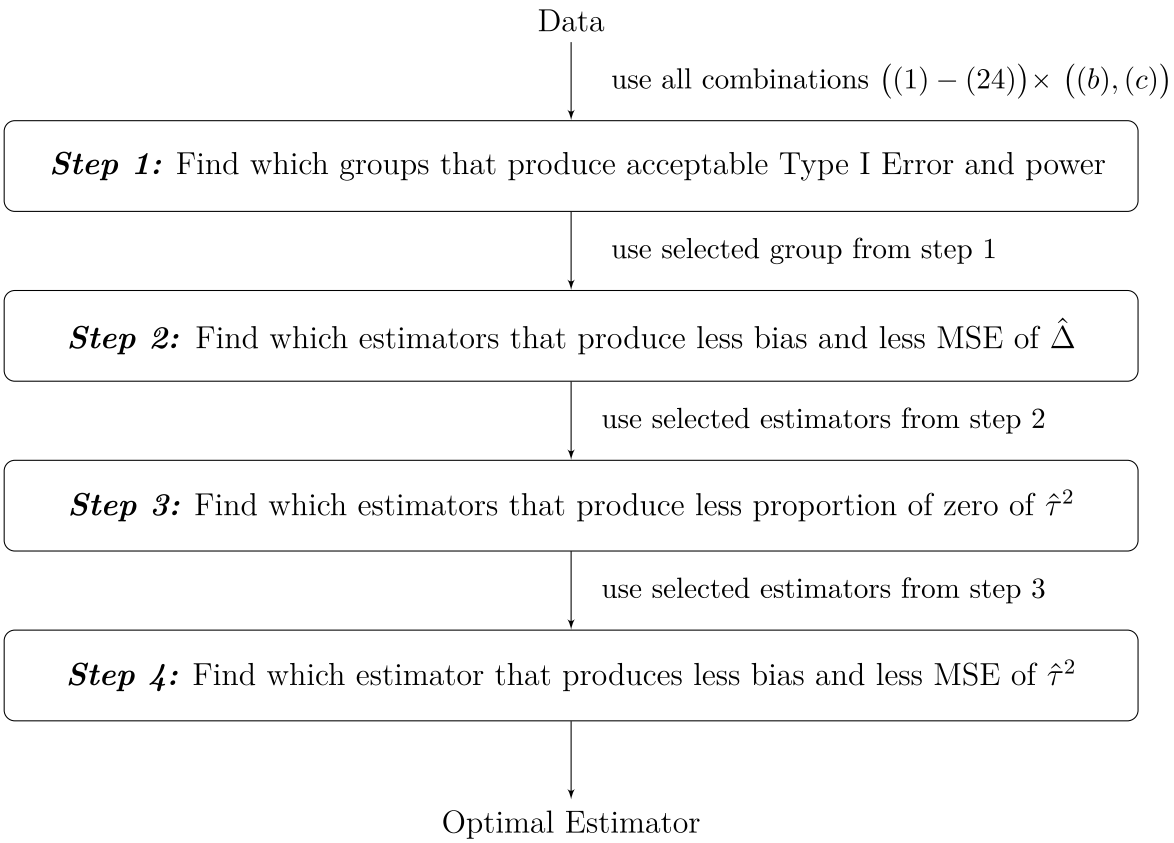

Data
use all combinations $\big((1)-(24)\big)\times\big((b),(c)\big)$
Step 1: Find which groups that produce acceptable Type I Error and power
use selected group from step 1
Step 2: Find which estimators that produce less bias and less MSE of $\hat{\Delta}$
use selected estimators from step 2
Step 3: Find which estimators that produce less proportion of zero of $\hat{\tau}^2$
use selected estimators from step 3
Step 4: Find which estimator that produces less bias and less MSE of $\hat{\tau}^2$
Optimal Estimator

Figure 2.2: *Structure of a data analysis (Alternative plan)*

2.5.4 Simulation Results

In all simulations, performance of ML, REML, and PM is assessed through not only the heterogeneity variance but also the effect on the overall effect estimate and Type 1 error. A summary of these methods is presented in Sections 2.5.4.1, 2.5.4.2, and 2.5.4.3. The comparison between these method is shown in Section 2.5.4.4.

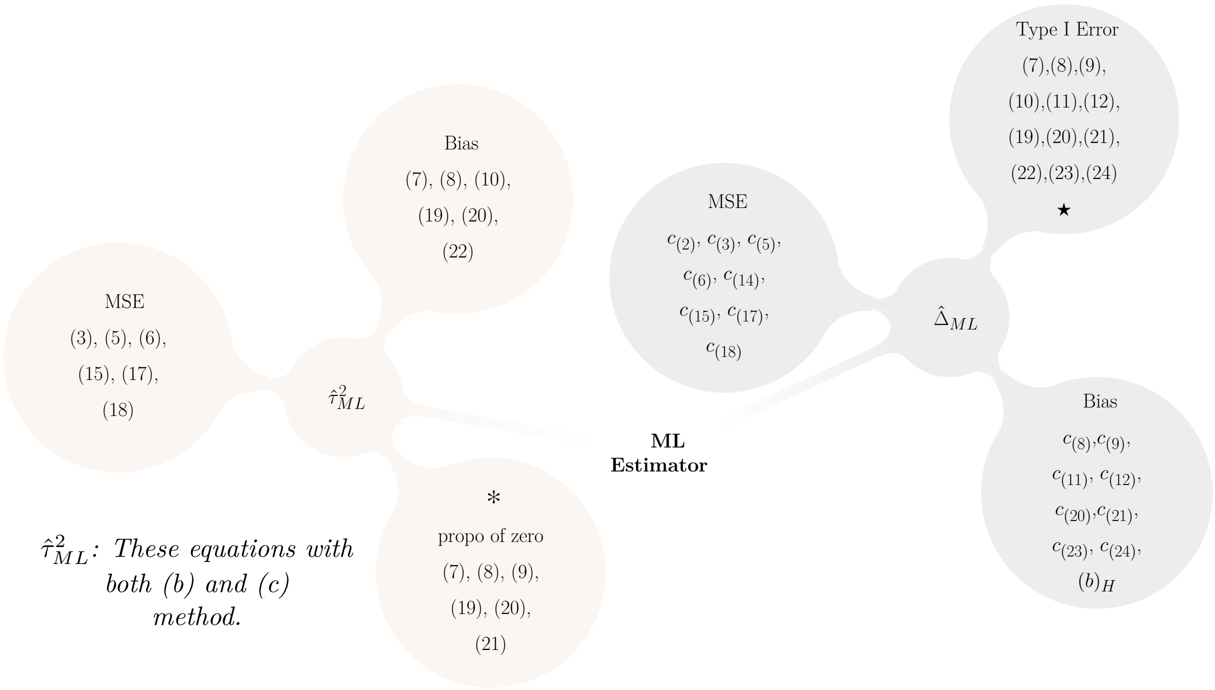
2.5.4.1 Properties of Between-Study Variance and The Overall Effect Size by Maximum Likelihood (ML)

For all combinations of the parameter value, we compare 48 heterogeneity variance estimators (τ^2_{ML}) in terms of bias, MSE, and proportion of zero estimates. See Section B.1.1. In general, the findings are consistent. First, τ^2_{ML} gives almost the same result by using either the (b) or

(c) method. Sometimes when testing small, but not necessarily equal, sample sizes with low to medium heterogeneity, the (b) method produces a slightly lower proportion of zeroes than the (c) method. On the other hand, the (c) method has less bias and less MSE in scenarios with medium effect size and low heterogeneity. Second, using s_j^2 that has individual effect size gives τ^2_{ML} less MSE but a larger proportion of zero. In terms of bias, the finding is inconsistent. Third, using Hedges' g estimators yields a smaller MSE than Cohen's d estimators, especially with a small sample size. Fourth, Hedges' g estimators ($s^2_{H.j}$) have less bias compared to Cohen's $d$estimators when bias is positive. However, when the bias is negative, Hedges' g estimators increase the bias. Thus, because most often τ^2_{ML} produces negative bias, especially with a high value of heterogeneity, Cohen's d estimators work better in estimating τ^2_{ML}. We aim to select the best 12 estimators from 48 estimators of τ^2_{ML}. Therefore, for each parameter combination, we mark all 12 estimators that produce smaller MSE, smaller bias, and a smaller proportion of zeroes (see Tables B.16, B.17, and B.18). The 12 columns with the most check marks will be selected as the best estimators. After running all simulations for each parameter combination, a summary of our recommendations is given in group below.

Regarding the the overall effect size, we compare the inverse variance (c) method and the unweighted (b) method. In general, we find when (c) uses s_j^2 that has the overall effect size term instead of individual effect size, $\hat{\Delta}_{ML}$ gives almost the same result provided by the (b) method. When both methods are estimated using Hedges' g estimator, they are almost unbiased with $\Delta = 0.5$. Also, using s_j^2 that has the overall effect size produces more MSE compared to using s_j^2 that has individual effect size, especially with small sample sizes. And Hedges' g estimator produces less bias and less MSE than Cohen d estimator. We aim to select the best six estimators from more than 26 estimators of $\hat{\Delta}_{ML}$. The same strategy for choosing the best between-study variance will be followed here for all combinations of the parameter value (see Tables B.19 and B.20). Sometimes more than six estimators are selected in one scenario multiple have the same result. The six columns with the most checks will be chosen as the best estimators. After running all simulations for each parameter combination, a summary of our recommendations is given in the group below

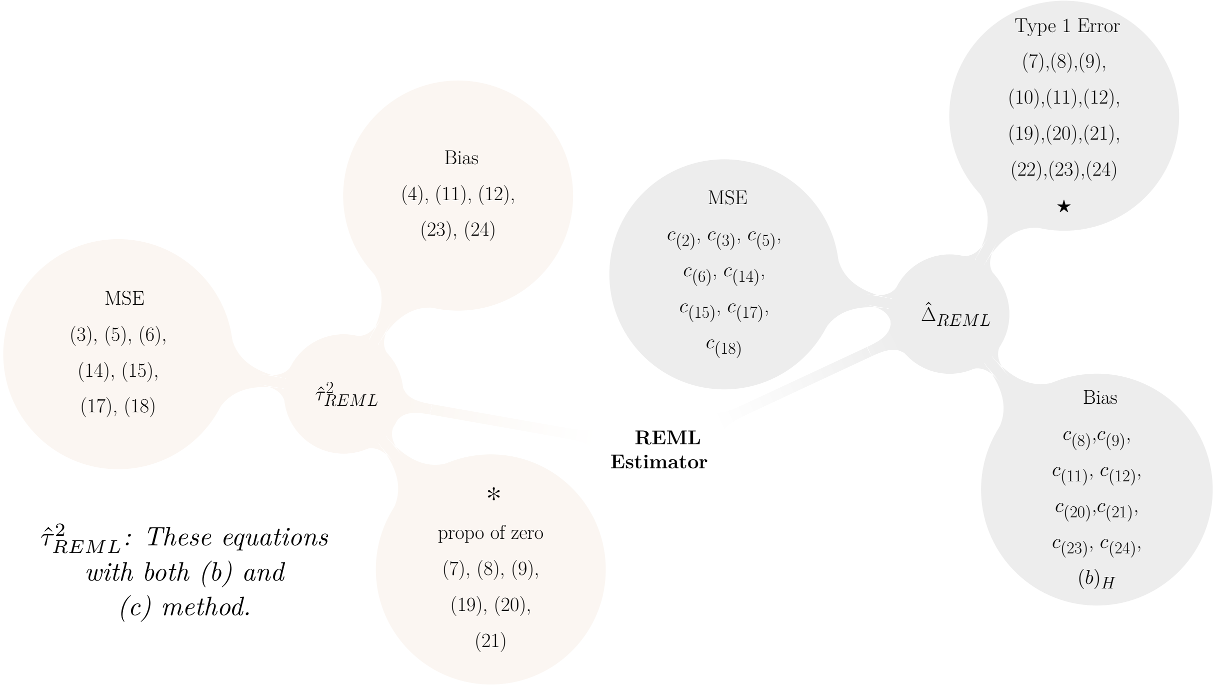
Concerning Type I Error, our simulations are based on t-distribution with Hartung-Knapp standard errors. The 12 estimators that provide acceptable Type I Error will be marked (see Table B.22.). Regarding power, there are no differences in results between all estimators except for the unweighted method, which gives high power. However, the unweighted method gives an inflated Type I error; they are removed from the comparison groups. Due to that, our comparison is based on the remaining, which all have the same result, so remove them from the comparison groups. Summary of best estimators in six groups by ML estimator is given



Here, the ★ indicates Type I Error from the Hartung-Knapp t-distribution (inverse variance method (c)) and $\circledast$ indicates proportion of zero. For details see Section B.1.2. In summary, we find that Estimator (8) and (20) are the best to use when estimating treatment heterogeneity using the ML method (see Table 2.2).

2.5.4.2 Properties of Between-Study Variance and Overall Effect Size by Restricted Maximum Likelihood (REML)

Our simulations for REML give a disappointing picture; we cannot follow the same strategy that was done for the ML estimator, see all the simulations in Section B.1.1. Tables B.2 and B.3 show the 12 estimators that produce less bias and less MSE. Similarly, Table B.4 shows the 12 estimators that produce less proportion of zero. Regarding overall effect size, see Tables B.5, B.6, B.7, and B.8 for bias, MSE, and Type I Error respectively. Regarding power, there are no differences in results between all estimators except for the unweighted method, which gives high power. However, the unweighted method gives an inflated Type I error; they are removed from the comparison groups. Due to that, our comparison is based on the remaining, which all have the same result, so remove them from the comparison groups. After running all simulations for each parameter combination, a summary of our recommendations is given below



$\hat{\tau}^2_{REML}$: These equations with both (b) and (c) method.

Again, the ★ indicates Type I Error from the Hartung-Knapp t-distribution (inverse variance method (c)) and ∗ indicates proportion of zero. For details see Section B.1.2. Given that our primary plan does not work perfectly, we move to the second one where we structure our simulation first on overall effect size because the primary usage of meta analysis is to determine whether an effect exists. The strategy of this simulation is to select the best method that gives an acceptable Type I Error instead of the most powerful or least biased estimators.

(1) Type I Error:

For each parameter combination, 48 Type I Errors by using t-distribution are compared. The best method will be selected based on 960 simulations

$$\overset{\overset{24((1)-(24))\times 2((b),(c))}{\nearrow}}{\textcircled{48}}(Type\ I\ error) \times 5(\tau^2) \times 4(N) = 960$$

We found that first, when the overall effect size variance uses within-study variance with individual effect size term instead of overall, Type I Error does not perform well in all scenarios under both the unweighted method (b) and inverse variance method (c). Second, with a large number of studies m and with medium $\tau^2 > 0.3$, (b) method produces a high Type I Error. It reached as high as 0.06 with small sample sizes, and 0.052 with large sample sizes, and dropped below 0.05. Third, with small m, it always performs better compared to others that use (b) method with a within-study variance that has overall effect size, and with (c) method with the either within-study variance that has individual or overall effect size. Because the (b) method performs poorly if we have a large number of studies, we will exclude this method from our comparisons. Concerning the (c) method and within-study variance with term individual effect size, Type I Error is always below 0.05, even with large m and N. Therefore, it also is excluded. On the other hand, when the variance of the overall effect size uses within-study variance that has overall effect size term, both methods (b) and (c) perform the same with equal sample size ($n_{1j} = n_{2j}$)), whether we have small or large n_{ij}. However, when sample sizes are not necessarily equal, the (c) method performs better with a small sample size and the (b) method performs slightly better with a large sample size. Thus, generally, using the (c) method that uses within-study variance with overall effect size term is recommended, which corresponds to the following estimators:

$$(7), (8), (9), (10), (11), (12), (19), (20), (21), (22), (23),\ \mathit{and}\ (24)$$

Regarding Type I Error using Hartung-Knapp standard errors, we compare 24 estimators using (c) method and omitting (b) method. The best method will be selected based on 480 simulations

$$\overset{24\ ((1)-(24))\times\ 1\ (c)}{\enclose{circle}{24}}\ (Type\ 1\ error) \times 5\ (\tau^2) \times 4\ (N) = 480$$

After running all simulations for each parameter combination, we achieved the same result from t-distribution using (c) method with within-study variance as outlined below. That is, the best estimators are:

$$(7), (8), (9), (10), (11), (12), (19), (20), (21), (22), (23),\ \mathit{and}\ (24)$$

(2) Overall Effect Size :

From Step 1, we selected the (c) method with within-study variance that has overall effect size, as define below:

$$(7)(c),\ (8)(c),\ (9)(c),\ (10)(c),\ (11)(c),\ (12)(c)$$

$$(19)(c),\ (20)(c),\ (21)(c),\ (22)(c),\ (23)(c),\ \mathit{and}\ \ (24)(c)$$

Next, we compare these 12 estimators as follows $(7)(c)$ $v.s$ $(8)(c)$ $v.s$ $(9)(c)$ $v.s$ $(10)(c)$ and so on. Our comparison is as shown.

$$\frac{\sum_{j=1}^{m} \hat{\delta}_{C.j}/\Big(\overset{(c)}{\enclose{circle}{\hat{\tau}^2_{REML}}} + \overset{(7)}{\enclose{circle}{s^2_{C.j}}}\Big)}{\sum_{j=1}^{m} 1/\big(\hat{\tau}^2_{REML} + s^2_{C.j}\big)} \quad v.s \quad \frac{\sum_{j=1}^{m} \hat{\delta}_{H.j}/\Big(\overset{(c)}{\enclose{circle}{\hat{\tau}^2_{REML}}} + \overset{(8)}{\enclose{circle}{s^2_{H.j}}}\Big)}{\sum_{j=1}^{m} 1/\big(\hat{\tau}^2_{REML} + s^2_{H.j}\big)} \quad v.s\ \ \ldots. \quad v.s\ldots$$

In general, Hedges' g estimators give the same result for both bias and MSE. When $\Delta = 0.5$, $\hat{\Delta}_{REML}$ is almost unbiased by Hedges' g estimators. Therefore, all estimators by Hedges' g are selected:

$$(8)(c),\ (9)(c),\ (11)(c),\ (12)(c)(20)(c),\ (21)(c),\ (23)(c),\ and\ \ (24)(c)$$

(3) Between-Study Variance

The eight estimators from Step 2 will be compared in term of heterogeneity variance estimators (bias, MSE, and proportion of zero estimates).

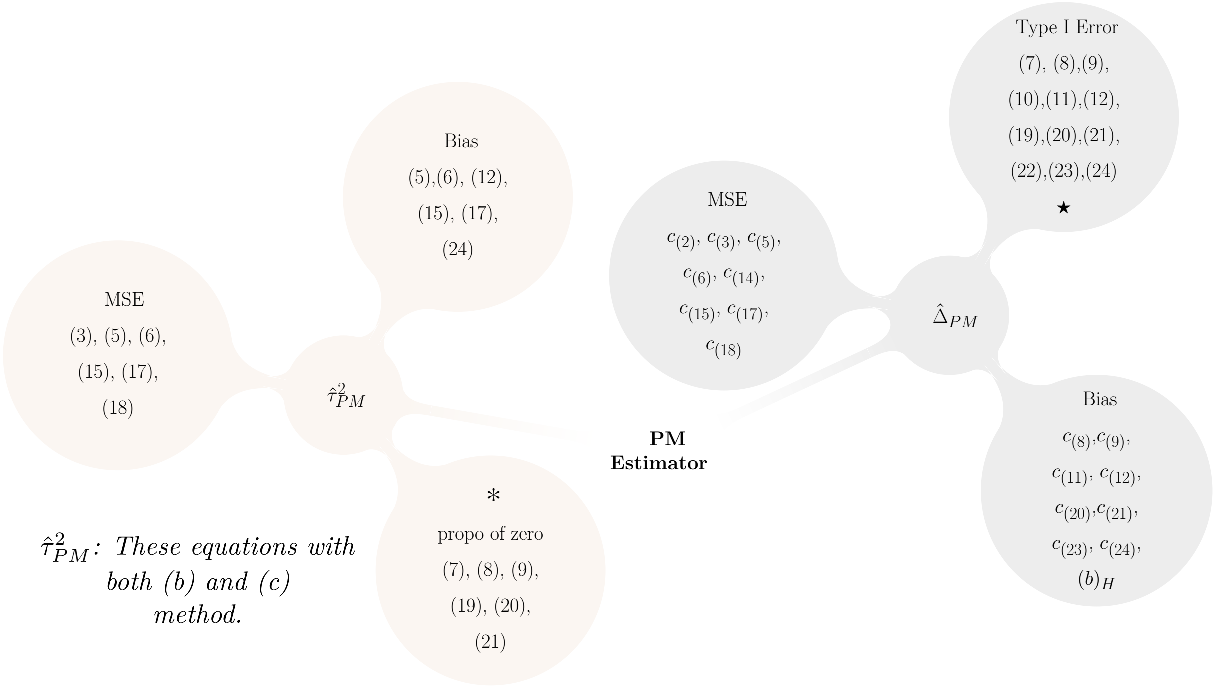
$$(8)(c)\ v.s\ (9)(c)\ v.s\ (11)(c)\ v.s\ (12)(c)\ v.s\ (20)(c)\ v.s\ (21)(c)\ v.s\ (23)(c)\ v.s\ (24)(c)$$

$$\frac{\sum_{j=1}^{m}\hat{w}_j{}^2((\hat{\delta}_{C.j}-\enclose{circle}{\hat{\Delta}_{REML}}\overset{(c)}{\nearrow})^2-\enclose{circle}{s^2_{C.j}}\overset{(8)}{\nearrow})}{\sum_{j=1}^{m}\hat{w}_j{}^2}+\frac{1}{\sum_{j=1}^{m}\hat{w}_j}\quad v.s\quad \frac{\sum_{j=1}^{m}\hat{w}_j{}^2((\hat{\delta}_{C.j}-\enclose{circle}{\hat{\Delta}_{REML}}\overset{(c)}{\nearrow})^2-\enclose{circle}{s^2_{C.j}}\overset{(9)}{\nearrow})}{\sum_{j=1}^{m}\hat{w}_j{}^2}+\frac{1}{\sum_{j=1}^{m}\hat{w}_j}$$

Because we aim to reduce the proportion of zeroes, we will start our comparisons under this criterion. We find that τ^2_{REML} produces a smaller proportion of zeroes by using either Estimator (8) or (20). Both estimators produce similar values for bias and MSE. That is, again, we find that Estimator (8) and (20) are the best to use when estimating treatment heterogeneity using the REML method.

2.5.4.3 Properties of Between-Study Variance and Overall Effect Size by Paule-Mandel (PM)

Our simulations for PM are disappointing; we cannot follow the same strategy that was done for the ML estimator. Tables B.9 and B.10 show the 12 estimators that produce less bias and less MSE. All the simulations can be found in Section B.1.1. Similarly, Table B.11 shows the 12 estimators that produce less proportion of zero. Regarding overall effect size, see Tables B.13, B.12, B.14, and B.15 for bias, MSE, and Type I Error. After running all simulations for each parameter combination, a summary of our recommendations is given below



Again, the ★ indicates Type I Error from the Hartung-Knapp t-distribution (inverse variance method (c)) and $\circledast$ indicates proportion of zero. For details see Section B.1.2. We follow the same strategy in the REML method to find the best estimator. However, because the performance of the PM estimator is nearly-identical to REML in term of the overall effect size (bias, MSE, and Type I Error) and between study variance (bias, MSE, and proportion of zero), we omit this here. We again find that Estimators (8) and (20) are recommended to use.

2.5.4.4 Discussion and Comparison

We find the following from our simulation study. First, simulations suggest using the alternative within-study variance Estimator (8) and (20) instead of the commonly used Estimator

(3) to have a better-performing heterogeneity variance estimators and the overall effect size. Second, although ML is identical with REML and PM under the overall effect size (in terms of bias, MSE, and Type I Error), we recommended against using ML because it is always produces a greater proportion of zeroes for $\hat{\tau}^2_{ML}$ compared to $\hat{\tau}^2_{REML}$, and $\hat{\tau}^2_{PM}$, and it has larger bias with large sample sizes). Third, the performance of the PM estimator is identical to REML in terms of the overall effect size (bias, MSE, and Type I Error) and between-study variance (bias, MSE, and proportion of zero), and both have overall good performance, with Estimator (8) and (20). Finally, Hartung-Knapp standard errors seem quite robust to inaccurate estimates of τ, and we recommend using these standard errors exclusively for random-effects meta analysis. Because the REML estimator is widely known, we will use it to compare it with best estimator under mixed model.

2.6 Simulation Studies on Mixed Model Approach

The same data sets that were generated in Section 2.5.1 will be used in this section. We analyze the data by using R via the `nlme` package (Pinheiro et al., 2017), `lme4` package (Bates et al., 2014), and `lmerTest` package (Kuznetsova et al., 2017). The performance of the the overall effect size will be assessed through their bias, MSE, and Type I Error. Hypothesis testing is conducted using Satterthswaite's degrees of freedom approximation, and is conducted using the R package `lmerTest` (Kuznetsova et al., 2017). For each simulation, we evaluate the performance of overall effect size of both ML and REML by using the *lmer* function. According to our simulations, REML and ML produce different Type I Errors (see Table 2.3).

For large sample sizes, the Type I Error of ML reaches 0.09 with a small number of studies ($m = 5$), and it never reaches below 0.06, even with large m and low τ^2. Thus, ML performs poorly because of inflated Type I Errors. However, REML provides an acceptable Type I Error. The range with small m is [0.05,0.055] and with large m is [0.0495,0.0505]. See Appendix B.2 for comparisons.

Table 2.3: *Performance of Overall Effect Size under Mixed Model: ML v.s REML*

Parameter			Overall Effect Size			
m	n_{ij}	τ^2	Bias	MSE	Type I Error under REML	Type I Error under ML
5	10	0.05	-0.0012	0.0516	0.0419	0.0522
		0.1	-8e-04	0.0620	0.0467	0.0619
		1.5	0.0033	0.3531	0.054	0.0882
	10-20	0.05	-0.0051	0.0387	0.0434	0.0578
		1.5	-0.0133	0.3355	0.0519	0.0852
10	10	0.05	-0.0005	0.0382	0.0394	0.0478
		0.1	-2e-04	0.0458	0.044	0.0560
		1.5	0.0026	0.2603	0.0523	0.0761
	10-20	0.05	-0.0015	0.0290	0.0449	0.0568
		1.5	-0.0029	0.2502	0.0508	0.0741
30	10	0.05	-0.0005	0.0283	0.0421	0.0485
		0.1	-3e-04	0.0339	0.0462	0.0553
		1.5	0.0015	0.1925	0.0513	0.0691
	10-20	0.05	-0.0007	0.0214	0.0461	0.0549
		1.5	-0.0014	0.1853	0.0507	0.0674

2.7 Comparison between Meta Analysis and Mixed Models

We used a simulation study under different scenarios of sample size setting rules, effect sizes, between-study heterogeneity, and number of studies to find the approach that performs best for analyzing multi-lab data. Based on results in Sections 2.6 and 2.5, REML for both meta-analyses and mixed models is generally indicated as having overall good performance. Therefore, REML under a random-effects meta-analysis will be compared with REML under a mixed model. For identical data, we make comparisons on the bias, MSE, and Type I Error in the overall effect size and based on bias, MSE, and the proportion of zero of between-study variance. The measure of between-study variance under mixed model is described as the standard deviations of environment-by-treatment interaction divided by the experimental error, for more details, see Section 2.2.

$$\hat{\tau}^2 = 2\,\frac{\hat{\sigma}_{\zeta}^2}{\hat{\sigma}_e^2} \tag{2.47}$$

From to our comparisons produced of 5,760 meta-analysis scenarios and 80 mixed model scenarios, we found that REML performs well under both meta-analysis and mixed model. The level of bias and MSE of the overall effect size when fitting a meta analysis is very similar when fitting a mixed model. However, based on between study variance, the mixed-model approach produces less bias and less MSE. We recommend analyzing the data of any multi-lab replication project using a meta-analysis for two reasons. First, the Type I Error under the mixed model is inflated slightly (reaches 0.054) when we have high heterogeneity. Second, the proportion of zero of τ^2 is a slightly higher than it is in meta-analysis (see Table 2.4).

When we have $\Delta = 0.5$, and we have equal sample sizes within lab, the level of bias of the overall effect size under meta analysis is smaller comparing to mixed model, and the MSE is higher than mixed model. With not necessary equal sample sizes, both bias and MSE under meta analysis is higher compared to mixed model. Note, the differences between these methods are very small, for example (when $m = 5$, $\tau^2 = 1.5$, and n_{ij}=10-20) bias and MSE of the overall effect size under meta analysis is -0.014 and 0.338 respectively, and bias and MSE of the overall effect size under mixed model is -0.011 and 0.336, respectively. Regarding between study variance, mixed model produces less bias and less MSE, yet produce higher proportion of zero compared to meta analysis with small τ^2. Overall, we find no substantial difference in the performance between the best-performing meta-analysis approaches and the mixed model approach for analyzing multi-lab data.

Table 2.4: *REML Method: Mixed model v.s Meta analysis with no effect*

			Mixed model						Meta analysis					
Parameter			Overall Effect size			Between-Study Variance			Overall Effect Size			Between-Study Variance		
m	n_{ij}	τ^2	Bias	MSE	Type I Error	Bias	MSE	Proportion of zero	Bias	MSE	Type I Error	Bias	MSE	Proportion of zero
5	10	0.05	-0.0012	0.0516	0.0419	0.047	0.026	0.46	-0.0011	0.0516	0.0499	0.059	0.033	0.437
		0.1	-8e-04	0.0620	0.0467	0.040	0.041	0.373	-7e-04	0.0622	0.0466	0.054	0.049	0.348
		0.3	3e-04	0.1036	0.0537	0.029	0.129	0.183	3e-04	0.1043	0.0495	0.047	0.148	0.173
		1.5	0.0033	0.3531	0.0545	0.029	1.557	0.024	0.0033	0.3568	0.0508	0.068	1.791	0.024
	10-20	0.05	-0.0051	0.0387	0.0434	0.027	0.014	0.425	-0.0052	0.0388	0.0498	0.033	0.016	0.413
		0.1	-0.0059	0.0490	0.0489	0.020	0.024	0.324	-0.0060	0.0492	0.0506	0.026	0.027	0.312
		0.3	-0.0079	0.0901	0.0543	0.008	0.095	0.135	-0.0080	0.0905	0.0497	0.017	0.102	0.130
		1.5	-0.0133	0.3355	0.0519	-0.004	1.363	0.015	-0.0138	0.3368	0.0504	0.016	1.487	0.014
	50	0.05	9e-04	0.0182	0.0562	0.004	0.004	0.225	9e-04	0.0182	0.0531	0.004	0.004	0.225
		1.5	0.0057	0.3118	0.0494	0.000	1.188	0.002	0.0057	0.3118	0.0492	0.010	1.244	0.002
	40-60	0.05	4e-04	0.0185	0.0532	0.003	0.004	0.218	4e-04	0.0185	0.0511	0.004	0.004	0.219
		1.5	0.0009	0.3123	0.0499	-0.006	1.182	0.002	0.0008	0.3129	0.0487	0.005	1.234	0.002
10	10	0.05	-0.0005	0.0382	0.0394	0.037	0.019	0.42	-0.0006	0.0384	0.0470	0.051	0.024	0.381
		0.1	-2e-04	0.0458	0.0442	0.029	0.029	0.311	-3e-04	0.0462	0.0457	0.045	0.036	0.281
		0.3	5e-04	0.0764	0.0500	0.020	0.094	0.121	4e-04	0.0774	0.0474	0.041	0.109	0.113
		1.5	0.0026	0.2603	0.0523	0.024	1.131	0.012	0.0024	0.2640	0.0498	0.068	1.305	0.012
	10-20	0.05	-0.0015	0.0290	0.0449	0.020	0.010	0.370	-0.0014	0.0292	0.0505	0.027	0.012	0.352
		0.1	-0.0016	0.0367	0.0506	0.014	0.018	0.254	-0.0015	0.0370	0.0521	0.022	0.020	0.240
		0.3	-0.0020	0.0673	0.0534	0.006	0.070	0.083	-0.0019	0.0679	0.0494	0.016	0.076	0.080
		1.5	-0.0029	0.2502	0.0508	0.001	0.995	0.008	-0.0027	0.2521	0.0500	0.023	1.088	0.007
	50	0.05	3e-04	0.0136	0.0518	0.002	0.003	0.156	3e-04	0.0136	0.0511	0.003	0.003	0.156
		1.5	0.0037	0.2337	0.0499	0.000	0.864	0.001	0.0035	0.2342	0.0502	0.009	0.901	0.001
	40-60	0.05	9e-04	0.0138	0.0516	0.002	0.003	0.153	9e-04	0.0139	0.0510	0.003	0.003	0.154

(To be continued)

m	n_{ij}	τ^2	Bias	MSE	Type I Error	Bias	MSE	Proportion of zero	Bias	MSE	Type I Error	Bias	MSE	Proportion of zero
		1.5	0.0021	0.2335	0.0505	0.002	0.863	0.001	0.0021	0.2342	0.0498	0.010	0.895	0.001
30	10	0.05	-0.0005	0.0283	0.0421	0.027	0.014	0.35	-0.0005	0.0285	0.0484	0.043	0.018	0.303
		0.1	-3e-04	0.0339	0.0462	0.021	0.022	0.235	-3e-04	0.0343	0.0477	0.039	0.027	0.206
		0.3	2e-04	0.0565	0.0514	0.014	0.069	0.081	1e-04	0.0573	0.0500	0.038	0.081	0.076
		1.5	0.0015	0.1925	0.0513	0.018	0.826	0.008	0.0013	0.1955	0.0498	0.067	0.958	0.008
	10-20	0.05	-0.0007	0.0214	0.0461	0.015	0.007	0.294	-0.0006	0.0216	0.0510	0.023	0.009	0.272
		0.1	-0.0008	0.0271	0.0506	0.010	0.013	0.181	-0.0007	0.0274	0.0517	0.019	0.015	0.168
		0.3	-0.0010	0.0498	0.0526	0.004	0.051	0.055	-0.0009	0.0503	0.0500	0.016	0.056	0.053
		1.5	-0.0014	0.1852	0.0507	0.002	0.729	0.005	-0.0014	0.1868	0.0503	0.029	0.799	0.005
	50	0.05	1e-04	0.0100	0.0496	0.002	0.002	0.106	1e-04	0.0100	0.0490	0.002	0.002	0.105
		1.5	0.0020	0.1724	0.0492	0.002	0.633	0.001	0.0019	0.1728	0.0492	0.010	0.659	0.001
	40-60	0.05	5e-04	0.0102	0.0504	0.002	0.002	0.103	5e-04	0.0102	0.0502	0.003	0.002	0.104
		1.5	0.0018	0.1732	0.0510	0.002	0.630	0.001	0.0018	0.1738	0.0503	0.010	0.654	0.001

Partial of the comparison are shown here, see Appendix B for further simulations

Table 2.5: *REML Method: Mixed model v.s Meta analysis with medium effect*

			Meta analysis						Mixed Model					
Parameter			Overall Effect Size			Between-Study Variance			Overall Effect Size			Between-Study Variance		
m	n_{ij}	τ^2	Bias	MSE	Power	Bias	MSE	Proportion of zero	Bias	MSE	Power	Bias	MSE	Proportion of zero
5	10	0.05	-0.002	0.057	0.378	0.062	0.035	0.439	0.005	0.055	0.413	0.047	0.026	0.464
		1.5	0.002	0.363	0.100	0.070	1.807	0.026	0.007	0.355	0.103	0.029	1.557	0.024
	10-20	0.05	-0.005	0.041	0.487	0.034	0.017	0.417	-0.001	0.040	0.509	0.027	0.014	0.425
		1.5	-0.014	0.338	0.100	0.016	1.489	0.015	-0.011	0.336	0.104	-0.004	1.363	0.015
10	10	0.05	-0.001	0.042	0.585	0.053	0.026	0.386	0.004	0.040	0.610	0.037	0.019	0.416
		1.5	0.002	0.270	0.143	0.069	1.318	0.013	0.005	0.262	0.147	0.024	1.131	0.012
	10-20	0.05	-0.002	0.031	0.690	0.029	0.012	0.357	0.001	0.030	0.703	0.020	0.010	0.370
		1.5	-0.003	0.253	0.151	0.024	1.091	0.008	-0.001	0.251	0.153	0.001	0.995	0.008
30	10	0.05	-0.001	0.031	0.723	0.046	0.019	0.308	0.003	0.030	0.740	0.027	0.014	0.349
		1.5	0.001	0.201	0.266	0.069	0.967	0.009	0.004	0.193	0.274	0.018	0.826	0.008
	10-20	0.05	-0.001	0.023	0.794	0.024	0.009	0.276	0.001	0.022	0.802	0.015	0.007	0.295
		1.5	-0.002	0.188	0.281	0.030	0.802	0.005	0.000	0.186	0.285	0.002	0.729	0.005

Partial of the comparison are shown here, see Appendix B for further simulations

2.8 Conclusion

This chapter aims to find the best practices for analyzing data from a multi-lab experiment. Specifically, we evaluate the validity of random-effects meta-analysis methods for estimating the effect-size for multi-lab data and compare the best-performing meta-analysis estimators to estimates from linear mixed models.

The meta-analysis literature indicates some concerns about the accuracy of the estimation of the heterogeneity variance and the methods for estimating the overall effect size. Therefore, we reviewed all methods for heterogeneity variance estimation, and focused our study on the three estimators that were either widely known and used in applications (maximum likelihood [ML]) or were positively evaluated (restricted maximum likelihood [REML] and Paule-Mandel [PM]). Given that estimates of both the overall effect and between-study variance are affected by within-study variance (s_H^2), and because previous literature showed conflicting recommendations about the accuracy form for estimating it (researchers always use simpler forms to approximate it), we also evaluated different ways of estimating the within-study variance. We compared 24 different estimators of s_j^2 , and numerous simulations were conducted to find the best estimator that provides accurate estimates of the heterogeneity variance and then the overall effect size.

First of all, results from simulated multi-lab studies showed superiority of use the Hartung-Knapp standard errors. These standard errors are robust to inaccurate estimates of treatment-by-lab variability, ensuring Type I Errors close to the nominal $\alpha = 0.05$ significance level.

Additionally, results showed better performance for estimators of the within-study variance different from the most commonly used ones. In general, we found convincing patterns that suggest avoiding within-study variances that use individual effect size because the correlation between effect sizes, study variance, and within-study variances will increase the bias for both the overall effect size sizes and between-study variance. In addition, the proportion of zero between-study variance increases, and Type I Error of the overall effect size will never reach 0.05 even with a large sample size and a large number of studies, except with low heterogeneity (=0.05), large sample sizes, and a large number of studies. Our findings confirm that within-study variances that use the unweighted method by the Hedges' g estimator are recommended. Thus, researchers should be careful about which estimators of within-study variance are used.

Of the three methods for estimating the variance of treatment heterogeneity, we recommend avoiding the use of ML even though it is identical with REML and PM under the overall effect size in terms of bias, MSE, and Type I Error. Our goal is not only to find the best-performing estimator for the overall effect size but also to have an accurate estimator of the between-study variance. ML always produces a higher proportion of zero-estimates for this variance. The performance of the PM estimator is identical to REML in terms of the overall effect size and between-study variance, and both are indicated as having overall good performance.

When comparing the meta-analysis approaches to those from mixed model approaches, we found no substantial difference between the two approaches in terms of accuracy of estimation and MSE, and both have an overall good performance in all the considered scenarios. However, we recommend the random-effects meta-analysis approach because the mixed model technique has a slightly higher proportion of zero estimates for the lab-by-treatment variability and may have a slightly inflated Type I Error under certain scenarios.

Chapter 3

A Sensitivity-Analysis Approach for Analyzing Multi-Lab Experiments

3.1 Introduction

Results from a large percentage of experimental studies have not been able to be replicated by follow-up studies—a phenomenon known as the *replicability crisis* (Wasserstein et al., 2019). For example, a large survey of 1,500 scientists and found that 70% of those polled had been unable to replicate others' results, and 50% could not replicate even their own results (Baker, 2016). Factors that may lead to difficulty in replicating results of a study include substandard research practices implemented within the study; insufficient sample sizes to obtain requisite power in the original study; differing research environments, including differences in expertise, protocol, location factors, and weather; publication bias; p-hacking; inappropriate analyses given the design of the study (for example, analyzing a cluster-randomized experiment as a completely randomized experiment); and deliberate fraud (Bello and Renter, 2018; Boos and Stefanski, 2011; Chen et al., 2021; Diener and Biswas-Diener, 2016; Francis, 2012; Gibson, 2021; Higgins et al., 2021; Ioannidis et al., 2009; Lewandowsky and Oberauer, 2020; of Sciences et al., 2019; Shiffrin et al., 2018).

Multi-lab experiments have emerged as a critical tool in the fight against the replicability crisis. In *multi-lab experiments*, the same experiment is independently performed across multiple facilities, with each facility following the same protocol when performing the experiment. Multi-lab experiments have both been used to test the likelihood of a proposed scientific finding being replicated prior to publication (Jaljuli et al., 2023; Kafkafi et al., 2005) and to verify or question the validity of a previously-published scientific finding (Ebersole et al., 2016, 2020; Klein et al., 2014b, 2018; Nieuwland et al., 2018).

There is currently no consensus as far as best practices for analyzing data from multi-lab experiments. Often, differences in treatment effects across labs are treated as a random variable. Under this setting, both random-effects meta-analytic approaches—in which effect sizes from each study are computed and aggregated to obtain an estimate of the overall effect size—and linear mixed model approaches—in which individual-level data are used to obtain estimates of treatment effects and effect sizes—are commonly used to analyze multi-lab data. In Chapter 2, we conclude that there is little difference between these approaches—with meta-analytic approaches performing slightly better than mixed-model approaches, see Section 2.7. However, both methods suffer from the same pitfall; estimation of the variability of treatment effects—otherwise known as across-lab variability—is inherently unreliable, especially when the number of studies and/or the number of observations within each study is small.

To combat this issue, we develop a sensitivity analysis approach for analyzing multi-lab data. In our approach, rather than estimating the across-lab variability, we consider a range of plausible values for this variability, and perform inferences on treatment effects across all values in the plausible range. We implement this method for both meta-analytic and mixed-model approaches. We then perform a simulation study to show the efficacy of our method and apply our method on a recent multi-lab study—Many Labs 4 (Klein et al., 2022).

This chapter is organized as follows. Section 3.2 introduces the environmental effect ratio (EER) in both approaches: mixed model and meta-analysis. In Section 3.4, we conduct simulation studies to examine a range of potential concerns of between-study variance. Sensitive analysis is used to determine the heterogeneity level that achieves the significance of effect size estimates. In Section 3.5, we apply a sensitive analysis to Many Labs 4 project. We conclude the findings in Section 3.6.

3.2 Notation and Preliminaries

Suppose we consider a multi-lab experiment performed across m labs, numbered 1 through m, with each lab performing the experiment independently. Units in the study are assigned to one of two treatment conditions, Treatment 1 or Treatment 2. We assume the following commonly-used model of response, which allows for treatment effects to vary randomly across studies:

$$Y_{ijk} = \mu_i + \theta_j + \zeta_{ij} + \varepsilon_{ijk} \qquad i = 1,2 \;\;,\; j = 1,\ldots,m \;\;,\; k = 1,,\ldots,n_{ij} \tag{3.1}$$

where μ_i is mean of the i^{th} treatment; n_{ij} is the number of units assigned to treatment i in lab j; θ_j represents a random source of variability common to all observations in study j; ζ_{ij} are lab-by-treatment interaction terms that represent random sources of variability unique to each treatment; and ε_{ijk} is the experimental error. We make the following distributional assumptions on the random terms:

$$\theta_j \sim N(0,\sigma_\theta^2),\; \zeta_{ij} \sim N(0,\sigma_\zeta^2),\; \varepsilon_{ij} \sim N(0,\sigma_e^2), \tag{3.2}$$

where these random variables are assumed to be mutually independent. Of note, we assume that the variance of the experimental errors σ_e^2 is the same across labs. Finally, determining that there is a significant treatment effect amounts to rejecting a null hypothesis $H_0 : \mu_1 = \mu_2$, or equivalently, rejecting a null hypothesis of a zero-valued *effect size* Δ:

$$H_0 : \Delta \equiv \frac{\mu_1 - \mu_2}{\sigma_e} = 0. \tag{3.3}$$

Under this model of response, most methods for estimating treatment effects or effect sizes from multi-lab experiments require estimation of the variance of the lab-by-treatment interaction σ^2_ζ. There are a wide variety of estimating this variance, include maximum likelihood (ML) and restricted maximum likelihood (REML). However, these methods tend to have significant pitfalls when sample sizes and/or the number of labs tend to be small. Specifically, estimates of this variance tend to be inaccurate and have a non-negligible chance of being 0. Underestimation of this variance may lead to misleading claims about the strength of the effect of treatment. Hence, methods that circumvent this estimation—for example, sensitivity analysis—may be appealing in practice, especially for small multi-lab studies.

A useful measure for the across lab variability is the environmental effect ratio (EER)—which is the ratio of the standard deviations of environment-by-treatment interaction and the experimental error (Higgins et al., 2021):

$$EER^2 = \frac{\sigma^2_\zeta}{\sigma^2_e}. \tag{3.4}$$

We will show that considering different values for the across-lab variability is equivalent to varying EER^2 in our sensitivity analysis approach.

3.3 Sensitivity Analysis for Mixed Linear Models

Mixed linear models and random-effects meta-analyses are the two most common approaches for analyzing multi-lab studies in which random lab-by-treatment interaction is present. We begin by developing our sensitivity analysis approach for mixed linear models and derive a closed-form expression for the special case of balanced designs.

3.3.1 Mixed Linear Models with a Known EER

To begin, note that equation (3.1) can be expressed in terms of matrices:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}\ , \qquad \mathbf{e} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{3.5}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{j}_{n_1} & 0 \\ 0 & \mathbf{j}_{n_2} \end{bmatrix}. \tag{3.6}$$

Here, $n_i = \sum_{j=1}^{m} n_{ij}$, $\beta = (\mu_1, \mu_2)'$, and $\mathbf{j}_{n_i}$ is a vector of n_i ones.

The covariance of the error term $\mathbf{e}$ can be decomposed as follows:

$$\mathbf{\Sigma} = \sigma_\theta^2 \mathbf{M}_\theta + \sigma_\zeta^2 \mathbf{M}_\zeta + \sigma_e^2 \mathbf{I}_n, \tag{3.7}$$

where

$$\mathbf{M}_\theta = \begin{bmatrix} \mathbf{J}_{n_{11}} & 0 & \cdots & 0 & \mathbf{J}_{n_{11}n_{21}} & 0 & \cdots & 0 \\ 0 & \mathbf{J}_{n_{12}} & \cdots & 0 & 0 & \mathbf{J}_{n_{12}n_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{J}_{n_{1m}} & 0 & 0 & \cdots & \mathbf{J}_{n_{1m}n_{2m}} \\ \mathbf{J}_{n_{21}n_{11}} & 0 & \cdots & 0 & \mathbf{J}_{n_{21}} & 0 & \cdots & 0 \\ 0 & \mathbf{J}_{n_{22}n_{12}} & \cdots & 0 & 0 & \mathbf{J}_{n_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{J}_{n_{2m}n_{1m}} & 0 & 0 & \cdots & \mathbf{J}_{n_{2m}} \end{bmatrix}, \tag{3.8}$$

and

$$\mathbf{M}_\zeta = \begin{bmatrix} \mathbf{J}_{n_{11}} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{J}_{n_{12}} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{J}_{n_{1m}} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \mathbf{J}_{n_{21}} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \mathbf{J}_{n_{22}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \mathbf{J}_{n_{2m}} \end{bmatrix}. \tag{3.9}$$

Here, the matrix $\mathbf{I}_k$ is the $k \times k$ identity matrix, $\mathbf{J}_k$ is the $k \times k$ ones matrix, and $\mathbf{J}_{k\ell}$ is the $k \times \ell$ ones matrix.

An estimate of the covariance $\mathbf{\Sigma}$ can be obtained by estimating the individual variance components σ^2_θ, σ^2_ζ, and σ^2_e and plugging them into 3.7. However, as mentioned previously, estimation of the lab-by-treatment interaction variance σ^2_ζ can be quite unreliable the sample sizes and/or the number of studies is small.

However, if the EER is known, we can bypass estimation of σ^2_ζ. In which case, we case we can rewrite (3.7) as

$$\begin{aligned} \mathbf{\Sigma} &= \sigma^2_\theta \mathbf{M}_\theta + EER^2 \sigma^2_e \mathbf{M}_\zeta + \sigma^2_e \mathbf{I}_n \\ &= \sigma^2_\theta \mathbf{M}_\theta + (EER^2 + 1)\sigma^2_e \mathbf{M}_e. \end{aligned} \tag{3.10}$$

where $\mathbf{M}_e$ is defined as follows:

$$\mathbf{M}_{e,ij} = \begin{cases} 1, & i = j, \\ EER^2/(1 + EER^2) & \text{units } i,\, j \text{ are in the same lab and have the same treatment status,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

This amounts to supposing a compound symmetry structure on the experimental errors with correlation $EER^2/(1 + EER^2)$ within each combination of lab and treatment. Statistical software can effectively estimate σ^2_e under this correlation structure, and we can obtain our estimated covariance matrix

$$\widehat{\mathbf{\Sigma}} = \hat{\sigma}^2_\theta \mathbf{M}_\theta + (EER^2 + 1)\hat{\sigma}^2_e \mathbf{M}_e. \tag{3.12}$$

To test for the presence of treatment effects, we use

$$T = \frac{\mathbf{a}'\hat{\beta}}{\sqrt{\mathbf{a}' \left(\mathbf{X}' \widehat{\mathbf{\Sigma}} \mathbf{X} \right)^{-1} \mathbf{a}}} \tag{3.13}$$

as our test statistic, where $\mathbf{a} = (1, -1)'$ and

$$\hat{\beta} = \left(\mathbf{X}'\hat{\mathbf{\Sigma}}\mathbf{X}\right)^{-1} \mathbf{X}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{y}. \tag{3.14}$$

Inferences are then obtained on treatment effects through using by using a t-distribution.

The degrees of freedom for this t-distribution will depend on how the sensitivity analysis is interpreted. If the plugged-in value of EER is viewed as a fixed parameter that eliminates the need to estimate across-lab variability, then using degrees of freedom equal to $\sum_{i=1}^{2}\sum_{j=1}^{m} n_{ij} - m$ may be preferred. However, if the plugged-in value of EER is viewed as a potential estimate of EER, then it may be preferable to compute degrees of freedom assuming a random across-lab effect. This can be done, for example, using a Satterthwaite approximation (Satterthwaite, 1941). Of the two, the Satterthwaite approximation is, by far, the more conservative option.

3.3.2 EER in Meta-Analysis

To do a study more than once using the mixed-model approach requires a large cost and time investment (Koole and Lakens, 2012; Lundwall, 2019; of Sciences et al., 2019). Therefore, meta-analysis is appropriate to use when experiments are studying the same treatments (Higgins et al., 2021). Because in most practical applications not all of the data of each individual study is available (only the summary data is available; (Olkin and Sampson, 1998), our focus is convert the quantity $\sigma_\zeta^2/\sigma_e^2$ in mixed model to meta analysis. The quantity $\sigma_\zeta^2/\sigma_e^2$ can be found by the variance of the difference between sample means ($\approx 2\sigma_\zeta^2$ with large sample sizes) divides by the error variance (Higgins et al., 2021).

$$Var\left(\frac{\mu_{1j} - \mu_{2j}}{\sigma_e}\right) = Var\left(\frac{(\mu_1 + \theta_j + \zeta_{1j}) - (\mu_2 + \theta_j + \zeta_{2j})}{\sigma_e}\right) = \frac{2\sigma_\zeta^2}{\sigma_e^2} \tag{3.15}$$

On the other hand, the variance of Cohen's population effect size (Higgins and Thompson, 2002) is

$$Var(\delta_j) = Var\left(\frac{\mu_{1j} - \mu_{2j}}{\sigma_e}\right) = \tau^2 \tag{3.16}$$

Thus, $\sigma_\zeta^2/\sigma_e^2$ has a connection with Cohen's d

$$\tau^2 = 2\,\frac{\sigma_\zeta^2}{\sigma_e^2} = 2EER^2 \tag{3.17}$$

we should use τ instead of τ^2 since both τ and the estimated effect are measured on the same scale (Cairns and Prendergast, 2020; Higgins et al., 2021).

$$\tau = \sqrt{2}\ EER \tag{3.18}$$

3.3.2.1 *P-Value*

Under the random-effects model each effect is weighted as

$$\hat{\Delta}_H = \frac{\sum_{j=1}^{m} \hat{w}_j\ \hat{\delta}_{H.j}}{\sum_{j=1}^{m} \hat{w}_j} \tag{3.19}$$

where

$$\hat{w}_j = \frac{1}{(s_{H.j}^2 + \hat{\tau}^2)} \tag{3.20}$$

From equation 3.17, we know that $\tau^2 = 2EER^2$, so $\hat{\tau}^2$ will be replaced with $2EER^2$. The within study variance is estimated as

$$s_{H.j}^2 = J_j^2 \tilde{n}_j + \frac{\hat{\Delta}_H^2}{2(n_{1j} + n_{2j})} \tag{3.21}$$

and an estimate of the variance of the overall effect-size is given as

$$V(\hat{\Delta}_H) = \frac{1}{\sum_{j=1}^{m} \hat{w}_j}. \tag{3.22}$$

For making inferences about the overall effect size, we use

$$T_H = \frac{\hat{\Delta}_H}{\sqrt{V(\hat{\Delta}_H)}} \tag{3.23}$$

and compare T_H against a t-distribution with $m-1$ degrees of freedom.

Alternatively, instead of using $V(\hat{\Delta}_H)$, we can also consider the Hartung-Knapp estimate of the variance (Hartung, 1999):

$$V_{HK}(\hat{\Delta}_H) = \frac{\sum_{j=1}^{m} \hat{w}_j (\hat{\delta}_{H.j} - \hat{\Delta}_H)^2}{(m-1)\sum_{j=1}^{m} \hat{w}_j} \tag{3.24}$$

This leads to the following test statistic:

$$T_{HK} = \frac{\hat{\Delta}_H}{\sqrt{V_{HK}(\hat{\Delta}_H)}}. \tag{3.25}$$

Again, comparisons are made against a t-distribution with $m-1$ degrees of freedom.

3.4 Evaluating the Replicability Under Different Values of EER

To illustrate the sensitivity of the replicability to EER, we need to generate data with the same experiments in several randomly selected environments. We consider the data set with a range of five to 30 studies. Each study compares a treatment group with a control group. For each data set, we generate sample sizes for each group; n_{1j} observations were generated by using $\mu_1 + \zeta_{1j} + \theta_j + e_{1jk}$ and n_{2j} observations by using $\mu_{2j} + \zeta_{2j} + \theta_j + e_{2jk}$. The data of ζ_{ij}, θ_j and e_{ijk} are generated as independent and identically distributed with $\zeta_{ij} \sim N(0, \sigma^2_\zeta)$, $\theta_j \sim N(0, \sigma^2_\theta)$, and $e_{ijk} \sim N(0, \sigma^2_e)$, respectively, $i = 1, 2$ and j denotes number of studies. The value of σ^2_e was set to 1. The categorized values of of Δ is defined as zero, low, medium, and high magnitude, 0, 0.2, 0.5, 0.8 respectively (Cohen, 2013). The values of τ^2 $(= 2EER^2$

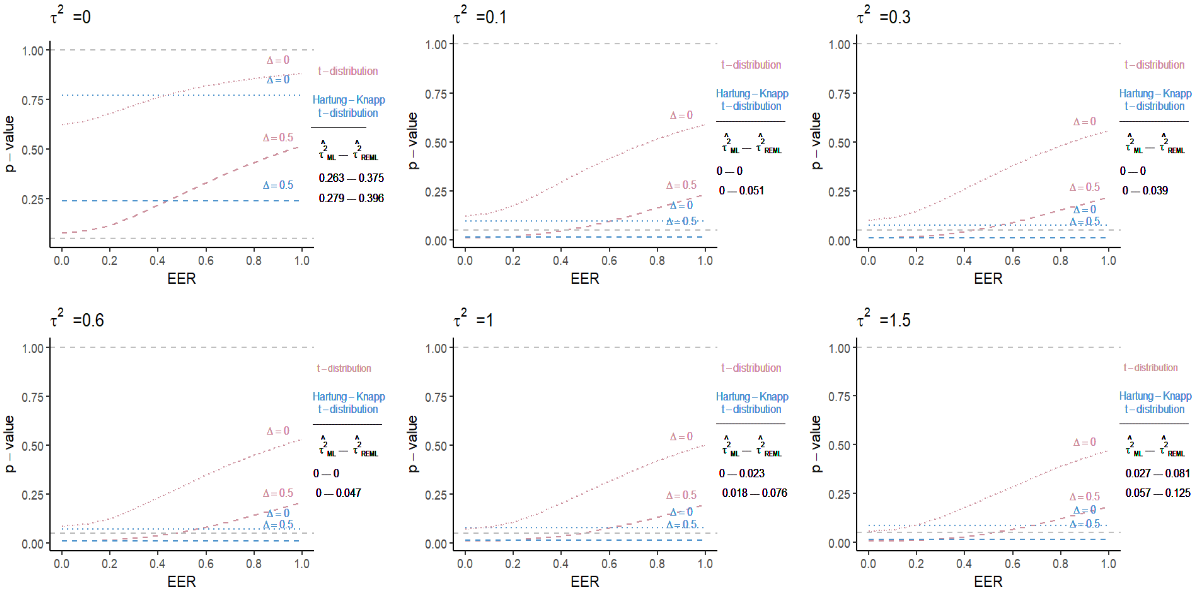
) is selected to be 0.05 (consider minimum), 0.1, 0.3, 0.6, and 1.5 (consider maximum). The study sample sizes n_{1j} and n_{2j} are generated in two ways: equal sample size $n_{1j} = n_{2j}$ (with 10 considered small and 50 considered large) and not equal sample size but in the same range (10 to 20 considered small and 40 to 60 considered large).

In Chapter 2, we conducted simulation studies under different scenarios to find the best estimators of two-stage meta-analysis and mixed model for analyzing the individual participants' data. We performed both fixed and random effects model using the REML method (see Chapter 2, Sections 2.5 and 2.6). However, according to our simulations in meta-analysis, we found that the better form for estimating within-study variance is by Hedges'g estimator $s^2_{H.j}$ which is as follows

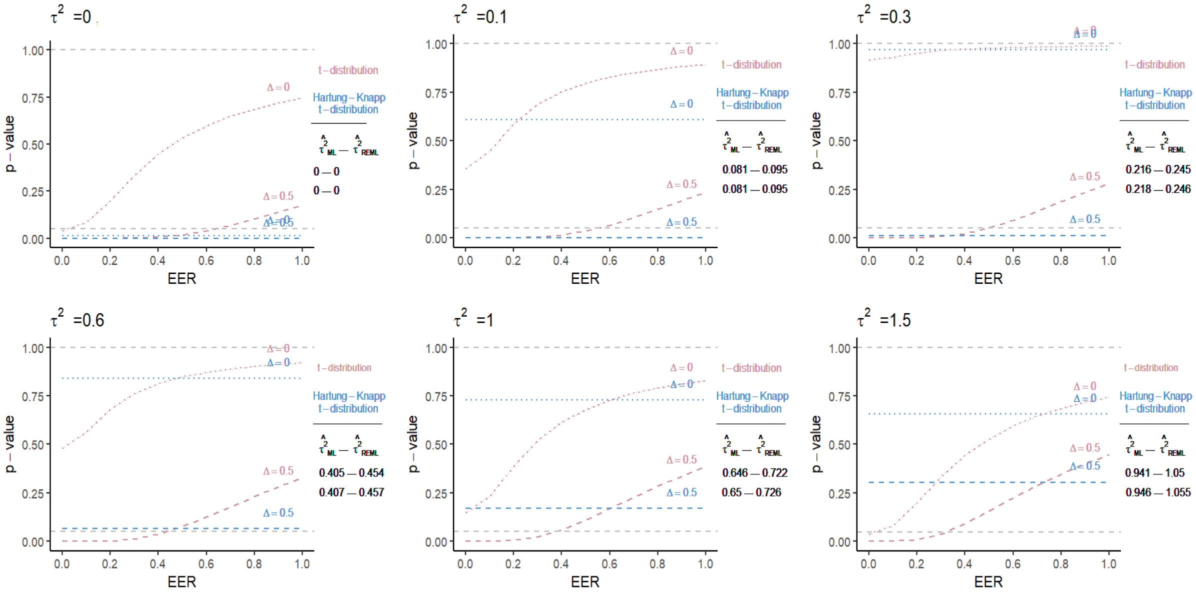
$$s^2_{H.j} = J^2_j \tilde{n}_j + \frac{\hat{\Delta}^2_{REML}}{2(n_{1j} + n_{2j})}.$$

3.4.1 Result and Discussion

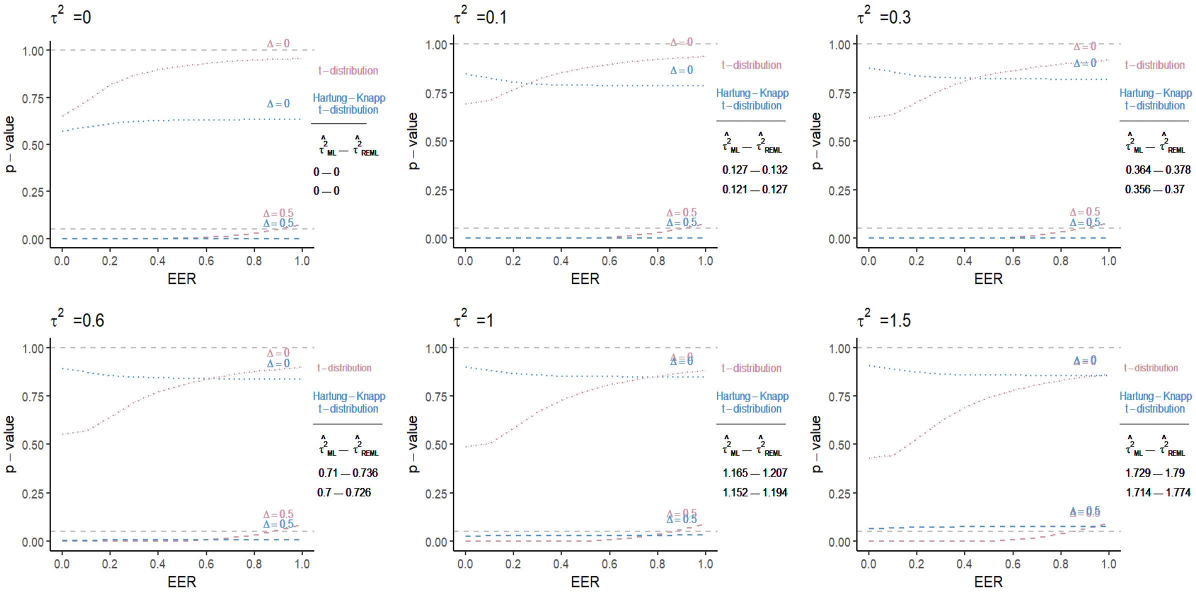
The generated data will be used under mixed model and meta-analysis. Under each parameter value combination, we first evaluate the performance of the estimation of between-study variance (REML and ML) in terms of giving false significant (nonsignificant) results. Second, we evaluate each method's performance in producing Type I Error. Third, the relationship between sample size, number of studies, heterogeneity, effect size, and statistically significant results are shown. We want to show which term plays an important role in replication and which does not have a significant effect. Fourth, we evaluate the work of sensitivity analysis by showing if it can help the researchers determine for what EER values the effect size estimate is statistically significant. All parameter value combinations can be found in the Appendix C. Subset of these simulations are given below. Figure 3.1 shows the comparison between t-distribution and Hartung-Knapp t-distribution under meta analysis, and Figure 3.2 shows the comparison under the mixed model.



m =5 , and n_{ij}=10

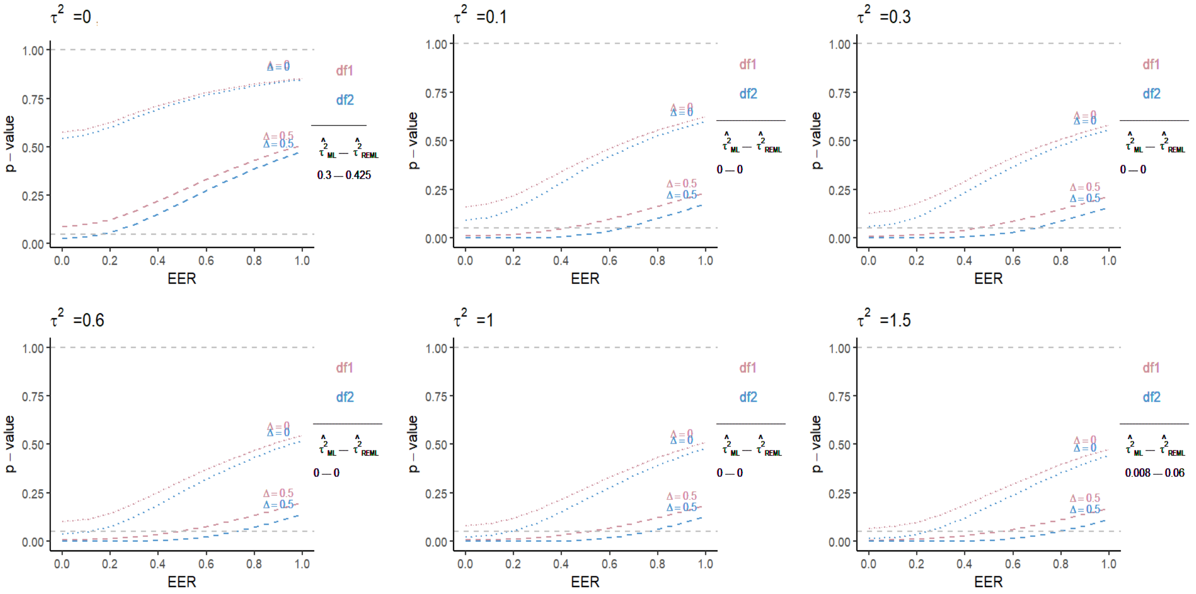


m =10 , and n_{ij}=50

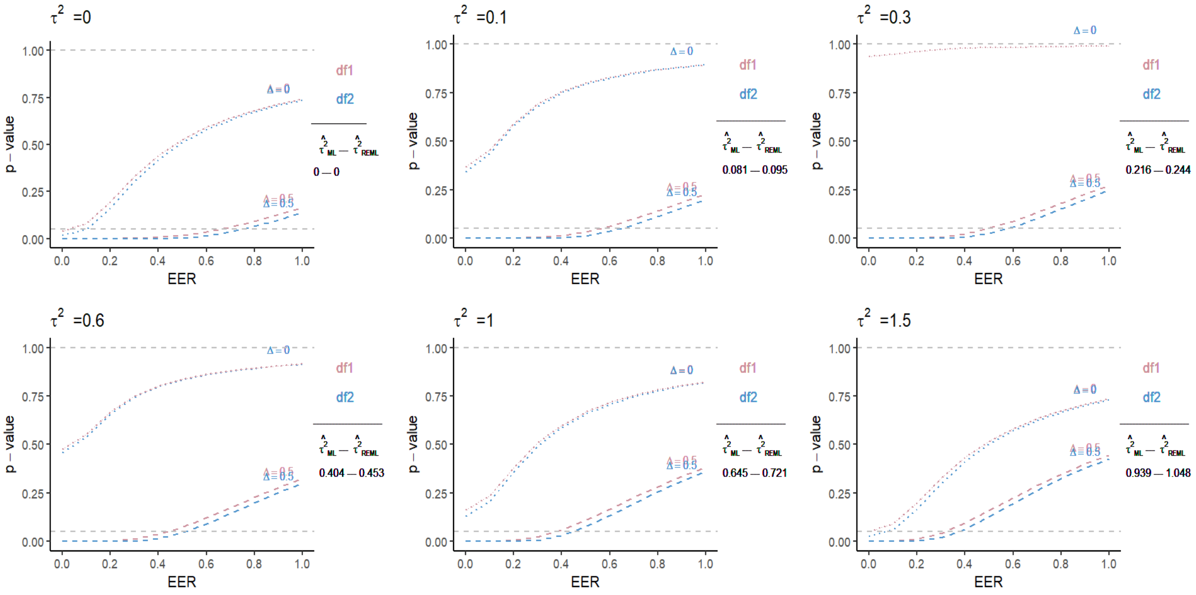


m =30 , and n_{ij}=40-60

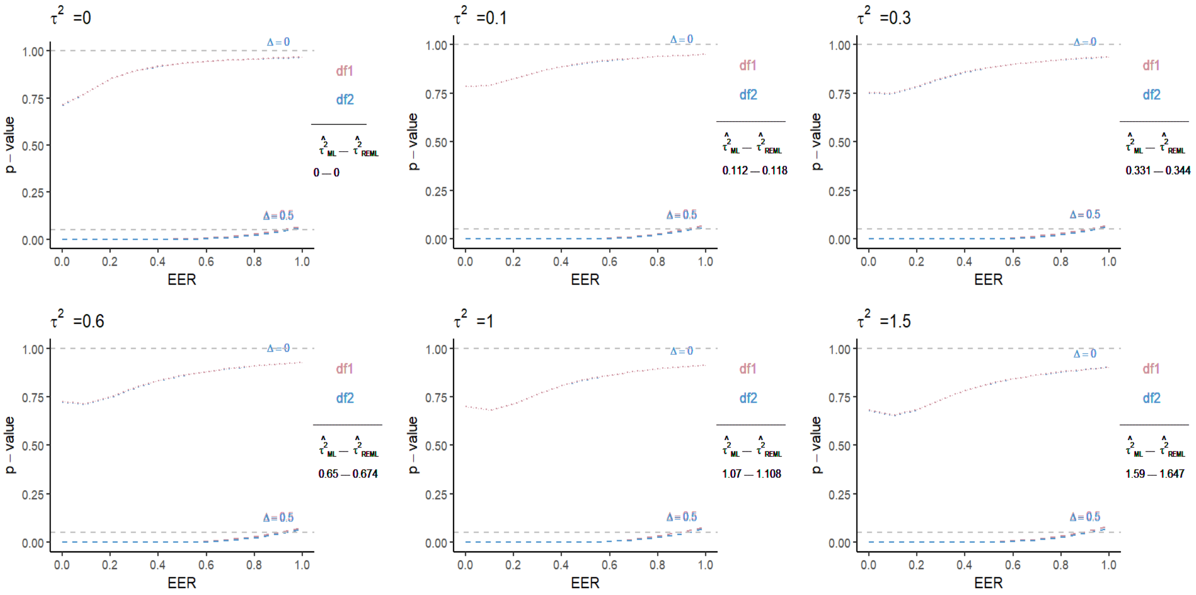
Figure 3.1: *Sensitivity analysis under meta analysis*



m =5 , and n_{ij}=10



m =10 , and n_{ij}=50



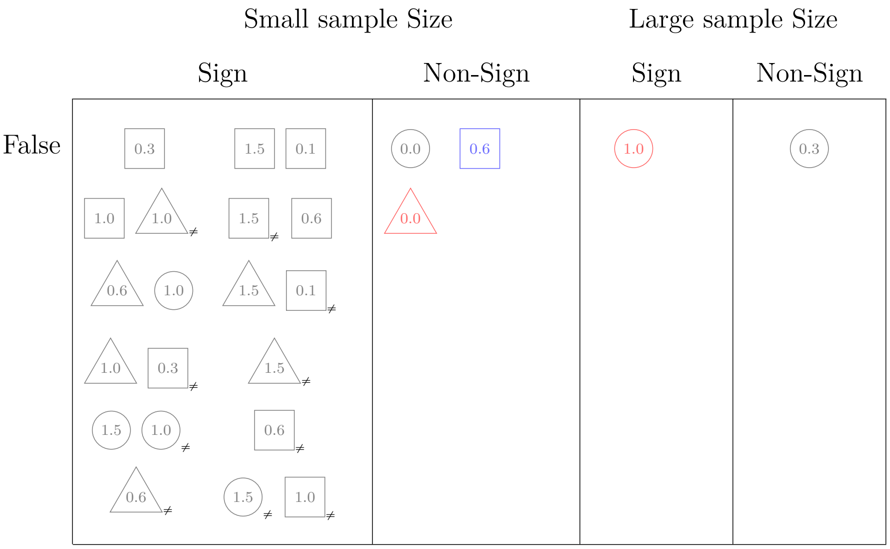
m =30 , and n_{ij}=40-60

Figure 3.2: *Sensitivity analysis under mixed model*

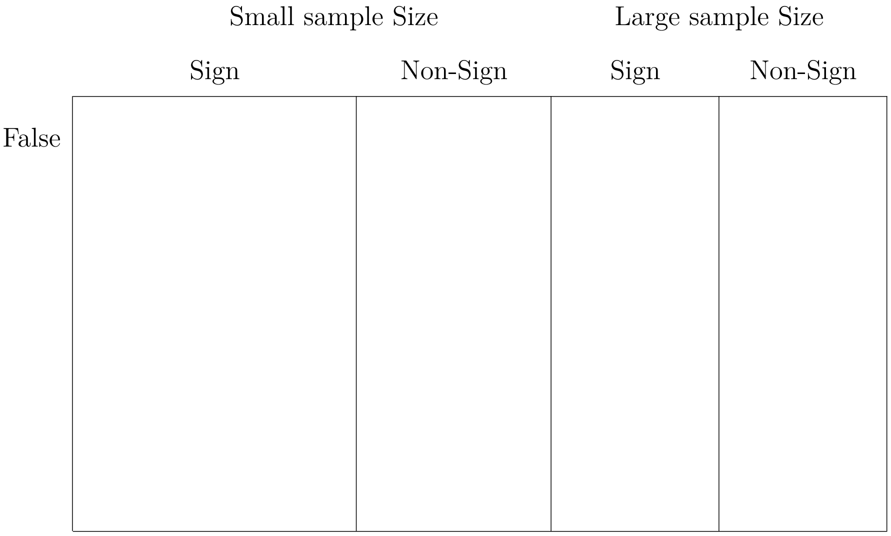
The x-axis for each plot is hypothetical values of the EER. The y-axis is the p-value when plugging in that hypothetical value of the EER into the estimator. Separate plots are given for data generated under various actual values of the τ, and plots also give estimates of this parameter under ML and REML methods for estimation. Note, p-values that are less than 0.05 when $\Delta = 0$ indicate a Type I Error and p-values that are greater than 0.05 when $\Delta = 0.5$ indicate a Type II Error.

These plots help determine for which values of EER the estimated effect size is significant, and can be used to judge how robust a result is to an increase in across-lab variability. However, when the true variability is large, but estimated to be close to 0 (which may happen in practice), results may look artificially more significant than they actually are. When sample sizes and/or the number of labs under study is small, the estimated across-lab variance tends to be much lower than the actual across-lab variance. Hence, both methods (mixed model and meta-analysis) under this setting are prone to report unreliable statistical inferences. The Hartung-Knapp standard errors (Hartung, 1999), however, tend to be robust across many hypothetical values of the EER. For more details see Appendix C.

We evaluate the performance of REML and ML under both methods. We find that p-values in the mixed model by Satterthwaite's approximation and meta-analysis using commonly used standard errors often have similar performances. Both methods produce a high rate of false significant findings when individual participant data has a small number of studies and small sample sizes—see Figures 3.3, and 3.4 (similarly with ML method). To mitigate the risk, we have two options: either we should have a medium-to-large number of studies (10 to 30), or we should use Hartung-Knapp standard errors, which reduce the chance of getting a falsely significant (or non-significant) result. When sample sizes are large, the probability of finding false results with either a small or large number of studies is negligible. Because in most practical applications, half or more of the meta-analysis often includes a small number of studies (Hedges and Vevea, 1998; Henmi and Copas, 2010; IntHout et al., 2015; Poole and Greenland, 1999; Seide et al., 2019), which have small sample sizes (Davey et al., 2011; IntHout et al., 2015), researchers should be cautious about published papers using meta-analysis. Note, high proportions of Type I Errors under meta-analysis has also been demonstrated by Borm and Donders (2009); Brok et al. (2008); Hu et al. (2007); Pereira and Ioannidis (2011); Thorlund et al. (2009, 2011); Wetterslev et al. (2008); Whitehead (1997).



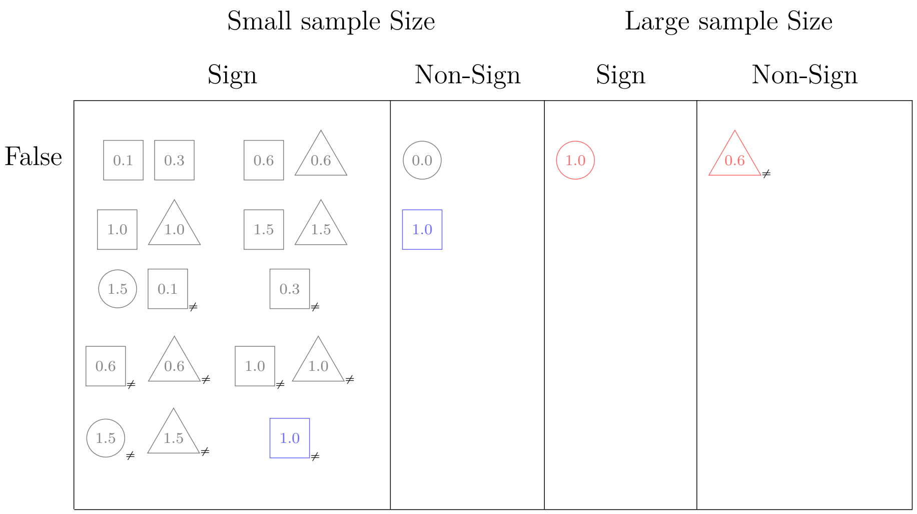
(a) **Meta-Analysis by t distribution**



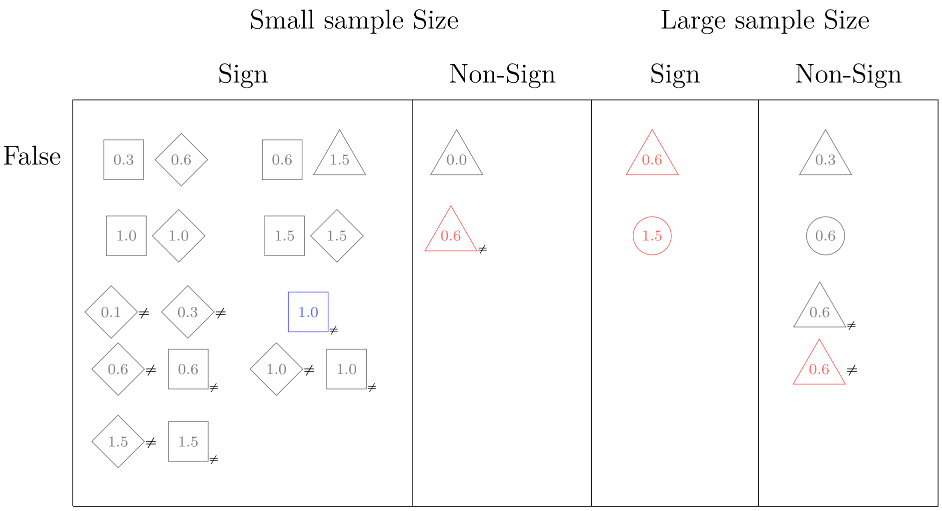
(b) **Meta-Analysis by Hartung-Knapp t distribution**

Figure 3.3: *Meta analysis: Evaluate the estimation of REML method*

Here, $\Delta = 0, 0.2, 0.5$ *and* 0.8 *indicates respectively as following* ◇, □, △, ○, *number inside the symbol indicates the value of* τ^2 *and sign* $\neq$ *indicates not necessary equal sample size.* ∘ *indicates small number of studies m=5,* ∘ *indicates medium number of studies m=10, and* ∘ *indicates large number of studies m=30. The Sign (Non-Sign) means the method estimates the between study variance and give significant result (non-significant result) wrongly.*



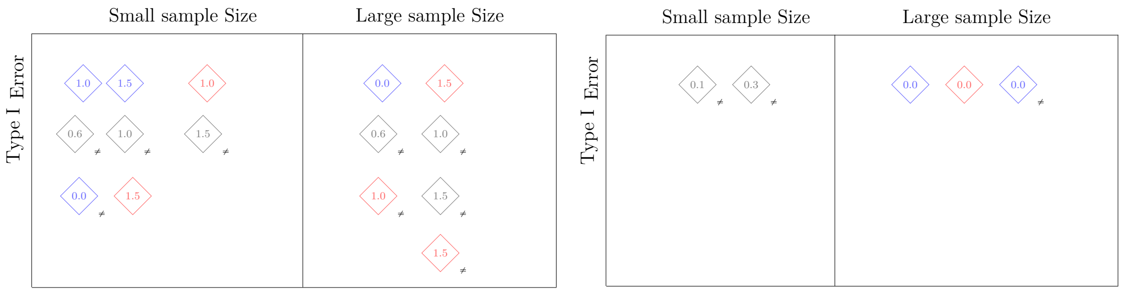
(a) Mixed Model by Satterthwaite approximation



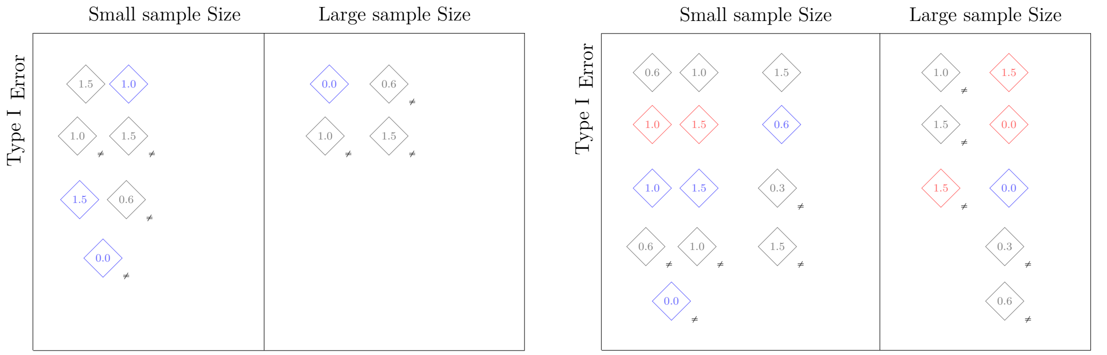
(b) Mixed Model by N-m-1 Degree of Freedom

Figure 3.4: *Mixed Model: Evaluate the estimation of REML method*

Here, $\Delta = 0, 0.2, 0.5$ *and* 0.8 *indicates respectively as following* ◇, □, △, ○, *number inside the symbol indicates the value of* τ^2 *and sign* $\neq$ *indicates not necessary equal sample size.* ∘ *indicates small number of studies m=5,* ∘ *indicates medium number of studies m=10, and* ∘ *indicates large number of studies m=30. The Sign (Non-Sign) means the method estimates the between study variance and give significant result (non-significant result) wrongly.*



(a) Meta-Analysis by t distribution **(b) Meta-Analysis by Hartung-Knapp t distribution**



(c) Mixed Model by Satterthwaite approximation **(d) Mixed Model by compound symmetry structure**

Figure 3.5: *Type I Error under meta analysis and mixed model*

The symbol ◇ indicates $\Delta = 0$*. Number inside the symbol indicates the value of* τ^2*, and sign* $\neq$ *indicates not necessary equal sample size.* ○ *indicates small number of studies m=5,* ○ *indicates medium number of studies m=10, and* ○ *indicates large number of studies m=30.*

Increasing the number of labs is much more effective than increasing the number of participants in each lab to improve power of tests for effect sizes. Even small sample sizes work better with medium and large numbers of studies. see Figures 3.6 and 3.7, and for more details see Appendix C. Hedges and Schauer (2019b) showed that the power usually rises more quickly if the number of labs is increased (at least 50 studies) instead of the participants per study (Hedges and Schauer, 2019b).

Finally, researchers should always be cautious about announcing getting significant results if their effect size from their study is small ($\leqslant 0.2$). This is especially the case with small sample sizes and a small number of studies because the estimation of between-study variance is inaccurate, which gives either zero or high which affect to change the result, except with

Hartung-Knapp method.

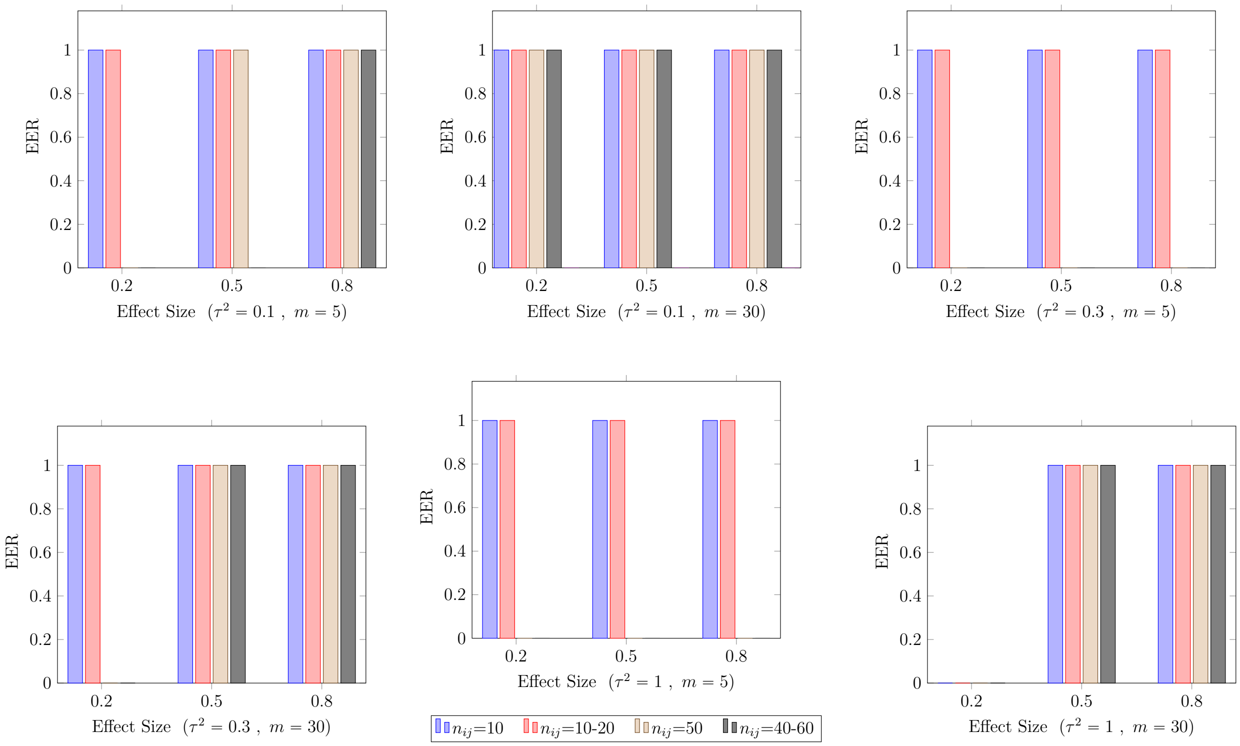


Figure 3.6: *Meta-Analysis by Hartung-Knapp t distribution under different scenarios. Each plot describes generated data with a specific level of heterogeneity ($\tau^2 = 0.1, 0.3,$ or 1), a specific number of studies (m=5 or 30), three various effect sizes (0.2, 0.5, and 0.8), and four sample sizes (n_{ij} = 10, 10-20, 50, and 40-60). Instead of estimating between-study variance, levels of EER between 0 and 1 are applied. Under meta-analysis $2EER^2$ is applied instead of τ^2. The bar in each plot measures the level of EER where the method gives significant results; thus, no bar means that even with EER=0, the method gives non-significant results.*

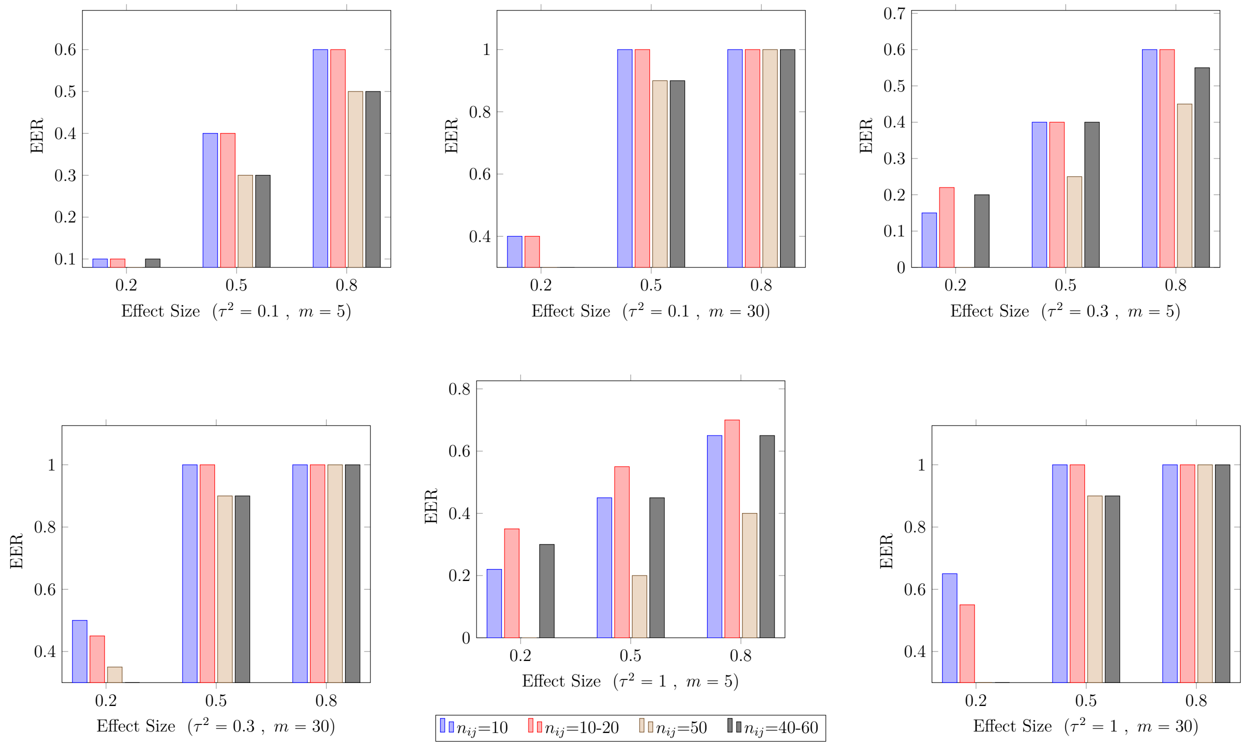


Figure 3.7: *Meta-Analysis by t distribution under different scenarios*

Each plot describes generated data with a specific level of heterogeneity ($\tau^2 = 0.1, 0.3$, or 1), a specific number of studies (m=5 or 30), three various effect sizes (0.2, 0.5, and 0.8), and four sample sizes (n_{ij} = 10, 10-20, 50, and 40-60). Instead of estimating between-study variance, levels of EER between 0 and 1 are applied. Under meta-analysis $2EER^2$ is applied instead of τ^2. The bar in each plot measures the level of EER where the method gives significant results; thus, no bar means that even with EER=0, the method gives non-significant results.

3.5 Application to Many Labs 4

Recently, the Many Labs 4 (ML4) project by Klein et al. (2019) used 21 labs to attempt to replicate an important effect in social psychology, the Mortality Salience Effect from Terror Management Theory by Greenberg et al. (1994). These labs were randomly assigned to either expert or in-house protocols. Researchers of the in-house labs constructed their own design independently without contacting anyone for advice—whether the original authors, other expert authors who are experts in this field, or ML4's teams. However, the researchers who worked in expert labs must follow the advice of TMT experts under three different exclusion sets. The main goal of the ML4 project was to determine whether the likelihood of successful replication increases when involving the original researchers in the study design.

Despite their efforts, the result did not replicate the under any conditions. Chatard et al. (2020) rebutted this finding and claimed that by imposing a strong restriction on the studies to be included in the analysis and by performing a meta-analysis on this subset of studies, the original result is replicated.

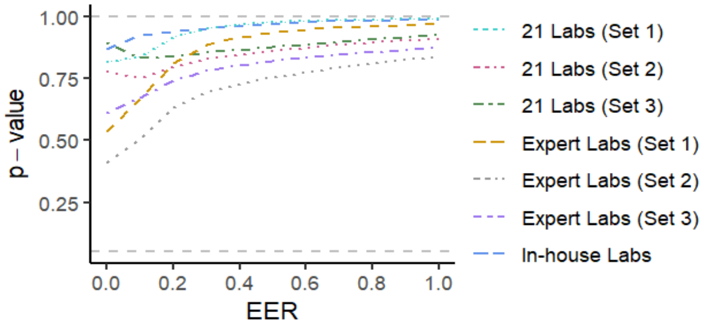
Klein et al. (2019, 2022), and Chatard et al. (2020) analyzed the data using random effects meta-analysis, which relies on estimating the variability in effect sizes across labs. The estimation of this parameter is typically unreliable and may lead them to a conflicting conclusion. Therefore, we will evaluate the replication of ML4 by applying a sensitivity-analysis approach across a range of values for the EER to determine the significance of effect size estimates for 21 labs. See Table 3.1 for the analyses performed.

Table 3.1: *Many Labs 4 under different analyses*

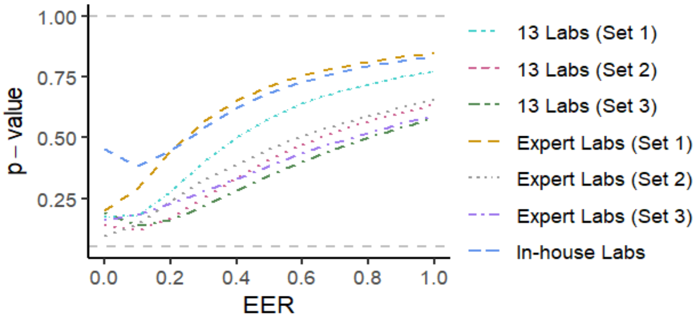
Key analysis by	Labs	Exclusion set	N-Criteria
Klein et al. (2019)	21 Labs	set 1- set 3	All
	Expert Labs (9)	set 1- set 3	All
	In-house Labs (13)	set 1- set 3	All
Chatard et al. (2020)	13 Labs	set 1- set 3	$N \geqslant 80$
	Expert Labs (6)	set 1- set 3	$N \geqslant 80$
	In-house Labs (7)	set 1- set 3	$N \geqslant 80$
Klein et al. (2022)	17 Labs	set 1- set 3	$N \geqslant 60$
	Expert Labs (7)	set 1- set 3	$N \geqslant 60$
	In-house Labs (10)	set 1- set 3	$N \geqslant 60$

Subgroups studied show no or small heterogeneity (EER between 0 and 0.18 by REML estimator), see Table 4.11. Given that the estimation could be inaccurate, sensitivity analysis on the EER will be applied. Under mixed model by compound symmetry structure, the data clearly provides evidence against a mortality salience effect, regardless of which labs (expert, in-house, or both) were used, which exclusion criteria were used, and whether they followed or deviated from the preregistration plan.

a) Key analysis by Klein et al. (2019)



b) Key analysis by Chatard et al. (2020)



c) Key analysis by Klein et al. (2022)

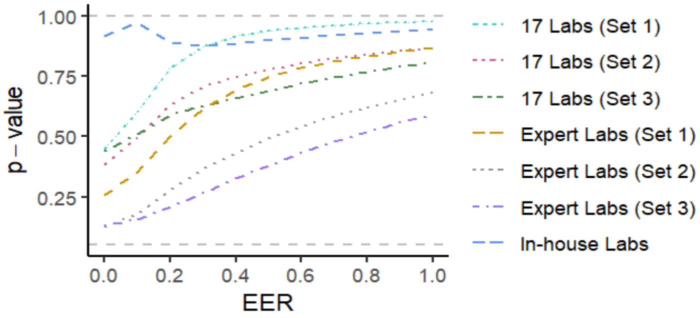
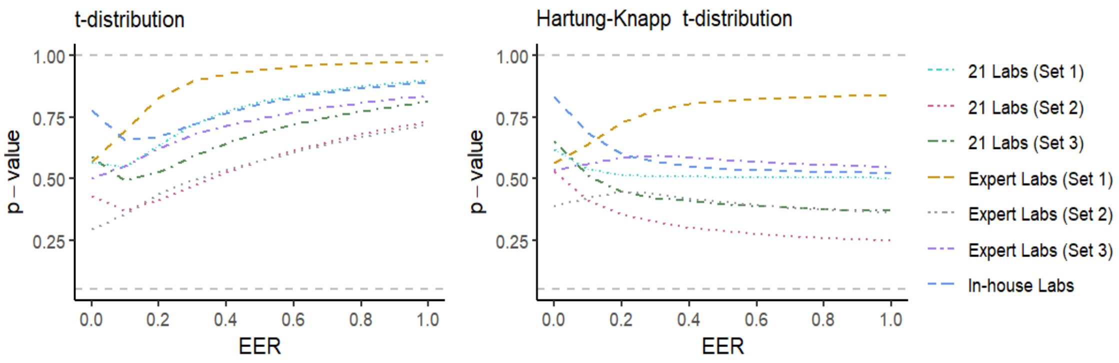


Figure 3.8: *Many Labs 4: Sensitivity analysis under mixed model*

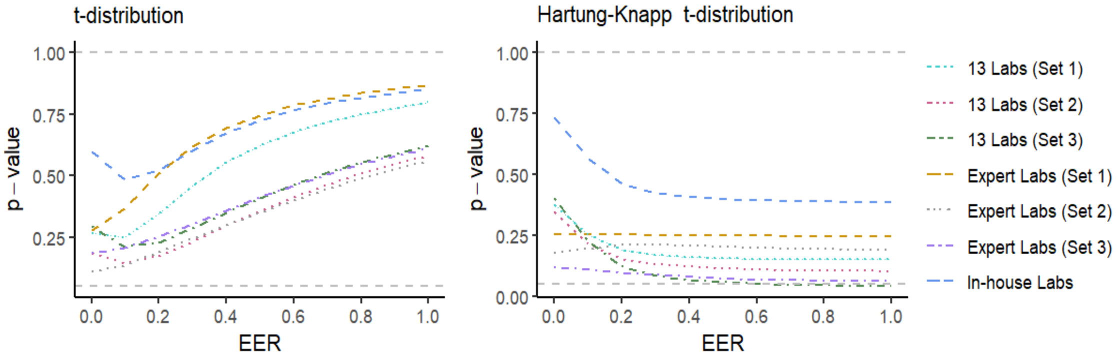
The random-effects meta-analysis technique by REML gives us a similar conclusion that has been found from the mixed model, see Figure 3.9. The mortality salience effect was non-significant even after considering the different heterogeneity values, except for 13 labs under Exclusion Set 3 and expert labs ($m = 7$) under Exclusion Set 3 using Hartung-Knapp standard errors, and this only becomes significant if the true EER is quite large (approximately 0.6 and 0.4 respectively, which is much less than the estimated EER).

Researchers may be pleased if they obtain significant results, even if their effect size is small. According to Cohen (2013), Chatard et al. (2020) stated that ML4 replicated their study with effect sizes between 0.1 and 0.27, considered "very low" and "low" respectively. However, they should be wary of announcing this result for several reasons: Their effect sizes are very low, their conclusion comes from a small number of studies that suffer from giving accurate results, and they estimate across-lab variability using ML estimators, which have received heavy criticism.

a) Key analysis by Klein et al. (2019)



b) Key analysis by Chatard et al. (2020)



c) Key analysis by Klein et al. (2022)

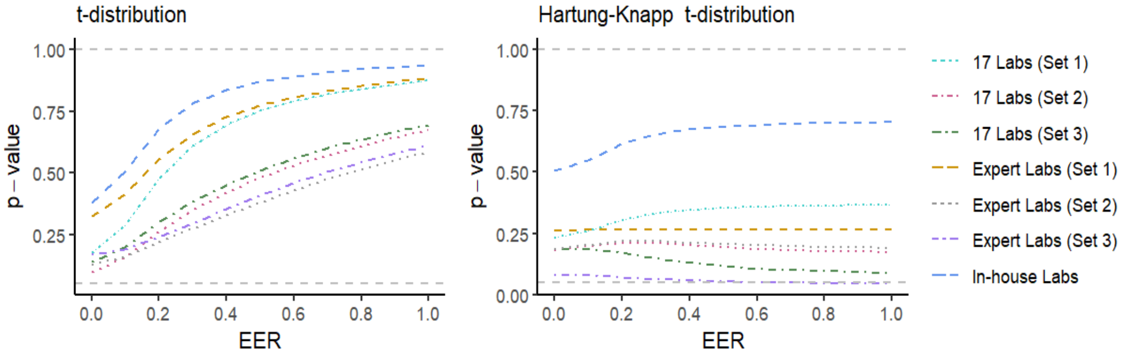


Figure 3.9: *Many Labs 4: Sensitive analysis under meta analysis*

Non-replication has been shown under both methods, even in carefully conducted exact replications, even with low heterogeneity, which is governing factor for successful replication.

3.6 Conclusion

When there is variability in the treatment effects across experiments, random effects meta-analysis and mixed models are often used to analyze multi-lab experiments. Because there is criticism about the estimation of between-study variance, we explore the use of sensitivity analysis procedures that consider a range potential values of this variability rather than estimate it. We develop these methods for both random effects meta analysis and mixed models.

Via simulation we show that the precision of the estimated between-study variance with a small number of studies and the small sample size is generally low, except the with the Hartung-Knapp method under meta-analysis. The probability of producing false significance (non-significance) is not negligible. In addition, mixed models and meta-analysis using more common standard errors give a high rate of false positives, even with large sample sizes.

In addition, we found that increasing the number of labs instead of the number of participants in each lab substantially increases the power to detect significant effect sizes. Lastly, researchers always should be wary about announcing significant results if the effect size from their study is small ($\Delta \leqslant 0.2$). This is especially true with small sample sizes and a small number of studies because the estimation of between-study variance is inaccurate—often either zero or high—which changes the result. Finally, while effective, we find that sensitivity analysis approaches can be misleading when the claimed or estimated across-lab variance is much lower than the actual across-lab variance, which can be possible when sample sizes and/or the number of labs under study is small.

Chapter 4

Re-examination of the Attempted Replication: Many Labs 4

4.1 Introduction

The *replicability crisis* refers to the phenomenon that many scientific findings cannot be replicated when independent follow-up studies attempt the same experiments that led to the original findings. For example, several hundred research groups in psychological science, which involved thousands of participants, tried to replicate the prior research findings (Stroebe, 2019). However, a substantial proportion of research has failed upon replication. There are myriad reasons why experiments may not replicate including lack of experience and expertise; substandard research practices; prohibitively small sample sizes; initial statistically significant results resulting from chance or that are only applicable to small subpopulations; different research environments; publication bias; *p*-hacking; and inappropriate analyses given the data (Bello and Renter, 2018; Boos and Stefanski, 2011; Chen et al., 2021; Diener and Biswas-Diener, 2016; Gibson, 2021; Higgins et al., 2021; Ioannidis et al., 2009; Lewandowsky and Oberauer, 2020; of Sciences et al., 2019; Shiffrin et al., 2018; Wasserstein et al., 2019).

Over the last decade, psychological science has trended towards team science, representing an important step forward for evaluating replicability (Hoogeveen et al., 2023). For example, the Many Labs projects (Ebersole et al., 2016, 2020; Klein et al., 2014b, 2018, 2022), the Open Science Collaboration project (Collaboration, 2015), and Registered Replication Reports (Simons et al., 2014) have all performed large replication studies involving many teams of scientists.

We focus our attention on a recent large-scale team science project involving 21 labs—Many Labs 4 (ML4) (Klein et al., 2022)—which attempted to replicate a study assessing impacts of the Mortality Salience Effect from Terror Management Theory (Greenberg et al., 1994). In particular, ML4 sought to replicate the finding from Study 1 in Greenberg et al. (1994) that critically thinking about one's own death can increase one's patriotic sentiment. Although Klein et al. (2022) attempted to increase the chance of replicating the result by implementing a multi-lab study and by involving the original authors in the study design, they failed to replicate the initial finding.

Of particular note, ML4 was very transparent about specifying their experimental protocol and making their data publicly available. The availability of ML4 data motivated many researchers to review and reanalyze the data. In particular, Chatard et al. (2020)—which includes Tom Pyszczynski as an author, who originally worked on the Greenberg et al. (1994) paper—rebutted against the finding in Klein et al. (2018), and argued that by imposing a strong restriction on the studies to be included in the analysis, and by performing a meta-analysis on this subset of studies, the original result is replicated. Because of these divergent findings, Hoogeveen et al. (2023) reanalyzed the data using different methodology, and was unable to replicate the original finding, corroborating the findings in Klein et al. (2018).

Because of the different conclusions in analyses from ML4, we perform our own reanalysis of these data. In our reanalysis, we investigate the quality of the ML4 data, the claims from the rebuttal in Chatard et al. (2020), and apply best practices for the analysis of multi-lab data using available meta-analysis and mixed-model techniques. Our analysis corroborates the original finding of (Klein et al., 2019) that the original experiment is unable to be replicated with any kind of consistency. Additionally, we identify issues with the analysis in the rebuttal by Chatard et al. (2020) that lead to inaccurate conclusions.

This chapter is organized as follows. Section 4.2 briefly describes Terror Management

Theory and the Mortality Salience Effect. In section 4.3, we detail ML4, its initial analysis, and the analysis of follow-up studies., Rebuttal ML4, and the follow-up studies. In section 4.4, we re-analyze ML4 using our methodology. We concludes our finding in Section 4.5.

4.2 Background of Terror Management Theory

Thoughts about death can create disquiet, impacting human thinking and behavior, which is the focus of terror management theory (TMT) research. TMT was introduced more than 35 years ago (Greenberg et al., 1986), and has been shown to affect ADD LIST Hundreds of publications have spawned after this topic; some with more than 1,000 citations (Klein et al., 2022). For an overview on TMT, see (Pyszczynski et al., 2015).

In particular, mortality salience hypotheses—psychological effects and behaviors of individuals when they are reminded of the inevitability of death—have been supported in more than 36 countries. Chatard et al. (2020) found that the effect of mortality salience on worldview defense (a classic finding from TMT) was replicated numerous times (e.g. in Arndt et al. (1997); Burke et al. (2010); Dechesne et al. (2003); Greenberg et al. (1994)).

In particular, a paper by Greenberg et al. (1994)—which has received more than 1,250 cites so far—performed a study (Study 1) to determine which types of mortality salience conditions had a stronger affect on worldview defense: subtle mortality salience inductions or blatant mortality salience inductions. The study was conducted as follows: a total of 21 male and 37 female students participated in this study, and groups of 3 to 5 participants were randomly randomly assigned to one of five conditions: subtle own death salient (thinking of your own death), subtle other's death salient (thinking of death of a loved one), deeper own death salient (thinking of your own death and imagining you having an advanced stage of cancer), deeper other's death salient (thinking of the death of a loved one and imagining that they have an advanced stage of cancer), and TV salient (thinking about watching TV, a control condition).

Participants participated in two "separate" studies. The first study varied by the treatment condition: participants wrote about their emotions felt when they thought about their treatment condition. For example, participants in the death salient condition wrote about what would happen to their real body as they were dying and once they were dead, and participants in the "TV salient" condition wrote about their emotions while watching television, and what they thought happened to their actual body as they watched television. The second study measured pro-American sentiment. Participants were asked to read a pro-American and an anti-American essay and asked to critique both essays and their authors. The response is a different in averages between sentiments for the pro-American and anti-American essays and authors. Results are summarized in Table 4.1.

Table 4.1: *Mean preference for the pro-USA target over the anti-USA target in Study 1 by Greenberg et al. (1994)*

Measure	Subtle own death salient	TV salient	Subtle other's death salient	Deeper own death salient	Deeper other's death salient
M	12.25	1.64	7.42	6.50	8.27
n	12	11	12	12	12

Study 1 of Greenberg et al. (1994) found that writing about your feelings when thinking about death increases your pro-American sentiment. The effect was strongest for the subtle own death salient treatment condition—comparing between this condition and the TV salient condition yielded a Cohen's $d = 1.34$, a t-statistic $t = 4.87$, and a p-value < 0.001.

4.3 Many Labs 4

Many Labs 4 (ML4) Klein et al. (2019) was a large, multi-lab study that sought to replicate Study 1 of Greenberg et al. (1994). To simplify the replication study, ML4 considered only comparisons between the subtle own death salient and the TV salient treatment conditions. In total, 21 labs participated in the replication study, and data were collected from a total of 2,228 participants.

One main goal of the ML4 project was to determine whether the likelihood of successful replication increases when involving the original researchers in the study design. Thus, the 21 labs were randomly assigned to either follow their own experimental protocol (in-house protocol) or to follow the protocol given by the original authors (expert protocol)—see Table D.1 for a detailed list of labs participating in the study. The in-house labs reconstructed the experiment in (Greenberg et al., 1994) following only the text of the original paper. This reconstruction was performed independently without seeking advice from anyone, including the original authors, other experts in the field, or other research teams participating in ML4, (see the instructions at `https://osf.io/drfg2/` for details).

Labs under the expert protocol condition followed the advice of TMT experts when reconstructing the experiment. Additionally, TMT experts required that data be analyzed under three increasingly-restrictive exclusion criteria. Exclusion Set 1 excluded all participants who did not complete both writing prompts and all six items evaluating the essay authors. Exclusion Set 2 excluded all participants that were excluded in Exclusion Set 1 and also all participants who did not identify as White or who indicated they were born outside the United States. Exclusion Set 3 excluded all participants in Exclusion Sets 1 and 2 and also excluded all participants who responded lower than 7 on the question "How important to you is your identity as an American?" Further details about these exclusion sets are found in Table 4.2. Note, analysis under these exclusion sets is not possible for in-house labs as these labs did not provide a questionnaire that would allow excluded participants to be identified.

We now detail four analyses of ML4. In Section 4.3.1 we describe the initial analysis by Klein et al. (2019). In Section 4.3.2 we consider the analysis in the rebuttal by Chatard et al. (2020). In Section 4.3.3, we detail the published analysis by Klein et al. (2022) and a follow-up analysis by Hoogeveen et al. (2023).

4.3.1 Many Labs 4 (Prior to Publication)

Klein et al. (2018) began their analysis by looking at treatment effects within each lab. Two sample t-tests were conducted to compare participants in the mortality salient (MS) condition with those in the TV salient condition. For expert protocol (or author-assisted (AA)) labs, analysis was conducted across all three exclusion sets. The study was interpreted as replicating if it yielded a significant effect ($p < 0.05$). Tables D.3 D.4 D.5 showed that all but one individual lab failed to replicate Study 1 under any of the three exclusion sets. The University of Illinois, which followed in-house (IH) protocol found a significant treatment effect (Hedges' $g = 0.74$, $p = 0.05$). The University of California at Riverside under Exclusion Set 2 nearly obtained a significant p-value (AA lab, Hedges' $g = 1.69$, $p \approx 0.05$) as well. However, there was very little consistency across labs—for example, under Exclusion Set 1, 14 of the 21 labs estimated a decrease in pro-American sentiment under the subtle mortality salient condition.

Next, random-effects meta-analysis with random terms estimated through the maximum likelihood method was used to estimate an overall effect size. Again, the study was interpreted as replicating if it yielded a significant effect ($p < 0.05$). Separate meta-analyses were conducted for the entire many-labs dataset, the in-house protocol labs, and the expert protocol labs. Results are found in Table 4.3. Neither the expert labs nor in-house labs successfully replicated Study 1 of Greenberg et al. (1994). Klein et al. (2019) stated that the failure could be due to the initial finding being a false positive, a gradual change in the population's pro-American sentiment over time, or necessary conditions that were present when the original experiment was conducted but were not fully understood or no longer exist.

Table 4.2: *Data exclusion as reported in preregistration plan*

Excluded by	Protocol	Exclusion
Klein et al. (2019)	Expert and In house	1) Excluded labs that has less than 60 participants
		2) Excluded all data that collected before the preregistration plan.
		3) Exclude participants who did not complete all six items evaluating the essay author
✠ TMT experts	Expert Labs	1) Exclusion Set 1: No Further Exclusions
		2) Exclusion Set 2: Excluded all participants who were not white Americans and were not born in the United States, including prior exclusion
		3) Exclusion Set 3: Excluded all participants who did not strongly identify with the US, including all prior exclusions as well

✠ *Exclusion Set 1 includes all exclusions considered for the in-house labs. Analysis under Exclusion Set 2 and 3 will only be applied to expert labs. In addition, within the expert labs(Exclusion Set 1, 2 and 3), participants must respond to both writing prompts to be included in the analyses.*

Table 4.3: *Random effects models as reported by a preliminary report of the ML4*

Labs	Exclusion Set	$\hat{\Delta}_H$	95% CI	p-value	$\hat{\tau}^2_{ML}$
All Labs (m=21)	Set 1	0.03	$[-0.06\ ,\ 0.12]$	0.58	0.003
	Set 2	0.06	$[-0.06\ ,\ 0.17]$	0.32	0.016
	Set 3	0.04	$[-0.07\ ,\ 0.16]$	0.48	0.01
Expert labs (m=9)	Set 1	0.043	$[-0.097,\ 0.18]$	0.55	0
	Set 2	0.13	$[-0.067,\ 0.33]$	0.2	0
	Set 3	0.11	$[-0.12\ ,\ 0.35]$	0.34	0
In-house labs (m=12)	All Sets	0.02	$[-0.1\ ,\ 0.15]$	0.72	0.01

Here, m indicates number of studies, $\hat{\Delta}_H$ is overall effect size by Hedges g estimator, and $\hat{\tau}^2_{ML}$ is between study variance.

4.3.2 Rebuttal Many Labs 4

When ML4 failed to replicate Study 1 from Greenberg et al. (1994), some of the researchers assisting the expert protocol labs were surprised (Chatard et al., 2020). As a result, Chatard et al. (2020) started examining the ML4 paper in detail, and looked at areas such as the data collection and the pre-registered plan. After review, they found that Klein et al. (2019) did not follow this pre-registered plan exactly, and some studies with small sample sizes were included in the analysis—the pre-analysis plan required all labs under study to obtain at least 60 participants. Chatard et al. (2020) reanalyzed the data under three different exclusions after excluding studies with small sample sizes. Their restriction on sample sizes was more stringent than indicated originally in the Klein et al. (2018) pre-analysis plan, requiring at least 80 participants in each lab. This left 13 studies to be included in their analysis (6 expert labs and 7 in-house labs). The total number of participants for each case becomes (as reported in their paper): Exclusion Set 1 ($N = 1782$), Exclusion Set 2 ($N = 1476$), and Exclusion Set 3 ($N = 1347$).

After Chatard et al. (2020) corrected the deviations as planned, they reanalyzed 13 of the 21 studies on one-tailed tests by random effect meta-analysis technique with maximum likelihood estimation, as was performed in the ML4 project. Chatard et al. (2020) found an evident replication of the original study, but this successful replication was only in the expert labs in Exclusion Sets 2 and 3 (see Table 4.4).

Table 4.4: *Random effects models as reported by Rebuttal-ML4*

Labs	Exclusion Set	$\hat{\Delta}_H$	90% CI	SE	Z	p-value
13 Labs	Set 1	0.06	[-0.03, 0.15]	0.05	1.17	0.13
	Set 2	0.13	[0.01, 0.26]	0.08	1.72	0.045
	Set 3	0.10	[-0.02, 0.22]	0.07	1.40	0.09
Expert Labs (m=6)	Set 1	0.10	[-0.04, 0.23]	0.08	1.21	0.12
	Set 2	0.27	[0.06, 0.48]	0.13	2.08	0.02
	Set 3	0.25	[0.02, 0.48]	0.14	1.76	0.04
In-House Labs (m=7)	All Sets	0.06	[-0.08, 0.19]	0.08	0.69	0.25

Here m indicates number of studies, $\hat{\Delta}_H$ indicates the overall effect size by Hedges g estimator, and p-values are one-tailed.

Chatard et al. (2020) argued that the original effect is robust among white American participants; the results indicated that TMT effect was successfully replicated when only white Americans were considered in the analysis (Exclusion Set 2) , despite effect sizes being much smaller than originally found in Study 1 in Greenberg et al. (1994). Moreover, the results indicated that the ML4 project was actually replicated when labs followed the advice by TMT experts (Expert Labs under Exclusion Set 2 and 3). Chatard et al. (2020) concluded that a replication study is most likely to succeed when researchers follow preregistered studies and follow the advice of researchers with considerable expertise and experience.

Additionally, Chatard et al. (2020) reanalyzed the data after including labs that collected at least 60 participants ($N \geqslant 60$). They obtained similar findings. Chatard et al. (2020) suggested that Klein et al. (2018) should closely follow their preregistered plans to avoid producing Type II Errors.

4.3.3 Follow-up Studies

Klein et al. (2022) stated that, when he posted a preprint on `https://psyarxiv.com/vef2c/`, he accidentally included studies with small sample sizes ($N < 60$). For that reason, Klein et al. (2022) reanalyzed the data and ensured strict adherence to their preregistration plan. In 2022, Klein et al. (2022) published their results by excluding the four labs that has less than 60 participants, and excluding some data from in-house labs that was collected before the analysis plan was preregistered (545 participants were excluded, see preregistration document on `https://osf.io/4xx6w`). Only 17 labs remained with a total 1,578 participants.

Again, aggregate data had been found, and p-value for each study was been calculated. Results are in Table D.7. Again, all individual labs fail to replicate Study 1 under three different exclusions, except one study (University of Illinois — IH). Then, separate meta-analyses were conducted for all 17 labs, the in-house labs only, and the expert labs only. Despite, strict adherence to their pre-registration plan, neither the expert labs nor in-house labs successfully replicated the original finding.

Table 4.5: *Random effects models as reported by ML4*

Labs	Execution Set	N	$\hat{\Delta}_H$	95% CI	p-value	$\hat{\tau}^2_{ML}$
17 Labs	Set 1	1550	0.07	$[-0.03, 0.17]$	0.187	0.001
	Set 2	1229	0.09	$[-0.03, 0.21]$	0.135	0.007
	Set 3	1076	0.09	$[-0.04, 0.22]$	0.193	0.01
Expert Labs (m=7)	Set 1	699	0.08	$[-0.07, 0.23]$	0.29	0
	Set 2	378	0.17	$[-0.04, 0.37]$	0.11	0
	Set 3	225	0.19	$[-0.07, 0.46]$	0.158	0
In-House Labs (m=10)	All Sets	851	0.05	$[-0.11, 0.22]$	0.5	0.02

Here, m indicates number of studies, $\hat{\Delta}_H$ indicates the overall effect size by Hedges g estimator , and $\hat{\tau}^2_{ML}$ is between study variance.

Hoogeveen et al. (2023) believed that applying conventional meta-analysis in ML4 project was inappropriate. They conducted a Bayesian hierarchical modeling instead of meta-analytic approach. This method was applied to different exclusion criteria. First analysis was made for the 17 labs with sufficiently large sample sizes under three exclusions as reported by Klein et al. (2022) but without excluding some data from in-house labs that was collected before the pre-analysis plan was finalized (545 participants were excluded). Hoogeveen et al. (2023) believed that because researchers of in-house labs constructed their own design independently, they were free to design their labs, there is no reason to exclude their observations. A second analysis was performed on the seven expert labs under three exclusions as reported by Klein et al. (2022). The sample sizes of these labs satisfied the recommendation by Chatard et al. (2020), ($N > 80$). A third analysis was performed across all studies (21 labs), applying the three exclusions to both expert and in-house labs, discarding missing values. Appendix D.2 for their results. All analyses gave results consistent with the absence of a mortality salience effect—all meta-analysis confidence intervals include zero. The data and the R code are available at (`https://github.com/SuzanneHoogeveen/ml4-reanalysis`).

4.4 Re-analysis of Many Labs

In this replication study, the original experiment was repeated across 21 labs, with some labs even involving the original authors from Greenberg et al. (1994) in their experiment. Despite strict protocols being followed, the results from the original study were not able to be replicated. Chatard et al. (2020) rebutted against this finding, and claims that by imposing a strong restriction on the studies to be included in the analysis, and by performing a meta-analysis on this subset of studies, the original result is replicated. Is the finding from Many Labs 4 accurate, or does the rebuttal by Chatard et al. (2020) hold significant merit? We aim to find the answer.

4.4.1 Many Labs 4 Data

We will carefully examine the ML4 paper before and after publication (Klein et al., 2019, 2022); their supplemental materials (full and aggregate data), and preregistered data plan.

The Rebuttal Many Labs 4 paper by (Chatard et al., 2020) and the data that was used in the reanalysis of the finding will be reviewed as well.

It is known that one factor in increasing the replicability power is that the sample size in the follow-up study must be greater than the initial study (Higgins et al., 2021). From 21 studies, there are 13 studies with sample sizes larger than 80 participants (2 studies have more than 200 participants, seven studies have more than 100 participants, and four studies have more than 80 participants). These studies are up to ten times larger than the original study; the total number of participants the original study is 23. Only the University of Illinois(IH — $N = 87$) successfully replicated the result, which is about 5% of the labs participating in the study—approximately the nominal significance level α. We know that large sample sizes alone do not guarantee replication if the different research environment for follow-up studies is high (Higgins et al., 2021), but how about all expert labs fail (9 studies)?

Before re-analyzing the data, we looked at summary statistics and the supplementary materials of Klein et al. (2018) to understand whether everything was performed correctly originally and to determine if there were issues with the re-analysis by Chatard et al. (2020). We found a few things: First, Chatard et al. (2020) excluded labs with less than 80 participants, which is never mentioned as a criterion for exclusion. They derived the exclusion criterion from the target sample size. However, the pre-registration plan stated that a lab should be included as long as it has equal to or greater than 60 participants. Moreover, Chatard et al. (2020) insisted on having large sample sizes in each lab, but their exclusion criteria actually reduced the sample sizes to below 25 in many studies.

Because there was debate about the pre-registration regarding the sample sizes, to increase the chance of replication, we will include both key analyses by Klein et al. (2019) and by Chatard et al. (2020). We re-estimate the data by using a random-effects two-stage meta-analysis by the maximum likelihood method, which ML4 used under the three different exclusions on 21 labs, 13 labs, expert labs, and in-house labs. However, when we tried to re-analyze the data, we found a slightly different result. Thus, we compared our data summary with that provided by Klein et al. (2018), and discovered some discrepancies. For example, for Exclusion Set 2, some participants who were not White Americans were included in the analysis, which led to a total number of 1,916 instead of 1,874. S imilarly, for Exclusion Set 3, some participants who did not strongly identify with the U.S. were included for a total of 1,723 instead of 1,693. The total number of participants in in-house labs is 1,421, which was correctly reported. We give corrected sample sizes in Table 4.6 and Table 4.7.

Fortunately for Klein et al. (2018), the main conclusions of his paper still held. It was more likely for us that Chatard et al. (2020) trusted the summaries of the aggregate data when performing their re-analysis. the data instead of the whole data. Therefore, we checked with the rebuttal's supplementary material at `https://osf.io/6v4kf/`.

After we re-analyze the data, we see that although the point estimates of expert labs are larger when compared to others, the width of confidence intervals are wider due to smaller sample sizes in each study. Thus, we conclude that ML4 still failed to replicate the initial finding and show no effect of mortality salience. The result is reported in Table 4.8. The Many Labs 4's failure, even though they attempted to avoid any potential causes of failure, moves us forward to ask the following questions: could this failure come from the method that they used? are the differences in the research environments making the exact replication of the original study impossible? We now aim to find answers to these questions.

Table 4.6: *Partial of summary 2 as reported by ML4*

Labs	AA/IH	N(TV)	N(MS)	Mean(TV)	Mean(MS)	SD (TV)	SD (MS)	Hedges' g	p-value
Ashland University	AA	~~26~~ 25	~~23~~ 22	1.83	1.39	1.85	2.02	-0.22	0.45
The College of New Jersey	AA	~~42~~ 9	~~59~~ 58	1.27	1.68	2.21	1.74	0.21	0.34
University of Kansas	AA	11	~~18~~ 17	1.76	0.8	2.41	2.49	-0.38	0.32
Occidental College	AA	~~18~~ 17	~~28~~ 21	0.57	0.79	1.74	2.76	0.09	0.76
Pace University	AA	~~34~~ 30	~~33~~ 29	1.69	1.53	2.01	2.19	-0.08	0.77
University of California, Riverside	AA	~~13~~ 6	~~13~~ 4	0.44	2.33	0.89	1.25	1.69	0.05
University of Wisconsin	AA	31	~~34~~ 32	0.88	0.86	1.68	2.27	-0.01	0.97
Virginia Commonwealth University	AA	12	~~28~~ 27	1.22	1.49	1.13	2	0.15	0.59

Table 4.7: *Partial of summary 3 as reported by ML4*

Labs	AA/IH	N(TV)	N(MS)	Mean(TV)	Mean(MS)	SD (TV)	SD (MS)	Hedges' g	p-value
Ashland University	AA	~~18~~ 17	14	2.43	1.76	1.85	2.29	-0.32	0.39
The College of New Jersey	AA	~~23~~ 21	~~32~~ 30	1.98	1.74	2.31	1.66	-0.12	0.69
University of Kansas	AA	6	~~11~~ 10	2.89	0.67	2.79	2.22	-0.87	0.13
Occidental College	AA	6	~~14~~ 10	0.5	2.1	1.38	3.42	0.51	0.21
Pace University	AA	~~17~~ 16	~~19~~ 16	2	2.21	2.21	2.16	0.09	0.79
University of California, Riverside	AA	~~10~~ 3	~~8~~ 3	1	2.33	0.88	1.53	1.05	0.28
University of Wisconsin	AA	19	~~23~~ 21	1.18	1.44	1.65	2.28	0.13	0.67
Virginia Commonwealth University	AA	7	~~16~~ 14	1.62	2.19	1.22	1.88	0.32	0.41

These tables are only a subset of the summary tables, see whole summary data in Tables D.4 and D.5. Red number indicates the incorrect number that was reported by Klein in his summary data. The correct number is defined above it

Table 4.8: *Re-analyze ML4 by ML method*

Key Analysis by	Labs	Exclusion Set	N	$\hat{\Delta}_H$	CI	p-value	$\hat{\tau}^2$
Klein et al. (2019)	All Labs (m=21)	Set 1	2220	0.03	$[-0.06, 0.12]$	0.58	0.003
		Set 2	1874	0.038	$[-0.05, 0.12]$	0.46	0.006
		Set 3	1693	0.031	$[-0.06, 0.12]$	0.57	0.008
	Expert Labs (m=9)	Set 1	799	0.04	$[-0.097, 0.18]$	0.55	0
		Set 2	453	0.09	$[-0.095, 0.28]$	0.3	0
		Set 3	272	0.08	$[-0.16, 0.32]$	0.5	0
	In-house labs (m=12)	All Sets	1421	0.02	$[-0.1 , 0.15]$	0.72	0.01
Chatard et al. (2020)	13 Labs	Set 1	1782	0.06	$[-0.03, 0.15]$	0.24	0.01
		Set 2	1476	0.09	$[-0.016, 0.20]$	0.16	0.016
		Set 3	1346	0.09	$[-0.03, 0.20]$	0.23	0.02
	Expert Labs (m=6)	Set 1	621	0.097	$[-0.04, 0.23]$	0.23	0
		Set 2*	315	0.2	$[0.01, 0.39]$	0.08	0
		Set 3	185	0.21	$[-0.04, 0.45]$	0.17	0
	In-House Labs (m=7)	All Sets	1161	0.057	$[-0.077, 0.19]$	0.49	0.0197

Here, m indicates number of studies, $\hat{\Delta}_H$ indicates the overall effect size by Hedges g estimator, and $\hat{\tau}^2$ is between study variance. Because one tail test was the Key analysis by Chatard et al. (2020) we use the same key analysis when we reanalysis 13 labs, expert labs (m=6), and In-house labs (m=7). $\ast$ indicates that the CI covered zero when we consider a two-tailed test with $\alpha = 0.05$, *[-0.0233,0.4291].*

4.4.2 Our Methodology

In our reanalysis, we want to include all different combinations of data exclusion criteria that derived by Klein et al. (2019), Klein et al. (2022), and Chatard et al. (2020) to increase the chance of finding an effect in some of these exclusion criteria. The following sections will be as follows. In section 4.4.2.1, we reexamine the performance of the ML estimator under meta-analysis that was used by Klein et al. (2019, 2022), and Chatard et al. (2020). Next, we perform an extensive simulation to evaluate the best estimator for performing meta-analysis using individual participant data (IPD) and for performing a mixed model. In section 4.4.2.2, an adjusted $p-value$, which accounts for changes in the research environment for a follow-up study, will be used. In section 4.4.2.3, a sensitivity-analysis approach for determining the significance of effect size estimates for meta-analysis and mixed-model estimators will be used.

4.4.2.1 Two-Stage IPD Meta-Analysis Approach and Mixed Model Approach

In Chapter 2, we conducted simulation studies under different scenarios of sample size setting rules, effect sizes, between-study heterogeneity, and numbers of studies to find the best approach (two-stage meta-analysis or mixed model) that performs better for analyzing the individual participants data. Based on simulations from a total of 5,760 meta-analysis scenarios (REML, ML, and PM) and 80 mixed-model scenarios (REML and ML), we found the following: First, the ML estimator should be avoided under both approaches, as the Type I errors are inflated, especially when the number of studies is small. For details, see Sections 2.5.4.1 and 2.6. On the other hand, REML performed better well under both meta-analysis and mixed model approaches (see Chapter 2, Sections 2.5 and 2.6). However, according to our simulations in meta-analysis, we found that the better form for estimating within-study variance is by Hedges'g estimator $s^2_{H.j}$, which is as follows

$$s^2_{H.j} = J^2_j \tilde{n}_j + \frac{\hat{\Delta}^2_{REML}}{2(n_{1j} + n_{2j})}$$

4.4.2.2 Adjusted $p - values$ for individual studies

The different environments can impede a successful replication, yet researchers do not consider it when measuring applicability. The different experimental environments contribute to a unique research environment in each experiment that can make it difficult for another researcher to repeat a study in the same way, even if both studies are conducted with high accuracy. The standard deviations of environment-by-treatment interaction divided by the experimental error should be considered; this dimensionless parameter is called the environmental effect ratio (EER) (Higgins et al., 2021).

$$EER = \frac{\sigma_\zeta}{\sigma_e} \tag{4.1}$$

There are several ways to estimate EER, see section 2.4 for more details. However, REML will be used because of its performance.

To make statistical inferences about $\mu_1 - \mu_2$, the usual p-values were used. Yet, this does not work here because these values do not account for changes in the research environment for a follow-up experiment. Therefore, the standard deviations of environment-by-treatment interaction divided by the experimental error should be considered; this dimensionless parameter is called the EER (Higgins et al., 2021). Given that $\hat{\Delta}$ follow-up studies has a noncentral t-distribution with degrees of freedom $df = n_1 + n_2 - 2$ and noncentrality parameter

$$\frac{\Delta\sqrt{n_h}}{\sqrt{1 + n_h EER^2}} \tag{4.2}$$

Higgins et al. (2021) derived the adjusted *p-values* that are functions of EER. The conditional p-values will be averaged across the environments to compute the adjusted *p-values*.

$$P\Big(\mid T \mid \geqslant \hat{\Delta}\sqrt{n_h} \mid follow-up\ study, \mu_1 - \mu_2 = 0\Big) = 2\Big(1 - G_0\Big(\frac{\hat{\Delta}\sqrt{n_h}}{\sqrt{1 + n_h E\hat{E}R^2}}\Big)\Big) \tag{4.3}$$

where $G_0()$ is the cdf of the central t-distribution, and

$$n_h = \frac{2}{1/n_1 + 1/n_2} \tag{4.4}$$

4.4.2.3 Sensitivity Analysis

Mixed models and meta-analysis approaches rely on estimating the variability in effect sizes across labs. However estimation of this variability is often inconsistent and unreliable when the number of labs included in the study are small. Thus, we consider a sensitivity analysis approach—which considers a range of potential values of this between-lab variance—for determining the significance of effect size estimates. For more details, see Section 3.4.

To make statistical inferences, the t distribution is used with both inverse-weighted standard errors and Hartung-Knapp standard errors. Regarding mixed-model method, the t-statistic with Satterthwaite's approximation, and the compound symmetry structure are used. A complete analysis was reported on the same prior settings, with varying exclusion criteria.

4.4.3 Results

From the current analysis, we found the two-stage meta-analysis approach and the linear mixed- model approach yielded the same conclusion. First, the effect size for follow-up studies was much smaller than original study (Effect size= 1.34). Second, there is some evidence of heterogeneity (τ^2_{REML}) in all labs, except in expert labs, compared to the previous result (τ^2_{ML}). Third, although the overall effect size is larger in the expert labs than in the others, there is evidence against an overall mortality salience effect whatever the exclusion set is considered. Therefore, the ML4 data was unable to replicate the result under any conditions, see Table 4.9. In addition, to assess robustness of this result, fixed effect meta analysis was used as well, and the result was unable to be replicated. See Table 4.10 for details.

Table 4.9: *Re-analysis ML4 under different key analyses by REML in meta-analysis and mixed model.*

				Meta-Analysis			Mixed Model		
Key analyses by	Labs	Execution Set	N-Criteria	$\hat{\Delta}_H$	p-value	$\hat{\tau}^2_{REML}$	$\hat{\Delta}_H$	p-value	$\hat{\tau}^2_{REML}$
Klein et al. (2019)	21 Labs	Set 1	All	0.031	0.558	0.010	0.011	0.844	0.021
		Set 2	All	0.052	0.423	0.017	0.019	0.763	0.025
		Set 3	All	0.043	0.514	0.019	0.015	0.836	0.037
	Expert Labs (m=9)	Set 1	All	0.042	0.565	0	0.044	0.536	0
		Set 2	All	0.106	0.386	0	0.079	0.408	0
		Set 3	All	0.084	0.533	0	0.056	0.689	0.020
	In-House Labs (m=12)	All Sets	All	0.034	0.671	0.025	-0.007	0.939	0.053
Chatard et al. (2020)	13 Labs	Set 1	$N \geqslant 80$	0.073	0.268	0.016	0.082	0.208	0.015
		Set 2	$N \geqslant 80$	0.118	0.194	0.030	0.110	0.155	0.021
		Set 3	$N \geqslant 80$	0.105	0.194	0.030	0.118	0.179	0.030
	Expert Labs (m=6)	Set 1	$N \geqslant 80$	0.098	0.253	0	0.103	0.200	0
		Set 2	$N \geqslant 80$	0.223	0.176	0	0.191	0.095	0
		Set 3	$N \geqslant 80$	0.224	0.117	0	0.208	0.163	0
	In-House Labs (m=7)	All Sets	$N \geqslant 80$	0.072	0.513	0.039	0.081	0.441	0.035
Klein et al. (2022)	17 Labs	Set 1	$N \geqslant 60$	0.070	0.247	0.010	0.030	0.658	0.030
		Set 2	$N \geqslant 60$	0.101	0.198	0.019	0.045	0.554	0.033
		Set 3	$N \geqslant 60$	0.096	0.186	0.021	0.049	0.560	0.040
	Expert Labs (m=7)	Set 1	$N \geqslant 60$	0.081	0.261	0	0.086	0.257	0
		Set 2	$N \geqslant 60$	0.183	0.187	0	0.160	0.126	0
		Set 3	$N \geqslant 60$	0.208	0.081	0	0.205	0.129	0
	In-House Labs (m=10)	All Sets	$N \geqslant 60$	0.054	0.574	0.036	-0.013	0.903	0.062

Here, m indicates number of studies. Under meta analysis, the p-value is calculated by t distribution using Hartung-Knapp standard errors. Under the mixed model, the p-value is calculated using the t distribution with a Satterthwaite approximation for the degrees of freedom.

Table 4.10: *Re-analyze ML4 under a fixed-effect meta-analysis*

Key Analysis by	Labs	Exclusion Set	$\hat{\Delta}_H$	95% CI	p-value
Klein et al. (2019)	All Labs (m=21)	Set 1	0.023	$[-0.061, \ 0.106]$	0.596
		Set 2	0.031	$[-0.061, \ 0.122]$	0.508
		Set 3	0.022	$[-0.074, \ 0.118]$	0.649
	Expert Labs (m=9)	Set 1	0.042	$[-0.098, \ 0.182]$	0.555
		Set 2	0.092	$[-0.095, \ 0.280]$	0.335
		Set 3	0.079	$[-0.163, \ 0.321]$	0.523
	In-house labs (m=12)	All Sets	0.012	$[-0.093, \ 0.116]$	0.826
Chatard et al. (2020)	All Labs (m=13)	Set 1	0.052	$[-0.041, \ 0.146]$	0.272
		Set 2	0.065	$[-0.038, \ 0.168]$	0.217
		Set 3	0.053	$[-0.055, \ 0.161]$	0.337
	Expert Labs (m=6)	Set 1	0.097	$[-0.062, \ 0.255]$	0.232
		Set 2	0.203	$[-0.023, \ 0.429]$	0.079
		Set 3	0.208	$[-0.086, \ 0.502]$	0.166
	In-house labs (m=7)	All Sets	0.029	$[-0.087, \ 0.145]$	0.627
Klein et al. (2022)	All Labs (m=17)	Set 1	0.069	$[-0.031, \ 0.169]$	0.178
		Set 2	0.092	$[-0.021, \ 0.205]$	0.111
		Set 3	0.087	$[-0.033, \ 0.208]$	0.156
	Expert Labs (m=7)	Set 1	0.080	$[-0.069, \ 0.229]$	0.292
		Set 2	0.166	$[-0.039, \ 0.372]$	0.113
		Set 3	0.194	$[-0.072, \ 0.460]$	0.153
	In-house labs (m=10)	All Sets	0.060	$[-0.076, \ 0.195]$	0.388

Analysis of the EER within subgroups suggest that expert-protocol labs had much smaller across-lab variability than in-house labs. See Table 4.11 for details. It is clear that expertise and experience mitigate the differences in the research environments. Small values of EER should ensure that the chance of replication in the follow-up experiment is large, especially because the sample size in the follow-up experiment is greater than in the original one.

Table 4.11: *EER under Different Exclusions by REML*

			EER		
Key analyses by	Labs	AA/IH	Exclusion Set 1	Exclusion Set 2	Exclusion Set 3
Klein et al. (2019)	21 Labs	both	0.10	0.11	0.14
	9 Labs	AA	2e-05	1e-04	0.10
	13 Labs	IH	0.16	0.16	0.16
Chatard et al. (2020)	13 Labs	both	0.09	0.10	0.12
	6 Labs	AA	2e-05	6e-05	6e-05
	7 Labs	IH	0.13	0.13	0.13
Klein et al. (2022)	17 Labs	both	0.12	0.13	0.14
	7 Labs	AA	9e-05	6e-05	7e-05
	10 Labs	IH	0.18	0.18	0.18

We re-analyze the individual studies using adjusted p-values assuming these EER values. data after accounting for the changes in the research environment when Greenberg et al. (1994) is redone. The adjusted p-value is used for the individual lab design, see Table 4.12 for details. Apart from which exclusion criteria were used, all the follow-up studies failed to replicate Study 1 under except for the University of Illinois lab.

Table 4.12: *12 Labs: Adjusted p-value for each study under 3 differed exclusions*

Labs*	AA/IH	adjusted p-value		
		Exclusion Set 1	Exclusion Set 2	Exclusion Set 3
AU	AA	0.2	0.511	0.445
UK	AA	0.832	0.376	0.14
UF	IH	0.374	0.4	0.459
APU	IH	0.801	0.804	0.81
OC	AA	0.717	0.799	0.354
UP	IH	0.504	0.518	0.549
BYU	AA	0.115	0.203	0.254
PU	IH	0.238	0.249	0.276
UW	IH	0.736	0.743	0.759
TCNJ	AA	0.297	0.429	0.726
PU	AA	0.744	0.806	0.819
UW	AA	0.952	0.977	0.725
UI	IH	0.005	0.007	0.012
PLU	IH	0.973	0.974	0.977
VCU	AA	0.953	0.699	0.528
IC	IH	0.824	0.832	0.85
UCR	AA	0.876	0.038	0.367
WU	IH	0.76	0.771	0.794
UK	IH	0.898	0.901	0.908
SOU	IH	0.172	0.177	0.192
WPI	IH	0.965	0.966	0.97

$*$ *Full name of each lab can be found in Section D.1.1*

4.5 Conclusion

A preregistered replication project involving 2,281 participants across 21 labs by (Klein et al., 2022) could not replicate an important effect in social psychology, the Mortality Salience Effect (MS) from Terror Management Theory (TMT), under any conditions. Klein et al. (2022) tried to avoid any potential causes of failure to replicate, including involving the original authors in their design. A re-analysis of this data by Chatard et al. (2020) contradicted the findings from the initial analysis by Klein et al. (2018). Due to these disparate findings, we perform our own analysis of this study.

There are several appealing properties of the ML4 study, one of which is that their data and pre-analysis plan are publicly available, making it possible for others to reproduce their findings and to catch any potential errors. We found that the aggregate data that was used to analyze Exclusion Sets 2 and 3 were incorrect, which ultimately negated the findings in Chatard et al. (2020). Note, when individual participant data is available, analysis of this data directly is preferable; if there are any errors in the reported summarized data, this can lead to incorrect conclusions.

We were unable to achieve any kind of significant result using the methodology used in Klein et al. (2022), and Chatard et al. (2020), or using best practices for analyzing multi-lab data described in Chapter 2. Moreover, a sensitivity analysis showed that replicability was not possible under any hypothesized value of the across-lab variability.

Chapter 5

Conclusion

5.1 Summary

Recently, a large-scale, preregistered study designed by Klein et al. (2022) tried to replicate an important effect in social psychology, the Mortality Salience Effect (MS) from Terror Management Theory (TMT) by Greenberg et al. (1994). To increase the chance of replication, 37 researchers cooperated and implemented a multi-laboratory studies design and involved the original authors in their study to leverage all possible expertise to identify a study suitable for replication. Although they strictly adhered to their preregistration plan and followed the advice from the experts of TMT, the data provided evidence against MS. Chatard et al. (2020) rebutted this finding and claimed that by imposing a strong restriction on the studies to be included in the analysis and by performing a meta-analysis on this subset of studies, the original result can be replicated. Which conclusion is correct by Klein et al. (2022) or Chatard et al. (2020)?

In Chapter 2, we perform an extensive literature review on random-effects meta-analysis methods and perform an extensive simulation to evaluate the best practices for performing meta-analysis when using individual participant data (IPD) instead of aggregate data from multi-lab studies. We then compare the best-performing meta-analysis estimators to those obtained from a mixed model. Overall, we considered 5,760 combinations of estimators and multi-lab experimental settings for meta-analysis and another 80 for mixed models. We found that the meta-analysis and linear mixed-model approaches yield similar results, with meta-analysis methods performing slightly better with Hartung-Knapp standard errors, which produces more inaccurate estimates of across-lab variability compared to other methods. Additionally, we found that both estimation methods suffer from the same significant pitfall— estimation of across-lab variability in treatment effects is often inconsistent and unreliable when the number of labs included in the study is small.

In Chapter 3, we develop a sensitivity-analysis approach for determining the significance of effect size estimates for both meta-analysis and mixed-model estimators. These methods allow researchers to consider a range of different values for the across-lab variance and help them determine for what values the estimate of the effect size is statistically significant. While effective, we found that these approaches can be misleading when the claimed or estimated across-lab variance is much lower than the actual across-lab variance, which can be possible when sample sizes or the number of labs under study is small.

In chapter 4, we reviewed ML4 (Klein et al., 2018, 2022) and rebuttal ML4 (Chatard et al., 2020) to get a clear picture of their analysis. We identified issues with the analysis in the rebuttal by Chatard et al. (2020) that led them to obtain inaccurate results. Next, we applied our methods to re-analyze the ML4 data under three different exclusions as specified by (Klein et al., 2018, 2022), and (Chatard et al., 2020) to increase the chance of finding an effect in some of these exclusion criteria. Our analysis corroborates the original finding of Klein et al. (2018) that the original experiment is unable to be replicated with any kind of consistency.

5.2 Future Research

It is commonly assumed that heterogeneity between studies in a meta-analysis is constant, which means they are from the same population. However, in reality, they may not be. In recent literature, the between study variance τ^2 is replaced to be variable across different subpopulations: τ_j^2. The idea of varying τ_j^2 was given in Thompson and Becker (2020) and Williams et al. (2021). Because there is a connection between τ and EER, the quantity of EER will be extended to be EER_j. We will focus on finding which variables may cause heterogeneity changes between studies. Additionally, we may aim to extend the replicability framework involving the EER to more than two treatment conditions. In the meta-analysis literature, this is known as network meta-analysis.

Bibliography

Ana C Alba, Paul E Alexander, Joanne Chang, John MacIsaac, Samantha DeFry, and Gordon H Guyatt. High statistical heterogeneity is more frequent in meta-analysis of continuous than binary outcomes. *Journal of clinical epidemiology*, 70:129–135, 2016.

Mike Allen and Raymond Preiss. Replication and meta-analysis: A necessary connection. *Journal of Social Behavior and Personality*, 8(6):9, 1993.

Jamie Arndt, Jeff Greenberg, Tom Pyszczynski, and Sheldon Solomon. Subliminal exposure to death-related stimuli increases defense of the cultural worldview. *Psychological Science*, 8(5):379–385, 1997.

Ilyas Bakbergenuly, David C Hoaglin, and Elena Kulinskaya. Simulation study of estimating between-study variance and overall effect in meta-analysis of odds-ratios. *arXiv preprint arXiv:1902.07154*, 2019.

Monya Baker. Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the'crisis rocking science and what they think will help. *Nature*, 533(7604): 452–455, 2016.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

Vincent Bauchau. Is there a" file drawer problem" in biological research? *Oikos*, pages 407–409, 1997.

Nora M Bello and David G Renter. Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. *Journal of dairy science*, 101 (7):5679–5701, 2018.

David C Berliner. Comment: Educational research: The hardest science of all. *Educational researcher*, 31(8):18–20, 2002.

William H Beyer. *Handbook of tables for probability and statistics.* Crc Press, 2019.

BJ Biggerstaff and RL Tweedie. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in medicine*, 16(7):753–768, 1997.

Dankmar Böhning, Uwe Malzahn, Ekkehart Dietz, Peter Schlattmann, Chukiat Viwatwongkasem, and Annibale Biggeri. Some general points in estimating heterogeneity variance with the dersimonian–laird estimator. *Biostatistics*, 3(4):445–457, 2002.

Douglas G Bonett. Replication-extension studies. *Current Directions in Psychological Science*, 21(6):409–412, 2012.

Dennis D Boos and Leonard A Stefanski. P-value precision and reproducibility. *The American Statistician*, 65(4):213–221, 2011.

Michael Borenstein, Julian PT Higgins, Larry V Hedges, and Hannah R Rothstein. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research synthesis methods*, 8(1):5–18, 2017.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis.* John Wiley & Sons, 2021.

George F Borm and A Rogier T Donders. Updating meta-analyses leads to larger type i errors than publication bias. *Journal of clinical epidemiology*, 62(8):825–830, 2009.

Sanford L Braver, Felix J Thoemmes, and Robert Rosenthal. Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, 9(3):333–342, 2014.

Jesper Brok, Kristian Thorlund, Christian Gluud, and Jørn Wetterslev. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *Journal of clinical epidemiology*, 61(8):763–769, 2008.

Brian L Burke, Andy Martens, and Erik H Faucher. Two decades of terror management theory: A meta-analysis of mortality salience research. *Personality and Social Psychology Review*, 14(2):155–195, 2010.

Nicholas R Buttrick, Balazs Aczel, Lena F Aeschbach, Bence E Bakos, Florian Brühlmann, Heather M Claypool, Joachim Hüffmeier, Marton Kovacs, Kurt Schuepfer, Peter Szecsi, et al. Many labs 5: registered replication of vohs and schooler (2008), experiment 1. *Advances in Methods and Practices in Psychological Science*, 3(3):429–438, 2020.

Maxwell Cairns and Luke Prendergast. On ratio measures of population heterogeneity for meta-analyses. *arXiv preprint arXiv:2009.10332*, 2020.

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.

Siri Carpenter. Harvard psychology researcher committed fraud, us investigation concludes. *Science*, 6, 2012.

Armand Chatard, Gilad Hirschberger, and Tom Pyszczynski. A word of caution about many labs 4: If you fail to follow your preregistered plan, you may fail to find a real effect. *Preprint Available on PsyArXiv*, 2020. URL `https://psyarxiv.com/ejubn/`.

Roy Chen, Yan Chen, and Yohanes E Riyanto. Best practices in replication: a case study of common information in coordination games. *Experimental Economics*, 24:2–30, 2021.

Yeojin Chung, Sophia Rabe-Hesketh, and In-Hee Choi. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in medicine*, 32(23):4071–4089, 2013.

Jacob Cohen. *Statistical power analysis for the behavioral sciences.* Academic press, 2013.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Harris Cooper, Larry V Hedges, and Jeffrey C Valentine. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation, 2019.

Jonathan Davey, Rebecca M Turner, Mike J Clarke, and Julian Higgins. Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC medical research methodology*, 11(1): 1–11, 2011.

Mark Dechesne, Tom Pyszczynski, Jamie Arndt, Sean Ransom, Kennon M Sheldon, Ad Van Knippenberg, and Jacques Janssen. Literal and symbolic immortality: The effect of evidence of literal immortality on self-esteem striving in response to mortality salience. *Journal of personality and social psychology*, 84(4):722, 2003.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

Edward Diener and Robert Biswas-Diener. The replication crisis in psychology, 2016.

C Patrick Doncaster and Rebecca Spake. Correction for bias in meta-analysis of little-replicated studies. *Methods in Ecology and Evolution*, 9(3):634–644, 2018.

Sorinel Dumitrescu. Estimates for the ratio of gamma functions using higher order roots. *Stud. Univ. Babeş-Bolyai Math*, 60:173–181, 2015.

Charles R Ebersole, Olivia E Atherton, Aimee L Belanger, Hayley M Skulborstad, Jill M Allen, Jonathan B Banks, Erica Baranski, Michael J Bernstein, Diane BV Bonfiglio, Leanne Boucher, et al. Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68–82, 2016.

Charles R Ebersole, Maya B Mathur, Erica Baranski, Diane-Jo Bart-Plange, Nicholas R Buttrick, Christopher R Chartier, Katherine S Corker, Martin Corley, Joshua K Hartshorne, Hans IJzerman, et al. Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, 3(3):309–331, 2020.

Dov Eden. From the editors: Replication, meta-analysis, scientific progress, and amj's publication policy. *Academy of Management Journal*, pages 841–846, 2002.

Matthias Egger, Martin Schneider, and George Davey Smith. Meta-analysis spurious precision? meta-analysis of observational studies. *Bmj*, 316(7125):140–144, 1998.

Matthias Egger, George Davey Smith, and Douglas Altman. *Systematic reviews in health care: meta-analysis in context*. John Wiley & Sons, 2008.

Martin Eisend and Farid Tarrahi. Meta-analysis selection bias in marketing research. *International Journal of Research in Marketing*, 31(3):317–326, 2014.

Scott R Eliason. *Maximum likelihood estimation: Logic and practice*. Sage, 1993.

Leandre R Fabrigar and Duane T Wegener. Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, 66:68–80, 2016.

Uljana Feest. Why replication is overrated. *Philosophy of Science*, 86(5):895–905, 2019.

A Field. Can meta-analysis be trusted? *PSYCHOLOGIST-LEICESTER-*, 16(12):642–645, 2003.

Alessandro Filazzola and James F Cahill Jr. Replication in field ecology: Identifying challenges and proposing solutions. *Methods in Ecology and Evolution*, 12(10):1780–1792, 2021.

Samuel C Fletcher. Replication is for meta-analysis. *Philosophy of Science*, pages 1–23, 2022.

Dean A Follmann and Michael A Proschan. Valid inference in random effects meta-analysis. *Biometrics*, 55(3):732–737, 1999.

Gregory Francis. Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, 19(6):975–991, 2012.

Hannah Fraser, Ashley Barnett, Timothy H Parker, and Fiona Fidler. The role of replication studies in ecology. *Ecology and Evolution*, 10(12):5197–5207, 2020.

Eric W Gibson. The role of p-values in judging the strength of evidence and realistic replication expectations. *Statistics in Biopharmaceutical Research*, 13(1):6–18, 2021.

Jeff Greenberg, Tom Pyszczynski, and Sheldon Solomon. The causes and consequences of a need for self-esteem: A terror management theory. In *Public self and private self*, pages 189–212. Springer, 1986.

Jeff Greenberg, Tom Pyszczynski, Sheldon Solomon, Linda Simon, and Michael Breus. Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of personality and social psychology*, 67(4):627, 1994.

Burak Kürsad Günhan, Christian Röver, and Tim Friede. Random-effects meta-analysis of few studies involving rare events. *Research Synthesis Methods*, 11(1):74–90, 2020.

Ruoling Guo, Max H Pittler, and Edzard Ernst. Hawthorn extract for treating chronic heart failure. *Cochrane Database of Systematic Reviews*, 37(1), 2008.

Johanna Gustafsson. Single case studies vs. multiple case studies: A comparative study, 2017.

Judith A Hall and Robert Rosenthal. Testing for moderator variables in meta-analysis: Issues and methods. *Communications Monographs*, 58(4):437–448, 1991.

Elizabeth A Hamman, Paula Pappalardo, James R Bence, Scott D Peacor, and Craig W Osenberg. Bias in meta-analyses using hedges'd. *Ecosphere*, 9(9):e02419, 2018.

Rebecca J Hardy and Simon G Thompson. A likelihood approach to meta-analysis with random effects. *Statistics in medicine*, 15(6):619–629, 1996.

Rebecca J Hardy and Simon G Thompson. Detecting and describing heterogeneity in meta-analysis. *Statistics in medicine*, 17(8):841–856, 1998.

J Hartung and KH Makambi. Positive estimation of the between-study variance in meta-analysis: theory and methods. *South African Statistical Journal*, 36(1):55–76, 2002.

Joachim Hartung. An alternative method for meta-analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 41(8):901–916, 1999.

Joachim Hartung and Kepher H Makambi. Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics-Simulation and Computation*, 32 (4):1179–1190, 2003.

Lee Harvey. Research fraud: a long-term problem exacerbated by the clamour for research grants, 2020.

Michael Harwell. An empirical study of hedge's homogeneity test. *Psychological methods*, 2 (2):219, 1997.

Larry V Hedges. Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2):107–128, 1981.

Larry V Hedges. Estimation of effect size from a series of independent experiments. *Psychological bulletin*, 92(2):490, 1982.

Larry V Hedges. A random effects model for effect sizes. *Psychological Bulletin*, 93(2):388, 1983.

Larry V Hedges and Ingram Olkin. *Statistical methods for meta-analysis*. Academic press, 2014.

Larry V Hedges and Jacob M Schauer. More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, 44(5): 543–570, 2019a.

Larry V Hedges and Jacob M Schauer. Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, 24(5):557, 2019b.

Larry V Hedges and Jack L Vevea. Fixed-and random-effects models in meta-analysis. *Psychological methods*, 3(4):486, 1998.

Masayuki Henmi and John B Copas. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in medicine*, 29(29):2969–2983, 2010.

Daniel J Hicks. Open science, the replication crisis, and environmental public health. *Accountability in Research*, 30(1):34–62, 2023.

James J Higgins, Michael J Higgins, and Jinguang Lin. From one environment to many: The problem of replicability of statistical inferences. *The American Statistician*, 75(3): 334–342, 2021.

Julian PT Higgins and Simon G Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558, 2002.

Julian PT Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

Johannes Hönekopp and Audrey Helen Linden. Heterogeneity estimates in a biased world. *PloS one*, 17(2):e0262809, 2022.

Suzanne Hoogeveen, Sophie W Berkhout, Quentin F Gronau, Eric-Jan Wagenmakers, and Julia M Haaf. Improving statistical analysis in team science: The case of a bayesian multiverse of many labs 4. *PsyArXiv*, 2023.

Mingxiu Hu, Joseph C Cappelleri, and KK Gordon Lan. Applying the law of iterated logarithm to control type i error in cumulative meta-analysis of binary outcomes. *Clinical Trials*, 4(4):329–340, 2007.

Tania B Huedo-Medina, Julio Sánchez-Meca, Fulgencio Marin-Martinez, and Juan Botella. Assessing heterogeneity in meta-analysis: Q statistic or i^2 index? *Psychological methods*, 11(2):193, 2006.

John E Hunter and Frank L Schmidt. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage, 2004.

Joanna IntHout, John PA Ioannidis, George F Borm, and Jelle J Goeman. Small studies are more heterogeneous than large ones: a meta-meta-analysis. *Journal of clinical epidemiology*, 68(8):860–869, 2015.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8): e124, 2005.

John PA Ioannidis. Meta-research: The art of getting it wrong. *Research synthesis methods*, 1(3-4):169–184, 2010.

John PA Ioannidis. Failure to replicate: sound the alarm. In *Cerebrum: The Dana forum on brain science*, volume 2015. Dana Foundation, 2015.

John PA Ioannidis. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3):485–514, 2016.

John PA Ioannidis, David B Allison, Catherine A Ball, Issa Coulibaly, Xiangqin Cui, Aedín C Culhane, Mario Falchi, Cesare Furlanello, Laurence Game, Giuseppe Jurman, et al. Repeatability of published microarray gene expression analyses. *Nature genetics*, 41(2):149–155, 2009.

Dan Jackson and Ian R White. When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, 60(6):1040–1058, 2018.

Dan Jackson, Martin Law, Gerta Rücker, and Guido Schwarzer. The hartung-knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? *Statistics in medicine*, 36(25):3923–3934, 2017.

Savita Jain, Suresh K Sharma, and Kanchan Jain. Meta-analysis of fixed, random and mixed effects models. *International Journal of Mathematical, Engineering and Management Sciences*, 4(1):199–218, 2019.

Iman Jaljuli, Neri Kafkafi, Eliezer Giladi, Ilan Golani, Illana Gozes, Elissa Chesler, Molly A Bogue, and Yoav Benjamini. Improving replicability using interaction with laboratories: a multi-lab experimental assessment. *bioRxiv*, pages 2021–12, 2021.

Iman Jaljuli, Neri Kafkafi, Eliezer Giladi, Ilan Golani, Illana Gozes, Elissa J Chesler, Molly A Bogue, and Yoav Benjamini. A multi-lab experimental assessment reveals that replicability can be improved by using empirical estimates of genotype-by-lab interaction. *Plos Biology*, 21(5):e3002082, 2023.

Neri Kafkafi, Yoav Benjamini, Anat Sakov, Greg I Elmer, and Ilan Golani. Genotype–environment interactions in mouse behavior: a way out of the problem. *Proceedings of the National Academy of Sciences*, 102(12):4619–4624, 2005.

Neri Kafkafi, Ilan Golani, Iman Jaljuli, Hugh Morgan, Tal Sarig, Hanno Würbel, Shay Yaacoby, and Yoav Benjamini. Addressing reproducibility in single-laboratory phenotyping experiments. *Nature methods*, 14(5):462–464, 2017.

Neri Kafkafi, Joseph Agassi, Elissa J Chesler, John C Crabbe, Wim E Crusio, David Eilam, Robert Gerlai, Ilan Golani, Alex Gomez-Marin, Ruth Heller, et al. Reproducibility and replicability of rodent phenotyping in preclinical studies. *Neuroscience & Biobehavioral Reviews*, 87:218–232, 2018.

Michael G Kenward and James H Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, pages 983–997, 1997.

Richard Klein, Kate Ratliff, Michelangelo Vianello, Reginald Adams Jr, Stěpán Bahník, Michael Bernstein, Konrad Bocian, Mark Brandt, Beach Brooks, Claudia Brumbaugh, et al. Data from investigating variation in replicability: A “many labs” replication project. *Journal of Open Psychology Data*, 2(1), 2014a.

Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, et al. Investigating variation in replicability. *Social psychology*, 2014b.

Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

Richard A Klein, Corey L Cook, Charles R Ebersole, Christine Vitiello, Brian A Nosek, Christopher R Chartier, Cody D Christopherson, Samuel Clay, Brian Collisson, Jarret Crawford, et al. Many labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Preprint Available on PsyArXiv*, 2019. URL `https://psyarxiv.com/vef2c/`.

Richard A Klein, Corey L Cook, Charles R Ebersole, Christine Vitiello, Brian A Nosek, Joseph Hilgard, Paul Hangsan Ahn, Abbie J Brady, Christopher R Chartier, Cody D Christopherson, et al. Many labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, 8(1):35271, 2022.

Stanley B Klein. What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology*, 24(3):326–338, 2014.

Guido Knapp and Joachim Hartung. Improved tests for a random effects meta-regression with a single covariate. *Statistics in medicine*, 22(17):2693–2710, 2003.

Spyros Konstantopoulos. Fixed and mixed effects models in meta-analysis. *Available on IZA Discussion Papers*, 2006. URL `https://www.econstor.eu/bitstream/10419/33946/1/51474698X.pdf`.

Evangelos Kontopantelis, David A Springate, and David Reeves. A re-analysis of the cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PloS one*, 8(7):e69930, 2013.

Sander L Koole and Daniël Lakens. Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, 7(6):608–614, 2012.

Julia Koricheva, Jessica Gurevitch, and Kerrie Mengersen. *Handbook of meta-analysis in ecology and evolution*. Princeton University Press, 2013.

Alexandra Kuznetsova, Per B Brockhoff, and Rune HB Christensen. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26, 2017.

Amanda Kvarven, Eirik Strømland, and Magnus Johannesson. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4 (4):423–434, 2020.

Anthony Lane, Olivier Luminet, Gideon Nave, and Moïra Mikolajczak. Is there a publication bias in behavioural intranasal oxytocin research on humans? opening the file drawer of one laboratory. *Journal of neuroendocrinology*, 28(4), 2016.

Dean Langan. *Estimating the Heterogeneity Variance in a Random-Effects Meta-Analysis*. PhD thesis, University of York, 2015.

Dean Langan, Julian PT Higgins, Dan Jackson, Jack Bowden, Areti Angeliki Veroniki, Evangelos Kontopantelis, Wolfgang Viechtbauer, and Mark Simmonds. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research synthesis methods*, 10(1):83–98, 2019.

Keith Laws. It's time for psychologists to put their house in order. *The Guardian*, 27, 2013.

Stephan Lewandowsky and Klaus Oberauer. Low replicability can support robust and efficient science. *Nature communications*, 11(1):1–12, 2020.

Molly Lewis, Maya B Mathur, Tyler J VanderWeele, and Michael C Frank. The puzzling relationship between multi-laboratory replications and meta-analyses of the published literature. *Royal Society Open Science*, 9(2):211499, 2022.

Lifeng Lin. Bias caused by sampling error in meta-analysis with small sample sizes. *PloS one*, 13(9):e0204056, 2018.

Lifeng Lin and Ariel M Aloe. Evaluation of various estimators for standardized mean difference in meta-analysis. *Statistics in Medicine*, 40(2):403–426, 2021.

Mark W Lipsey and David B Wilson. *Practical meta-analysis.* SAGE publications, Inc, 2001.

Eric Loken and Andrew Gelman. Measurement error and the replication crisis. *Science*, 355 (6325):584–585, 2017.

Steven G Luke. Evaluating significance in linear mixed-effects models in r. *Behavior research methods*, 49:1494–1502, 2017.

Rebecca A Lundwall. Changing institutional incentives to foster sound scientific practices: One department. *Infant Behavior and Development*, 55:69–76, 2019.

Fulgencio Marin-Martinez and Julio Sánchez-Meca. Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, 70(1):56–73, 2010.

Scott E Maxwell, Michael Y Lau, and George S Howard. Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487, 2015.

Jonathon McPhetres, Nihan Albayrak-Aydemir, Ana Barbosa Mendes, Elvina C Chow, Patricio Gonzalez-Marquez, Erin Loukras, Annika Maus, Aoife O'Mahony, Christina Pomareda, Maximilian A Primbs, et al. A decade of theory as reflected in psychological science (2009–2019). *PloS one*, 16(3):e0247986, 2021.

Blakeley B McShane, Ulf Böckenholt, and Karsten T Hansen. Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5):730–749, 2016.

Blakeley B McShane, Jennifer L Tackett, Ulf Böckenholt, and Andrew Gelman. Large-scale replication projects in contemporary psychological research. *The American Statistician*, 73(sup1):99–105, 2019.

Kerrie Mengersen and Christopher H Schmid. 10. maximum likelihood approaches to meta-analysis. In *Handbook of Meta-analysis in Ecology and Evolution*, pages 125–144. Princeton University Press, 2013.

Milliken and Dallas E. Johnson Johnson, George A. Milliken. *Analysis of messy data, volume I: Designed Experiments, Second Edition.* Chapman and Hall/CRC, 2009.

M Mittlböck and H Heinzl. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Statistics in medicine*, 25(24):4321–4333, 2006.

Cristinel Mortici. New approximation formulas for evaluating the ratio of gamma functions. *Mathematical and Computer Modelling*, 52(1-2):425–433, 2010.

Michael Muthukrishna and Joseph Henrich. A problem in theory. *Nature Human Behaviour*, 3(3):221–229, 2019.

Priscilla Nagarajan, Bharath Garla, M Taranath, and I Nagarajan. The file drawer effect: A call for meticulous methodology and tolerance for non-significant results. *Indian Journal of Anesthesia*, 61(6), 2017.

Leif D Nelson, Joseph Simmons, and Uri Simonsohn. Psychology's renaissance. *Annual review of psychology*, 69:511–534, 2018.

Mante S Nieuwland, Stephen Politzer-Ahles, Evelien Heyselaar, Katrien Segaert, Emily Darley, Nina Kazanina, Sarah Von Grebmer Zu Wolfsthurn, Federica Bartolozzi, Vita Kogan, Aine Ito, et al. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, 7:e33468, 2018.

Putri W Novianti, Kit CB Roes, and Ingeborg van der Tweel. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemporary clinical trials*, 37 (1):129–138, 2014.

National Academies of Sciences, Engineering, Medicine, et al. *Reproducibility and replicability in science*. National Academies Press, 2019.

Ingram Olkin and Allan Sampson. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*, pages 317–322, 1998.

Thammarat Panityakul, Chinnaphong Bumrungsup, and Guido Knapp. On estimating residual heterogeneity in random-effects meta-regression: A comparative study. *J. Stat. Theory Appl.*, 12(3):253–265, 2013.

Robert C Paule and John Mandel. Consensus values and weighting factors. *Journal of research of the National Bureau of Standards*, 87(5):377, 1982.

Tiago V Pereira and John PA Ioannidis. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of clinical epidemiology*, 64(10): 1060–1069, 2011.

Terri Pigott. *Advances in meta-analysis*. Springer Science & Business Media, 2012.

José Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, Siem Heisterkamp, Bert Van Willigen, and R Maintainer. Package 'nlme'. *Linear and nonlinear mixed effects models, version*, 3(1):274, 2017.

Charles Poole and Sander Greenland. Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology*, 150(5):469–475, 1999.

Stephen M Powers and Stephanie E Hampton. Open science, reproducibility, and transparency in ecology. *Ecological applications*, 29(1):e01822, 2019.

Tom Pyszczynski, Sheldon Solomon, and Jeff Greenberg. Thirty years of terror management theory: From genesis to revelation. In *Advances in experimental social psychology*, volume 52, pages 1–70. Elsevier, 2015.

Brian Resnick. More social science studies just failed to replicate. here's why this is good. *Vox. com*, 2018.

Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.

María Rubio-Aparicio, José Antonio López-López, Wolfgang Viechtbauer, Fulgencio Marín-Martínez, Juan Botella, and Julio Sánchez-Meca. Testing categorical moderators in mixed-effects meta-analysis in the presence of heteroscedasticity. *The Journal of Experimental Education*, 88(2):288–310, 2020.

Andrew L Rukhin. Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):451–469, 2013.

Andrew L Rukhin, Brad J Biggerstaff, and Mark G Vangel. Restricted maximum likelihood estimation of a common mean and the mandel–paule algorithm. *Journal of Statistical Planning and Inference*, 83(2):319–330, 2000.

Julio Sánchez-Meca and Fulgencio Marín-Martínez. Homogeneity tests in meta-analysis: A monte carlo comparison of statistical power and type i error. *Quality and Quantity*, 31(4): 385–399, 1997.

Julio Sánchez-Meca and Fulgencio Marin-Martinez. Weighting by inverse variance or by sample size in meta-analysis: A simulation study. *Educational and Psychological Measurement*, 58(2):211–220, 1998.

Julio Sánchez-Meca and Fulgencio Marín-Martínez. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological methods*, 13(1):31, 2008.

Patarawan Sangnawakij, Dankmar Böhning, Sa-Aat Niwitpong, Stephen Adams, Michael Stanton, and Heinz Holling. Meta-analysis without study-specific variance information: Heterogeneity case. *Statistical Methods in Medical Research*, 28(1):196–210, 2019.

Antonio Sartal, Miguel González-Loureiro, and Xosé H Vázquez. Meta-analyses in management: What can we learn from clinical research? *BRQ Business Research Quarterly*, 24 (1):91–111, 2021.

Franklin E Satterthwaite. Synthesis of variance. *Psychometrika*, 6(5):309–316, 1941.

Svenja E Seide, Christian Röver, and Tim Friede. Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. *BMC medical research methodology*, 19(1):1–14, 2019.

Doron J Shahar et al. Minimizing the variance of a weighted average. *Open Journal of Statistics*, 7(02):216, 2017.

Donald Sharpe and Sarena Poets. Meta-analysis as a response to the replication crisis. *Canadian Psychology/Psychologie canadienne*, 61(4):377, 2020.

Richard M Shiffrin, Katy Börner, and Stephen M Stigler. Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, 115(11): 2632–2639, 2018.

Kurex Sidik and Jeffrey N Jonkman. Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis*, 50(12):3681–3701, 2006.

Kurex Sidik and Jeffrey N Jonkman. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in medicine*, 26(9):1964–1981, 2007.

Lianne Siegel, M Hassan Murad, and Haitao Chu. Estimating the reference range from a meta-analysis. *Research synthesis methods*, 12(2):148–160, 2021.

Daniel J Simons, Alex O Holcombe, and Barbara A Spellman. An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, 9(5):552–555, 2014.

Bimal K Sinha, Joachim Hartung, and Guido Knapp. *Statistical meta-analysis with applications*. John Wiley & Sons, 2011.

Dalene Stangl and Donald A Berry. *Meta-analysis in medicine and health policy*. CRC Press, 2000.

Tom D Stanley, Evan C Carter, and Hristos Doucouliagos. What meta-analyses reveal about the replicability of psychological research. *Psychological bulletin*, 144(12):1325, 2018.

Theodore D Sterling. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association*, 54(285):30–34, 1959.

Wolfgang Stroebe. What can we learn from many labs replications? *Basic and Applied Social Psychology*, 41(2):91–103, 2019.

Wolfgang Stroebe and Fritz Strack. The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9(1):59–71, 2014.

Wolfgang Stroebe, Tom Postmes, and Russell Spears. Scientific misconduct and the myth of self-correction in science. *Perspectives on psychological science*, 7(6):670–688, 2012.

Shonosuke Sugasawa and Hisashi Noma. A unified method for improved inference in random effects meta-analysis. *Biostatistics*, 22(1):114–130, 2021.

Giovanni Domenico Tebala. What is the future of biomedical research? *Medical hypotheses*, 85(4):488–490, 2015.

Christopher G Thompson and Betsy Jane Becker. A group-specific prior distribution for effect-size heterogeneity in meta-analysis. *Behavior research methods*, 52(5), 2020.

Kristian Thorlund, PJ Devereaux, Jørn Wetterslev, Gordon Guyatt, John PA Ioannidis, Lehana Thabane, Lise-Lotte Gluud, Bodil Als-Nielsen, and Christian Gluud. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International journal of epidemiology*, 38(1):276–286, 2009.

Kristian Thorlund, Georgina Imberger, Michael Walsh, Rong Chu, Christian Gluud, Jørn Wetterslev, Gordon Guyatt, Philip J Devereaux, and Lehana Thabane. The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis—a simulation study. *PloS one*, 6(10):e25491, 2011.

Rebecca M Turner, Jonathan Davey, Mike J Clarke, Simon G Thompson, and Julian PT Higgins. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the cochrane database of systematic reviews. *International journal of epidemiology*, 41(3): 818–827, 2012.

Rebecca M Turner, Sheila M Bird, and Julian PT Higgins. The impact of study size on meta-analyses: examination of underpowered studies in cochrane reviews. *PloS one*, 8(3): e59202, 2013.

Robbie CM van Aert and Dan Jackson. Multistep estimators of the between-study variance: The relationship with the paule-mandel estimator. *Statistics in medicine*, 37(17):2616–2629, 2018.

Sara van Erp, Josine Verhagen, Raoul PPP Grasman, and Eric-Jan Wagenmakers. Estimates of between-study heterogeneity for 705 meta-analyses reported in psychological bulletin from 1990–2013. *Journal of Open Psychology Data*, 5(1), 2017.

Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian PT Higgins, Dean Langan, and Georgia Salanti. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, 7(1):55–79, 2016.

Wolfgang Viechtbauer. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293, 2005.

Wolfgang Viechtbauer and Maintainer Wolfgang Viechtbauer. Package 'metafor'. *The Comprehensive R Archive Network. Package 'metafor'. http://cran. r-project. org/web/packages/metafor/metafor. pdf*, 2015.

Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. Moving to a world beyond "$p < 0.05$", 2019.

George N Watson. A note on gamma functions. *Edinburgh Mathematical Notes*, 42:7–9, 1959.

Jørn Wetterslev, Kristian Thorlund, Jesper Brok, and Christian Gluud. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of clinical epidemiology*, 61(1):64–75, 2008.

Anne Whitehead. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in medicine*, 16(24):2901–2913, 1997.

Donald R Williams, Josue E Rodriguez, and Paul-Christian Bürkner. Putting variation into variance: Modeling between-study heterogeneity in meta-analysis. *Available on PsyArXiv*, 2021. URL `file:///C:/Users/ali-h/Downloads/pre_print%20(3).pdf`.

Xiaohuan Xue. Improved approximations of hedges'g. *arXiv preprint arXiv:2003.06675*, 2020.

Joanne Yaffe. From the editor—do we have a replication crisis in social work research? *Journal of Social Work Education*, 55(1):1–4, 2019.

Appendix A

Additional Material for Meta-Analysis

A.1 Methods for Estimating the Between-Study Variance

There are 20 methods for estimating the between-study variance τ^2 in meta-analysis publication, and the majority of these estimators are based on the method of moments. A complete list of methods is given:

Table A.1: *Overview of the Estimators for the Between-Study Variance*

Estimator	**Abbreviations**
DerSimonian and Laird	DL
Positive DerSimonian and Laird	DL_p
Two-step DerSimonian and Laird	DL_2
Cochran's ANOVA	CA
Two-step Cochran's ANOVA	PM_{CA}
Paule and Mandel	PM
Hartung and Makambi	HM

(To be continued)

Estimator	**Abbreviations**
Hunter and Schmidt	HS
Maximum likelihood	ML
Restricted maximum likelihood	$REML$
Approximate restricted maximum likelihood	$AREML$
Sidik and Jonkman	SJ
Sidik-Jonkman (CA initial estimate)	SJ_{CA}
Bayes estimators Rukhin Bayes	RB
Positive Rukhin Bayes	RB_p
Bayes Modal	BM
Bootstrap DerSimonian-Laird	DL_B
Malzahn, Böhning and Holling	MBH
Rukhin (zero prior)	$B0$
Rukhin (simple)	BP
Rukhin (alternate)	SB
Full Bayes	FB
Approximate Bayes	AB

A.2 Large Sample Size Outcomes

A.2.1 Hedges Approximation (J)

Because Cohen's d is biased when the true SMD $\neq 0$, Hedges (1982, 1983) proposed a small-sample bias-corrected estimator, J. Because J is a special case of Wallis ratio when $x = \frac{v}{2} - 1$, the Wallis ratio will be used to prove that J converges to 1 when $v \to \infty$, where

$\upsilon = n_1 + n_2 - 2$. The Wallis ratio is given as

$$\frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})} \tag{A.1}$$

In particular about the Wallis ratio, Watson (1959) proposed the following inequality

$$\sqrt{x+\frac{1}{4}} < \frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})} \leqslant \sqrt{x+\frac{1}{\pi}} \quad , \quad x \geqslant 0 \tag{A.2}$$

then by letting $x = \frac{\upsilon}{2} - 1$, we get

$$\sqrt{\frac{\upsilon}{2} - 1 + \frac{1}{4}} < \frac{\Gamma(\frac{\upsilon}{2} - 1 + 1)}{\Gamma(\frac{\upsilon}{2} - 1 + \frac{1}{2})} \leqslant \sqrt{\frac{\upsilon}{2} - 1 + \frac{1}{\pi}}$$

$$\sqrt{\frac{\upsilon}{2} - \frac{3}{4}} < \frac{\Gamma(\frac{\upsilon}{2})}{\Gamma(\frac{\upsilon-1}{2})} \leqslant \sqrt{\frac{\upsilon-2}{2} + \frac{1}{\pi}}$$

$$\sqrt{\frac{2\upsilon-3}{4}} < \frac{\Gamma(\frac{\upsilon}{2})}{\Gamma(\frac{\upsilon-1}{2})} \leqslant \sqrt{\frac{\pi(\upsilon-2)+2}{2\pi}}$$

now, divide it by $\sqrt{\frac{\upsilon}{2}}$

$$\sqrt{\frac{2\upsilon-3}{4}} \Bigg/ \sqrt{\frac{\upsilon}{2}} < \frac{\Gamma(\frac{\upsilon}{2})}{\sqrt{\frac{\upsilon}{2}}\Gamma(\frac{\upsilon-1}{2})} \leqslant \sqrt{\frac{\pi(\upsilon-2)+2}{2\pi}} \Bigg/ \sqrt{\frac{\upsilon}{2}}$$

$$\sqrt{\frac{2(2\upsilon-3)}{4\upsilon}} < J \leqslant \sqrt{\frac{2\pi(\upsilon-2)+4}{2\pi\upsilon}} \tag{A.3}$$

Therefore, when $\upsilon \to \infty$, $J \to 1$.

A.2.2 Within-Study Variance

Using properties of the non-central t-distribution, the exact variance of the within-study variance of $\hat{\delta}_{H.j}$ conditional on the study-level effect size δ_j is

$$\sigma_j^2 = Var(\hat{\delta}_{H.j}|\delta_j) = \frac{J_j^2\ (n_{1j} + n_{2j} - 2)\ (\tilde{n}_j + \delta_j^2)}{(n_{1j} + n_{2j} - 4)} - \delta_j^2 \tag{A.4}$$

where $\tilde{n_j} = 1/n_{1j} + 1/n_{2j}$

$$\sigma_j^2 = \frac{J_j^2\ (n_{1j} + n_{2j} - 2)\ \tilde{n}_j}{(n_{1j} + n_{2j} - 4)} + \delta_j^2 \left(\frac{J_j^2\ (n_{1j} + n_{2j} - 2)}{(n_{1j} + n_{2j} - 4)} - 1 \right)$$

$$= \frac{(n_{1j} + n_{2j} - 2)}{(n_{1j} + n_{2j} - 4)} \left(J_j^2\ \tilde{n}_j + \frac{J_j^2\ \delta_j^2}{(n_{1j} + n_{2j} - 2)} \left((n_{1j} + n_{2j} - 2) - \frac{(n_{1j} + n_{2j} - 4)}{J_j^2} \right) \right)$$

$$= \frac{n_{1j} + n_{2j} - 2}{n_{1j} + n_{2j} - 4} \left(J_j^2\ \tilde{n}_j + \frac{\gamma_j\ J_j^2\ \delta_j^2}{(n_{1j} + n_{2j} - 2)} \right) \tag{A.5}$$

where

$$\gamma_j = (n_{1j} + n_{2j} - 2) - \frac{(n_{1j} + n_{2j} - 4)}{J_j^2}. \tag{A.6}$$

with large sample sizes, we have that

$$\frac{(n_{1j} + n_{2j} - 2)}{(n_{1j} + n_{2j} - 4)} \rightarrow 1$$

$$(n_{1j} + n_{2j} - 2) \approx (n_{1j} + n_{2j})$$

In Section A.2.1, we showed that $J_j \rightarrow 1$ with large sample size; therefore, $J_j^2 \rightarrow 1$. Similarly, after simplifying γ_j by using Hedges approximation of J $\Big(1 - 3/\big(4(n_{1j} + n_{1j} - 2) - 1\big)\Big)$ and considering large sample size, $\gamma_j \rightarrow 0.5$. Therefore, in practice, the most commonly-used estimator of σ_j^2 is

$$s_j^2 = \tilde{n}_j + \frac{\hat{\delta}_{H.j}^2}{2\ (n_{1j} + n_{2j})} \tag{A.7}$$

Appendix B

Simulation-Based Comparison of Estimators for Meta-Analysis and Mixed Model for Chapter 2

The plan of these simulations is to find the better approach (meta-analysis or mixed model) that researchers should use when conducting multi-laboratory replication studies (MLRS), see meta-analysis simulations in section B.1, and mixed model simulations in section B.2. The same data sets were generated in both methods.

Under meta-analysis, the simulations for all methods (REML, PM and ML) are carried out in R using our own code. 5,760 meta-analysis scenarios are conducted to perform a comprehensive comparison, but only a subset is presented. We use 10,000 repetitions for each combination. The performance of REML, PM, and ML is assessed through the heterogeneity variance and the effect on the overall effect estimate and Type 1 error. All simulations can be found in Sections B.1.1 and B.1.2.

To perform a comprehensive comparison under mixed model, 80 mixed model scenarios are conducted. All simulations can be found in section B.2.

B.1 Meta Analysis

B.1.1 Simulation Studies

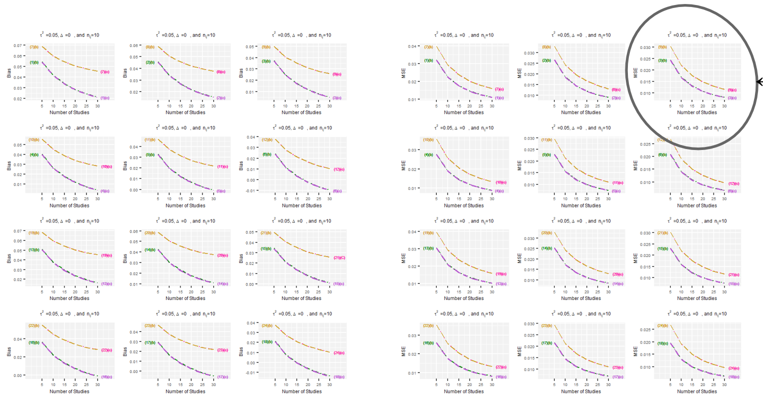
Our simulations focus on both between-study variance and the overall effect. In this section, there are 5,760 different simulations.

$$3\ (Method) \times 24\ \big((1) - (24)\big) \times 2\ \big((b), (c)\big) \times 5\ (\tau^2) \times 4\ (N) \times 2\ (\Delta) = 5,760$$

Because the heterogeneity variance estimators ($\hat{\tau}^2$) and the overall effect size ($\hat{\Delta}$) rely on s_j^2, we decide to make the comparison based on s_j^2, for further explanation see Section 2.5.3. For all combinations of the parameter value, we compare heterogeneity variance estimators in terms of bias, MSE, and proportion of zero estimates, and overall Effect Size is also compared based on bias, MSE, and Type 1 Error. It can be challenging to provide all simulations (5,760 scenarios); therefore, only a subset is provided (a subset of simulations with small sample sizes is provided), and all comparisons are shown in Section B.1.2.

Explanation of the Figure:

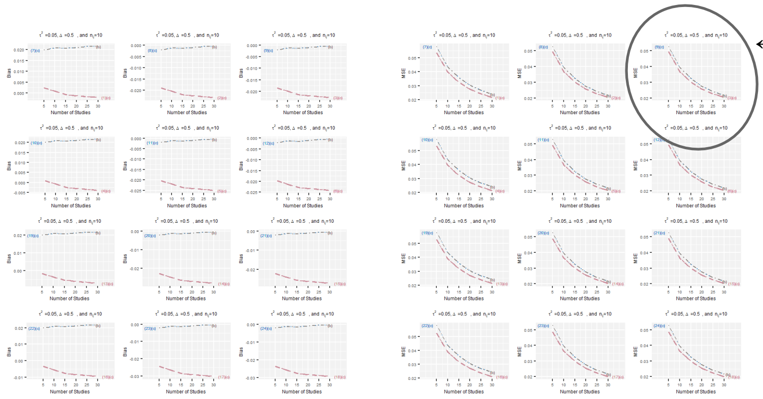
Under between-study variance, in term of bias, MSE, and proportion of zero, the figures for each simulation will be represented as following:



Each plot represents 4 comparisons. (c) method with two different s_j^2: (the difference only on the effect size term); one has individual effect sizes and the second has summary effect sizes, similar with (b) method. See Section 2.5.3 for further details.

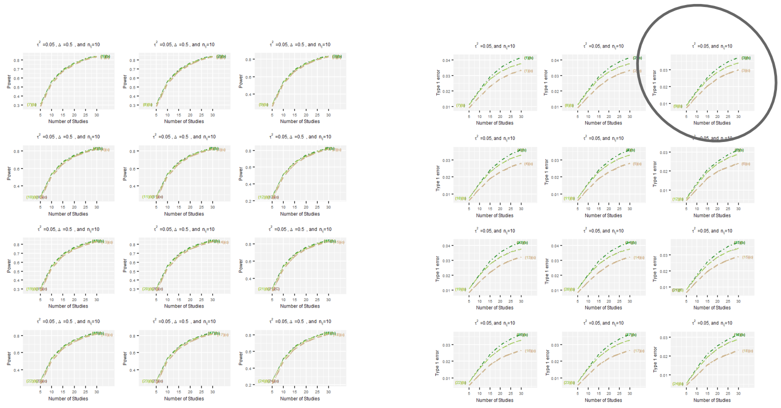
The total number of comparisons for each simulation is 48; there are 40 different scenarios.

Regarding bias and MSE, the bias is on the left side (3 columns), and the mean squared error is on the right side (3 columns). For the proportion of zero, the left-hand-side (3 columns) is under no effect size, and the right-hand-side (3 columns) is under medium effect size. Regarding overall effect size in terms of bias, and MSE, the figures for each simulation will be represented as follows:



⟵ Each plot represents 3 comparisons. (c) method with two different s_j^2: (the difference only on the effect size term). One has individual effect sizes and the second has summary effect sizes. each time, we use (c) method by Hedges estimator (Cohen d estimator), we will use (b) method with Hedges estimator (Cohen d estimator). See Section 2.5.3 for further details.

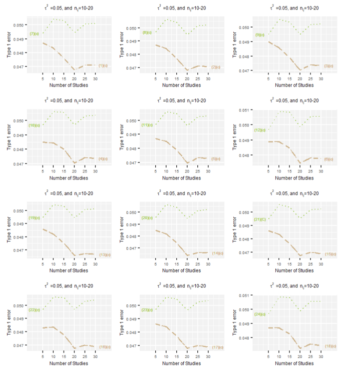
The total number of comparisons for each simulation is 26, 40 different scenarios. The bias is on the left side (3 columns), and the mean squared error is on the right side (3 columns). Regarding type 1 error and power, the figures for each simulation by t distribution will be represented as follows:

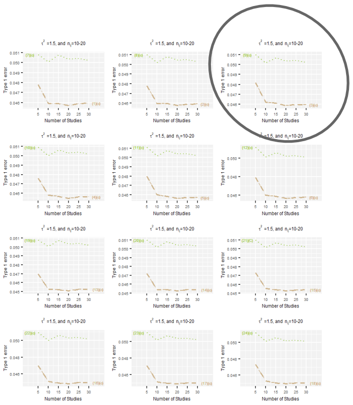


⟵ Each plot represents 4 comparisons. (c) method with two different s_j^2: (the difference only on the effect size term). One has individual effect sizes and the second has summary effect sizes, similar with (b) method. See Section 2.5.3 for further details.

The total number of comparisons for each simulation is 48; there are 40 different scenarios. The power is on the left side (3 columns), and the type I error is on the right side (3 columns).

Regarding type I error Hartung-Knapp t-distribution, the figures for each simulation will be represented as follows:



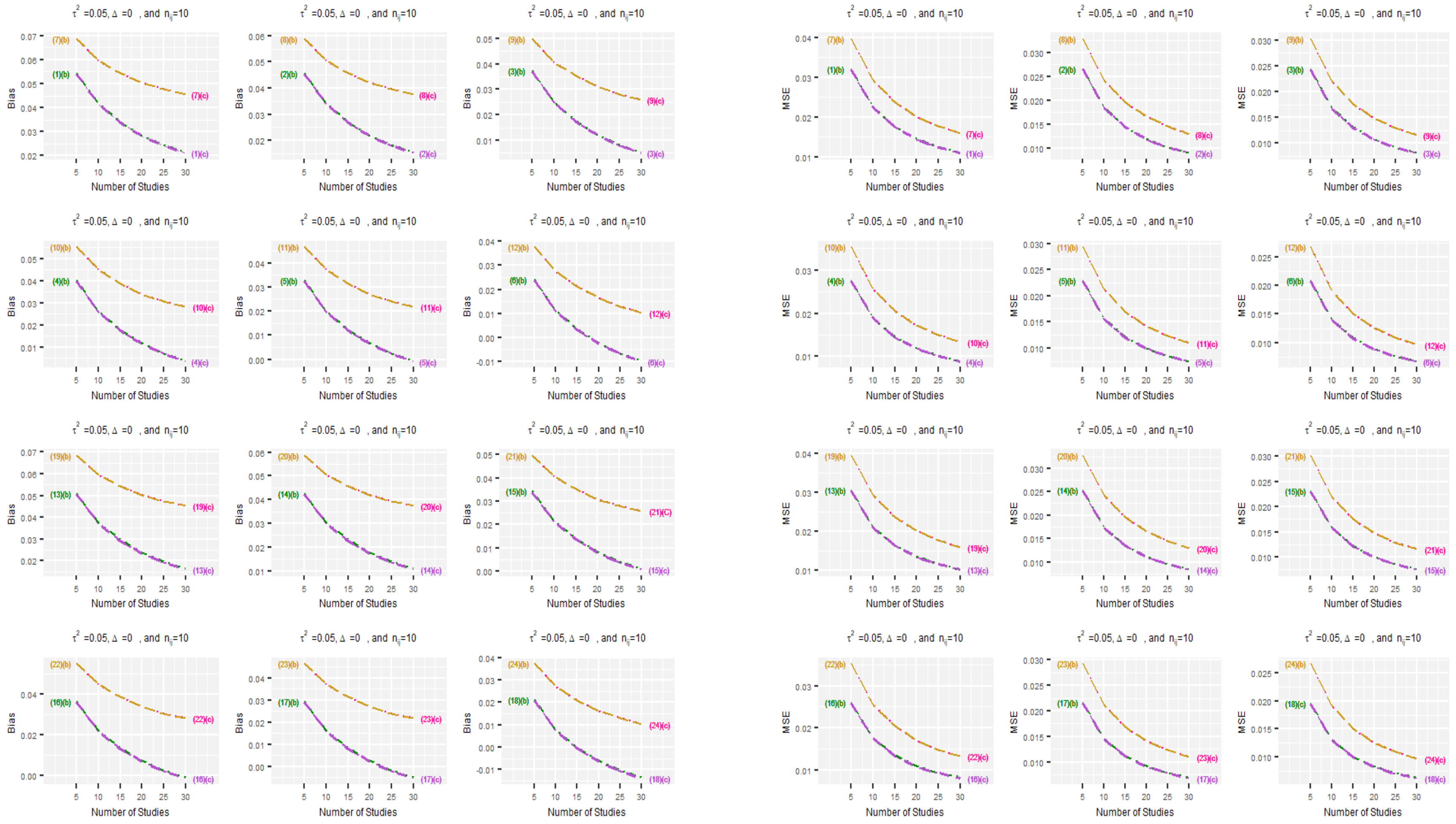


⟵ Each plot represents 2 comparisons. (c) method with two different s_j^2: (the difference only on the effect size term). One has individual effect sizes and the second has summary effect sizes, ee Section 2.5.3 for further details.

The total number of comparisons for each simulation is 24; there are 40 different scenarios.

The numbers ((1)-(24)) beside the lines indicate the numbers of equations of within-study variances, see Table 2.2. The letters ((b)-(c)) indicate the type of averaging overall effect size: (b) indicates the unweighted method, and (c) indicates the inverse variance method. Section 2.5.2 describes how these estimators $\big($ ((1)-(24)) and ((b)-(c)) $\big)$ are applied in ML, REML and PM methods. The τ^2 indicates heterogeneity variance, Δ indicate overall effect size, and n_{ij} indicates sample sizes for each group (n_{1j} for treatment 1 and n_{2j} for treatment 2).

Between-Study Variance: REML Method

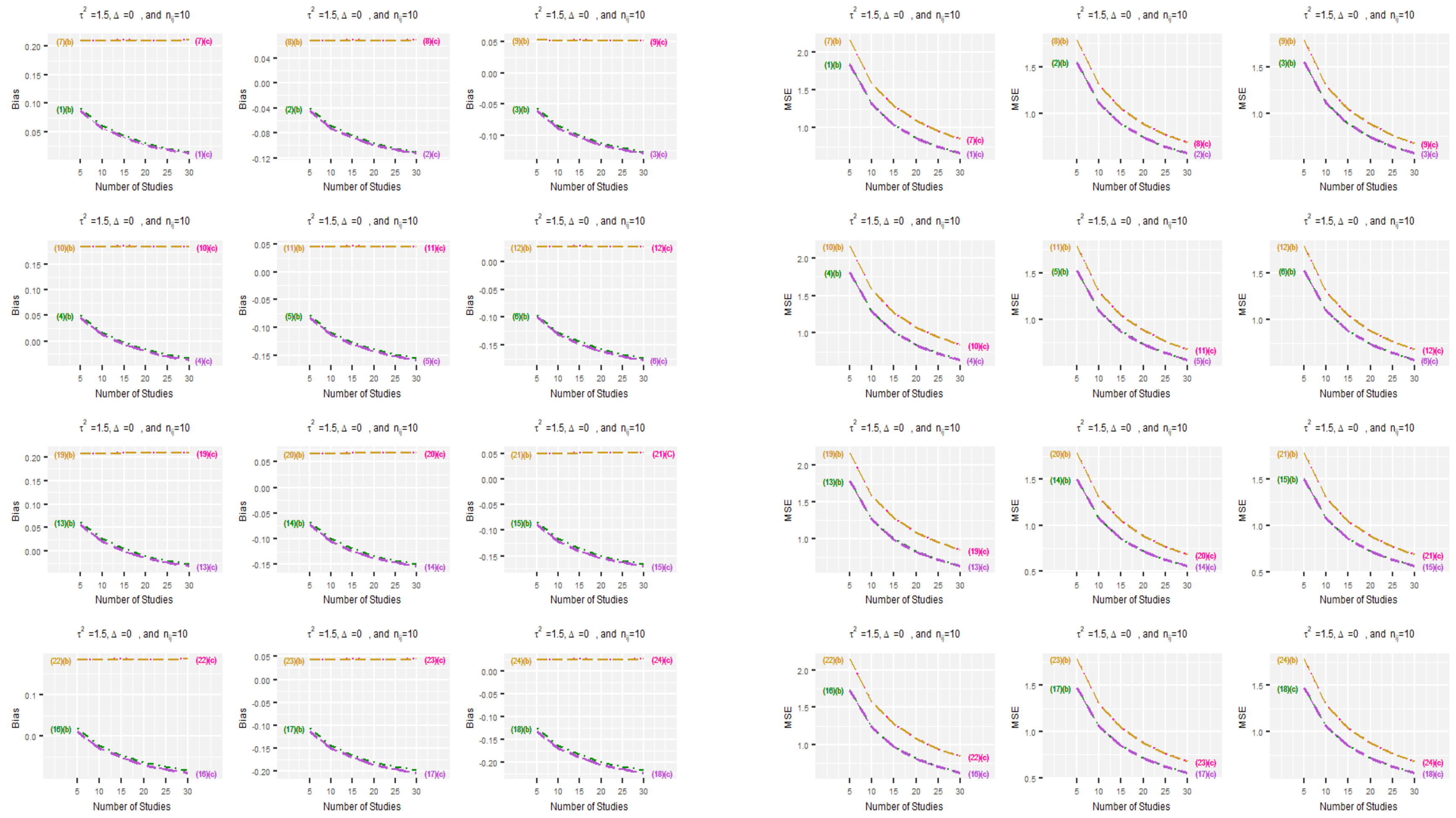


(a) Bias

(b) MES

Figure B.1: *REML Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and very low heterogeneity, for further details see Section* B.1.1.

Between-Study Variance: REML Method

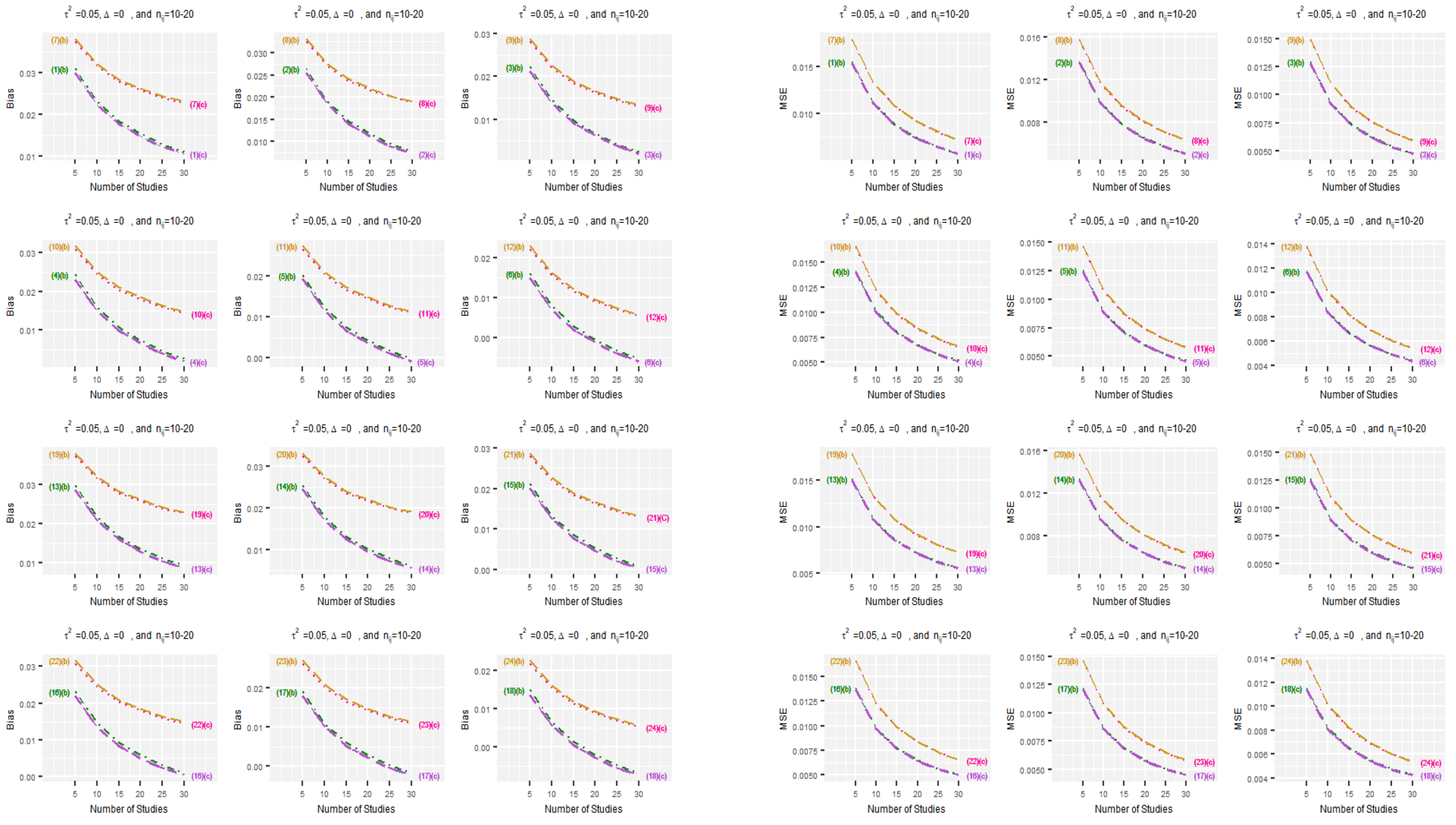


(a) Bias

(b) MES

Figure B.2: *REML Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and large heterogeneity, for further details see Section* B.1.1.

Between-Study Variance: REML Method

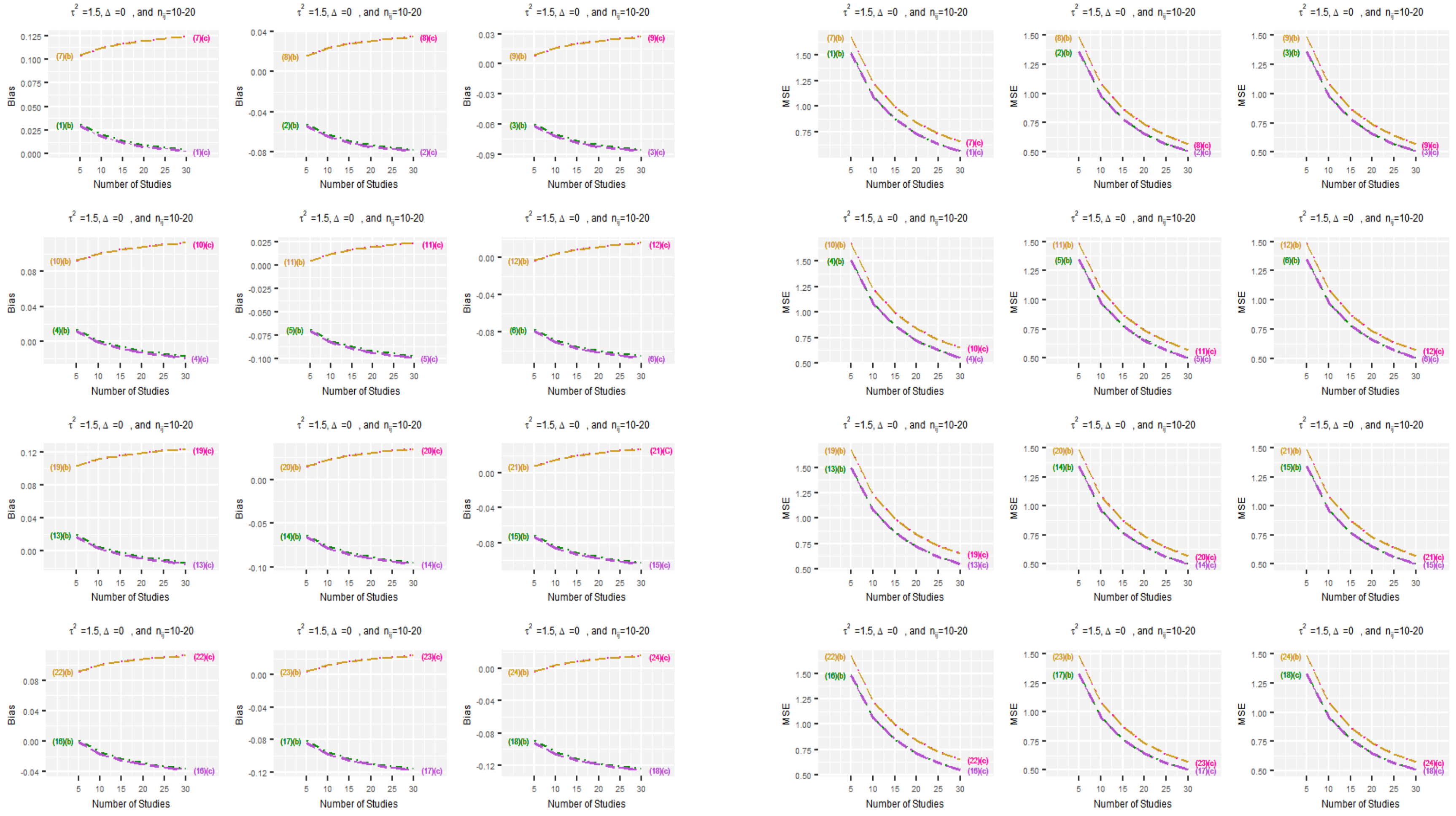


(a) Bias

(b) MES

Figure B.3: *REML Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes (n_{1j} not necessary equal to n_{2j}), no effect size, and very low heterogeneity, for further details see Section B.1.1*

Between-Study Variance: REML Method



(a) Bias

(b) MES

Figure B.4: *REML Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes (n_{1j} not necessary equal to n_{2j}), no effect size, and large heterogeneity, for further details see Section B.1.1.*

Proportion of zero heterogeneity variance estimates: REML Method

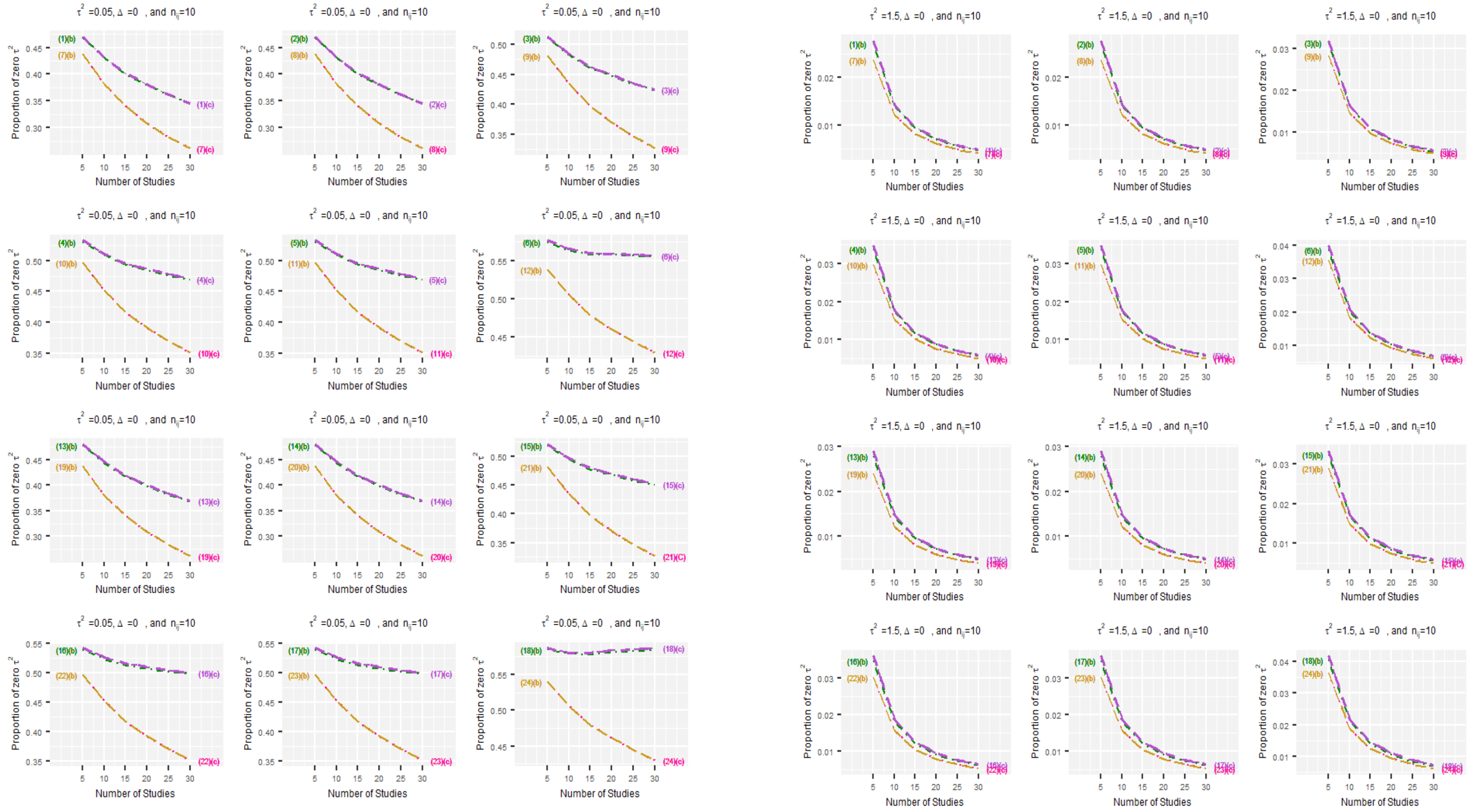


Figure B.5: *REML Method: 48 between-study variance estimators are compared through their proportion of zero with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and very low and large heterogeneity, for further details see Section B.1.1*

Proportion of zero heterogeneity variance estimates: REML Method

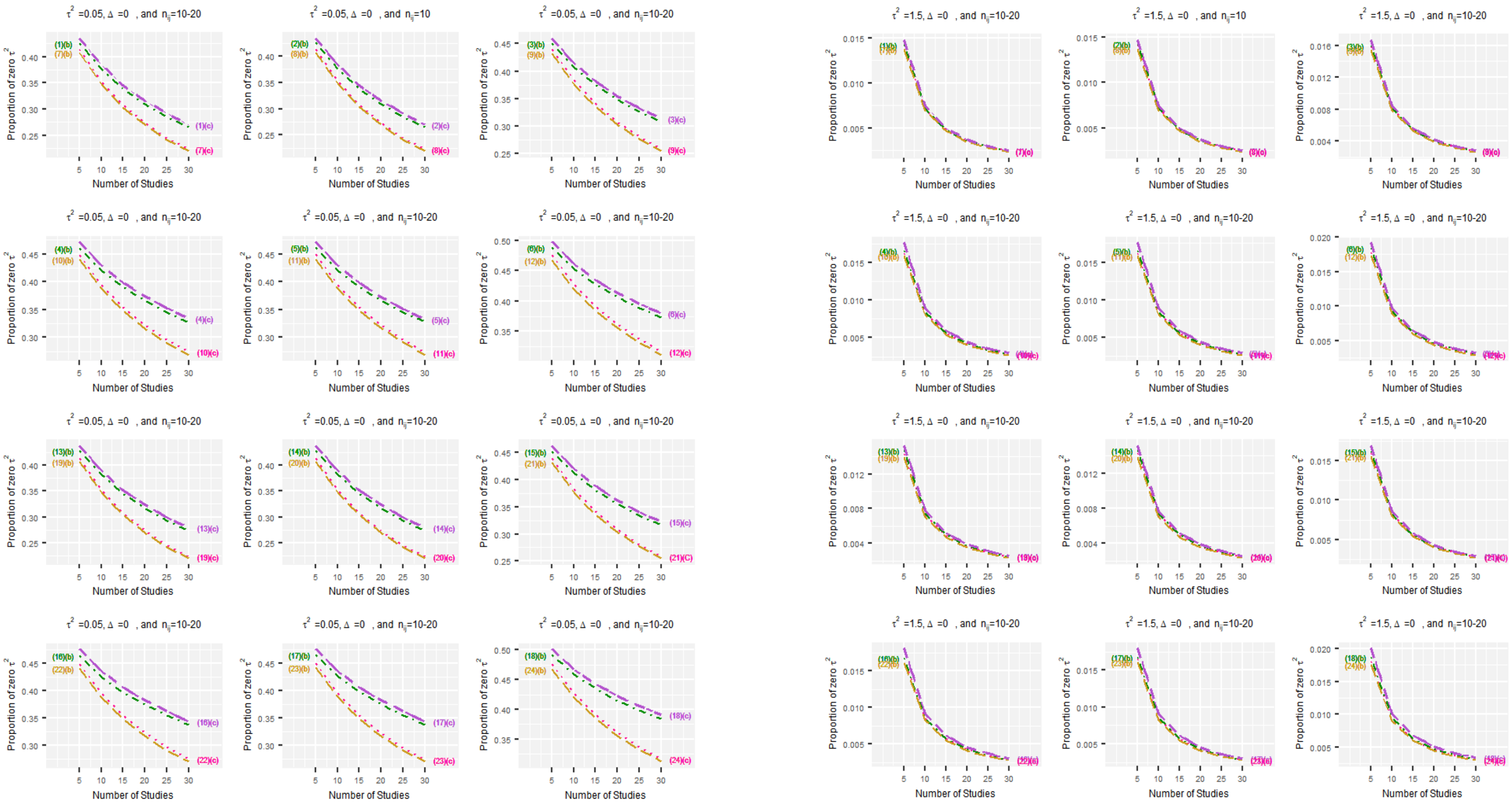
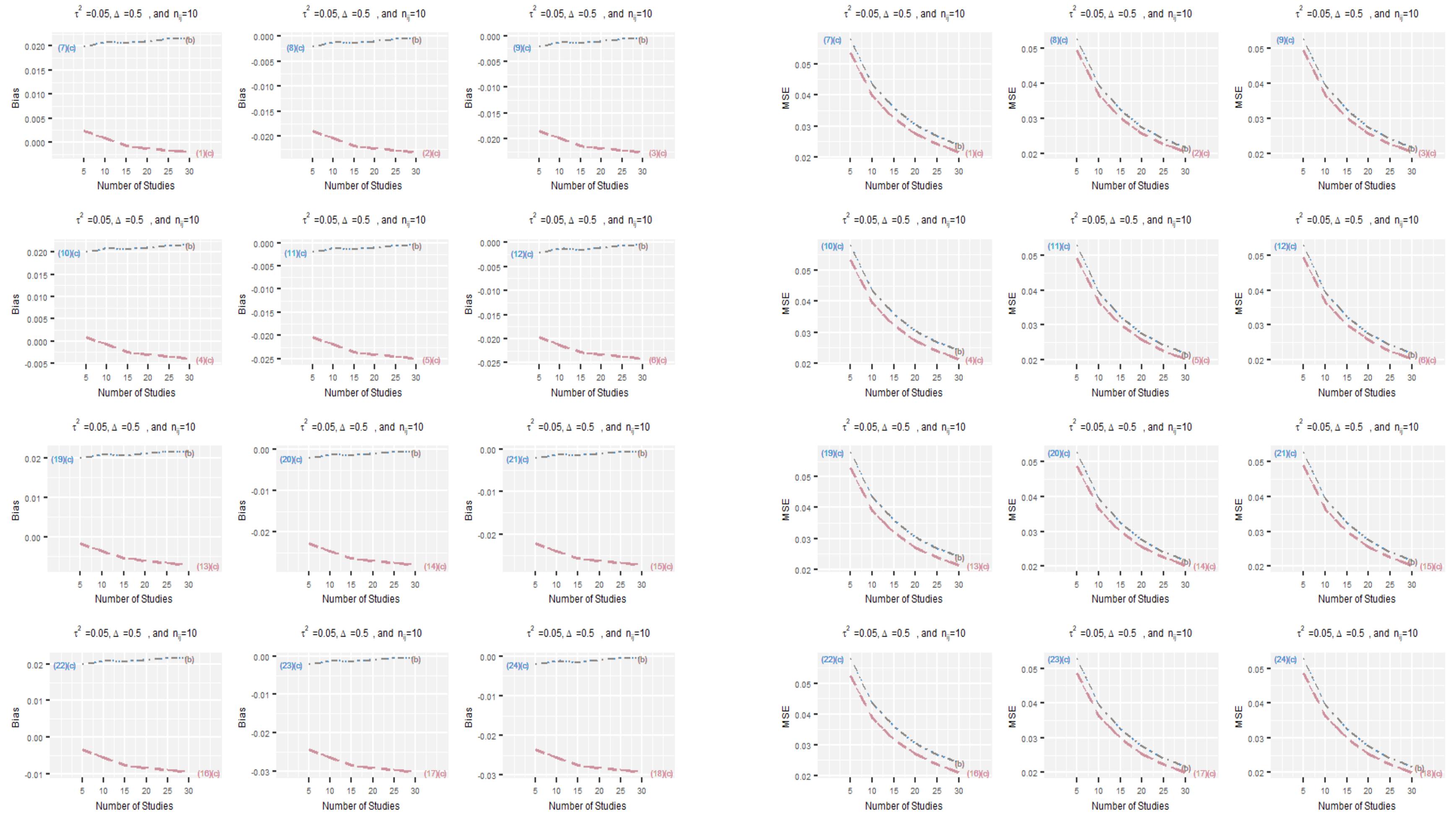


Figure B.6: *REML Method: 48 between-study variance estimators are compared through their proportion of zero with small sample sizes (n_{1j} not necessary equal to n_{2j}), no effect size, and very low and large heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: REML Method

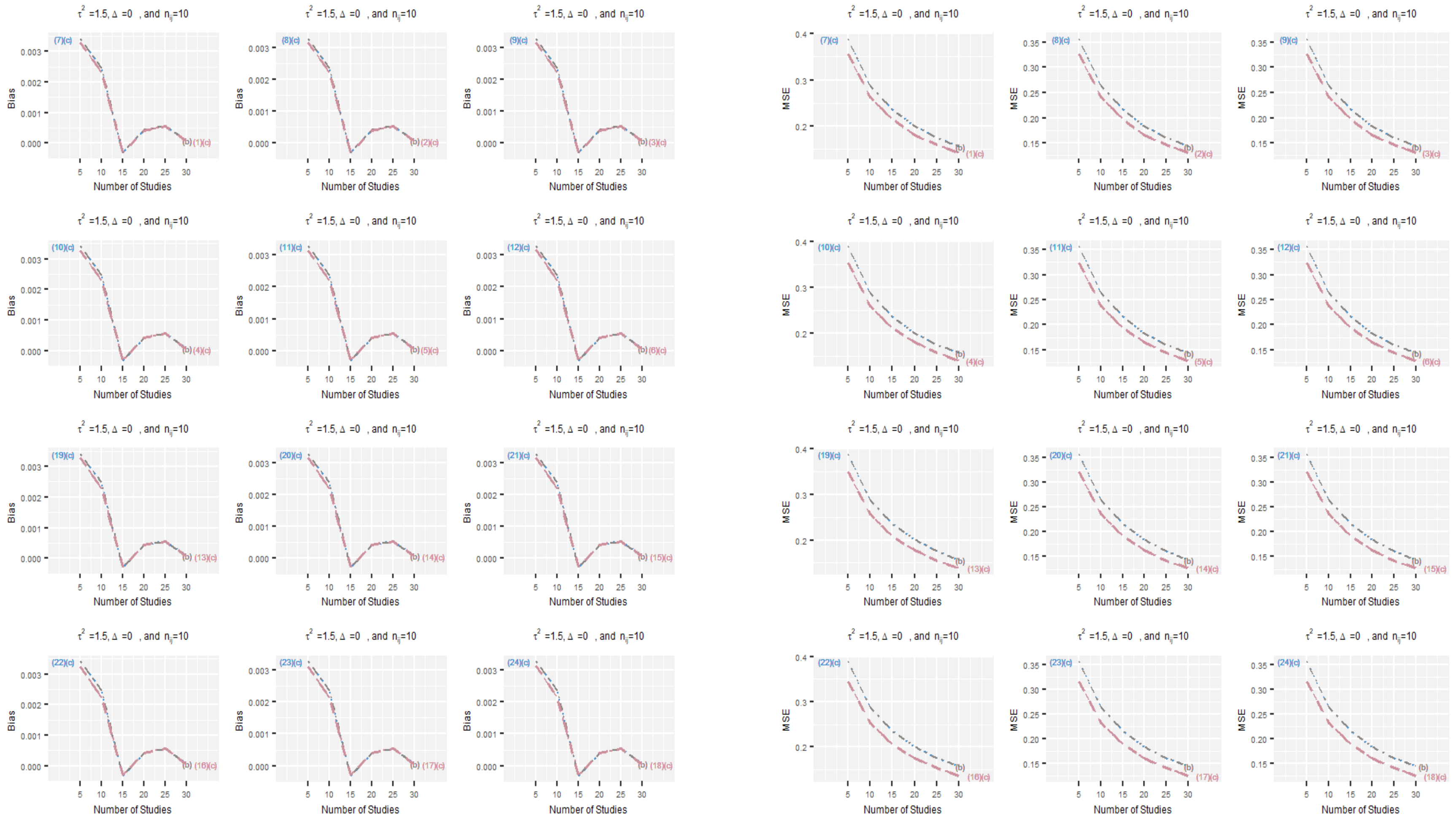


(a) Bias

(b) MES

Figure B.7: *REML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes* ($n_{1j} = n_{2j}$)*, medium effect sizes , and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: REML Method

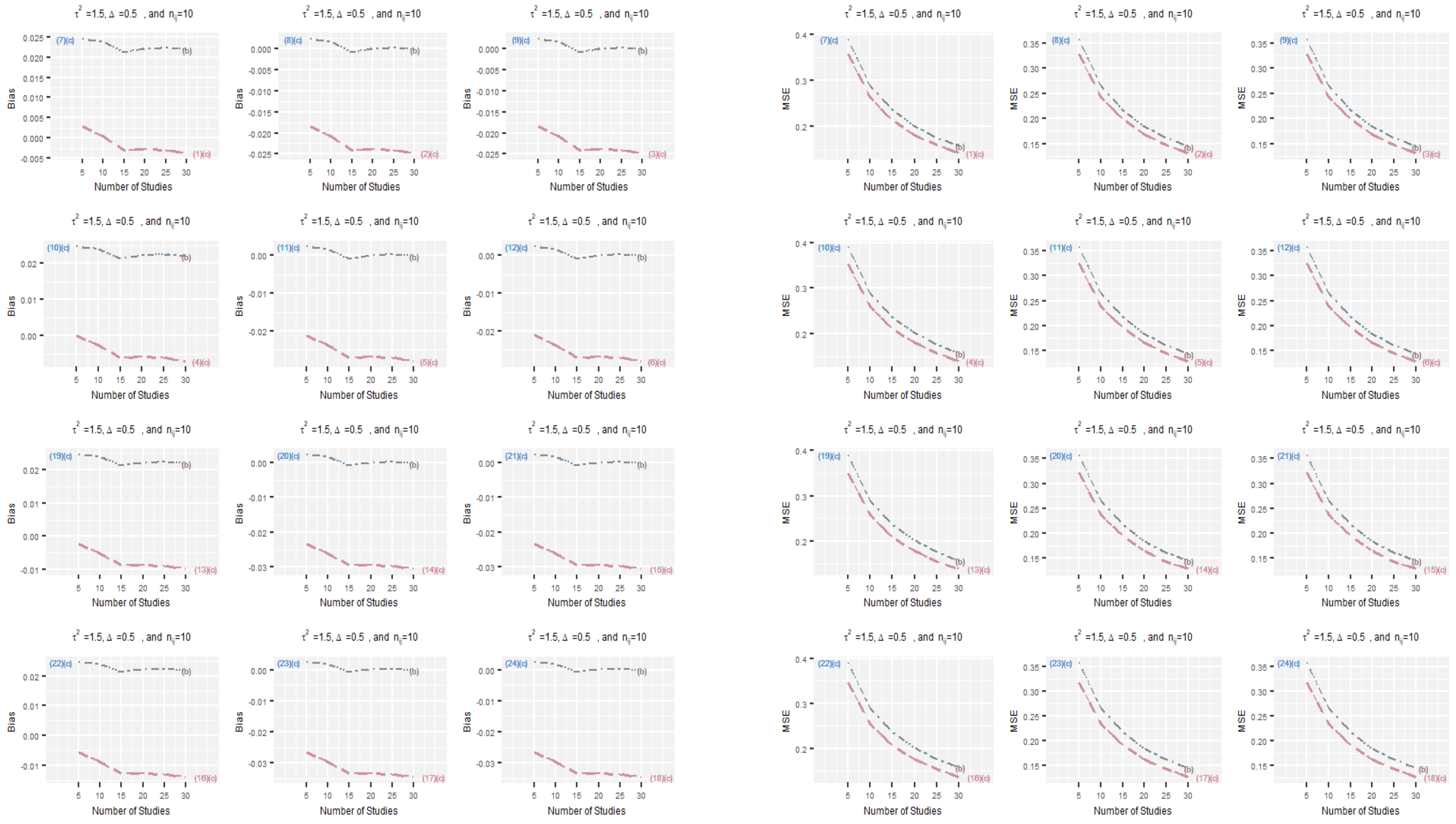


(a) Bias

(b) MES

Figure B.8: *25 overall effect size estimators are compared through their bias and mean squared error by using REML method, with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect sizes , and large heterogeneity, for further details see Section* B.1.1.

Overall Effect Size: REML Method

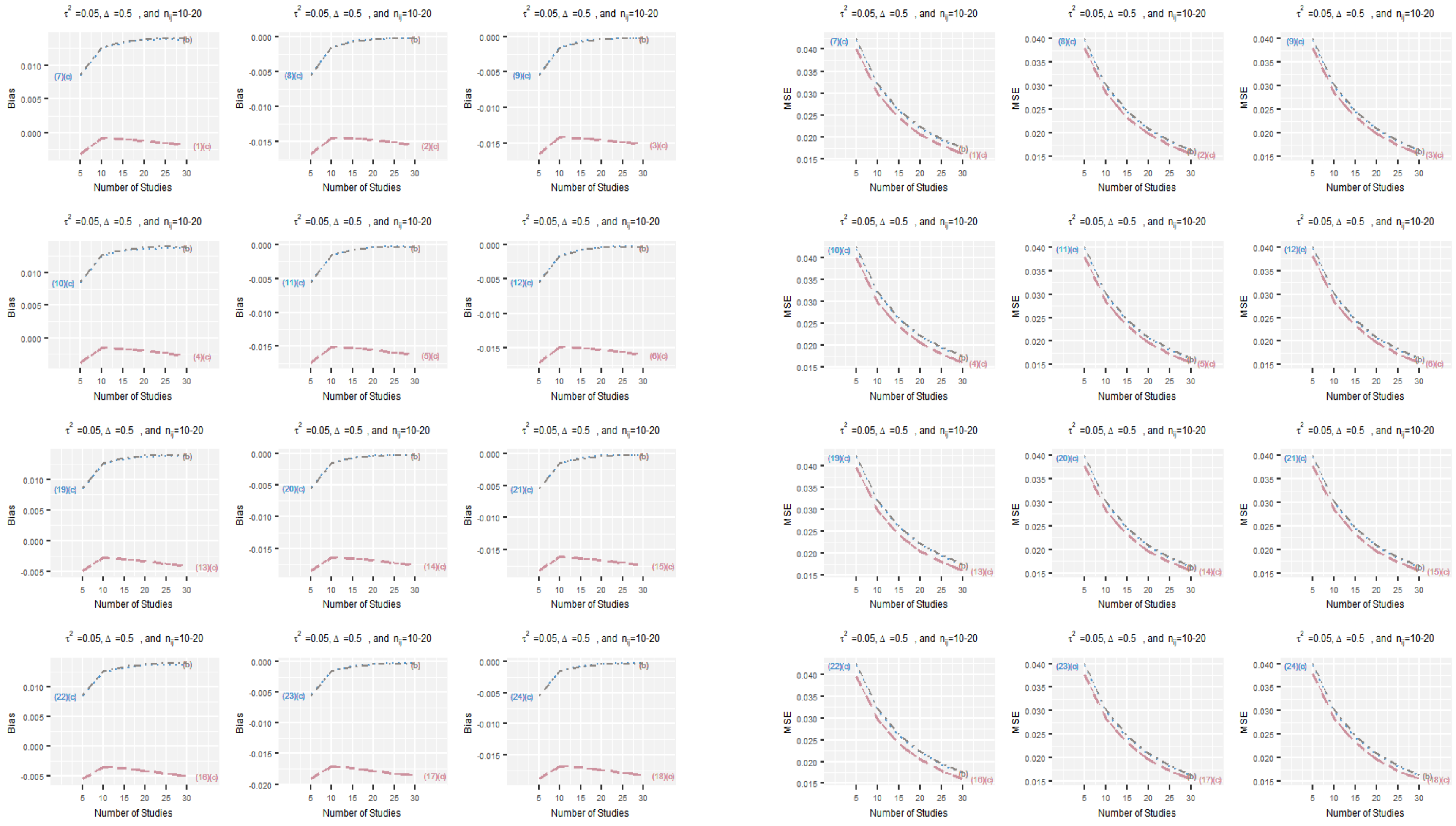


(a) Bias

(b) MES

Figure B.9: *25 overall effect size estimators are compared through their bias and mean squared error by using REML method, with small sample sizes* ($n_{1j} = n_{2j}$)*, medium effect sizes , and large heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: REML Method

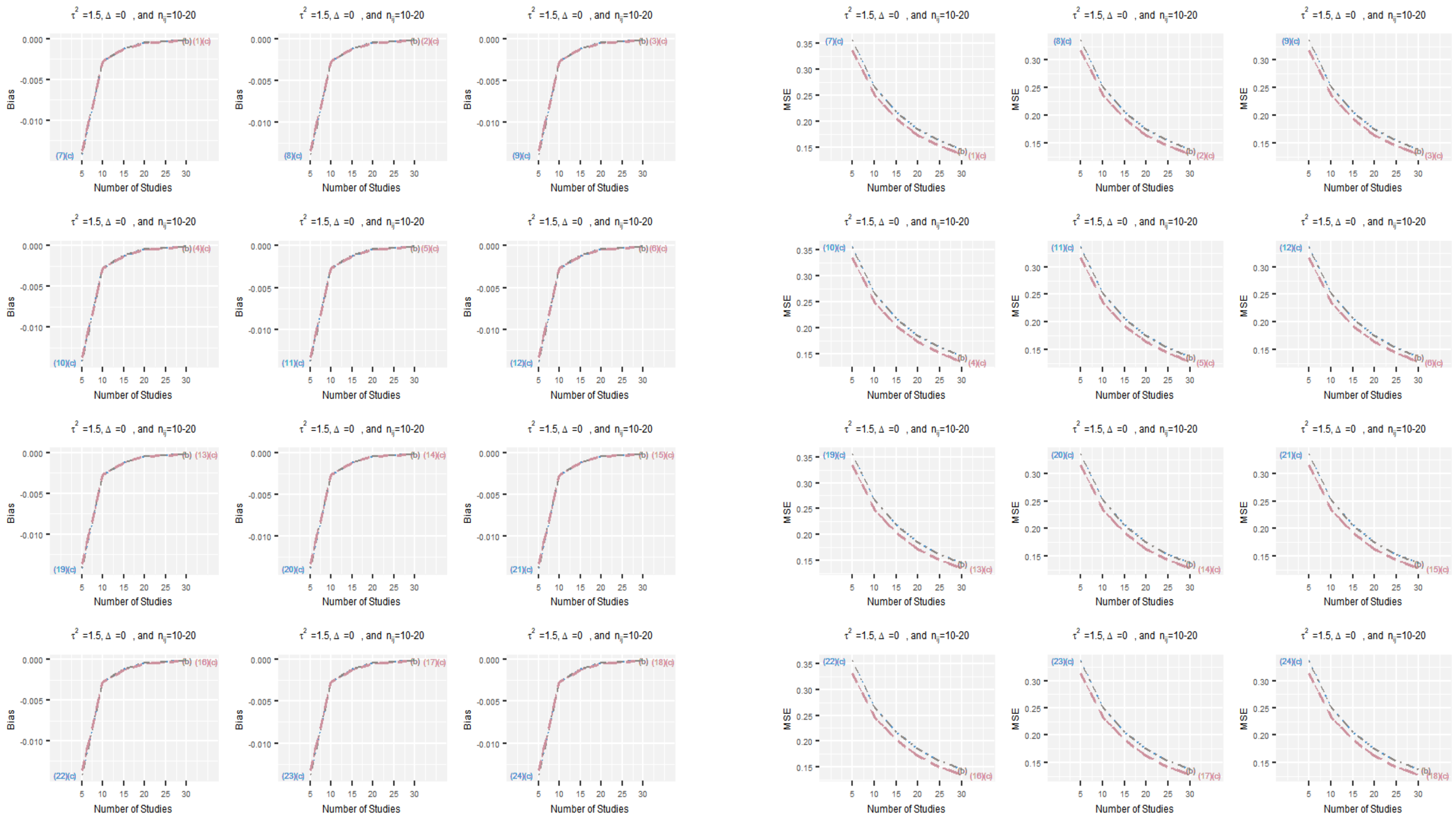


(a) Bias

(b) MES

Figure B.10: *REML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes* (n_{1j} *not necessary equal to* n_{2j})*, medium effect size, and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: REML Method

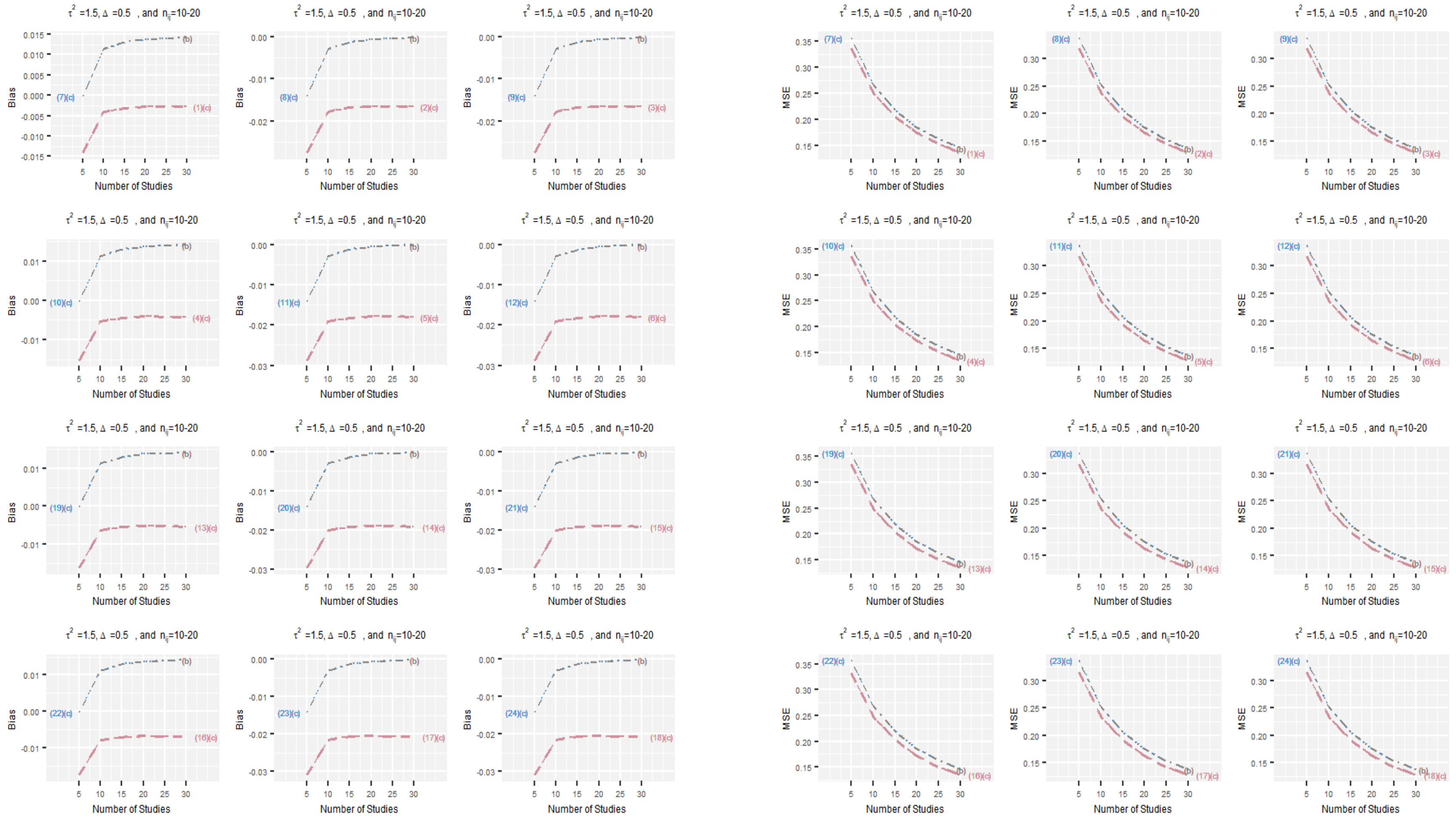


(a) Bias

(b) MES

Figure B.11: *REML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes (n_{1j} not necessary equal to n_{2j}), no effect size, and large heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: REML Method

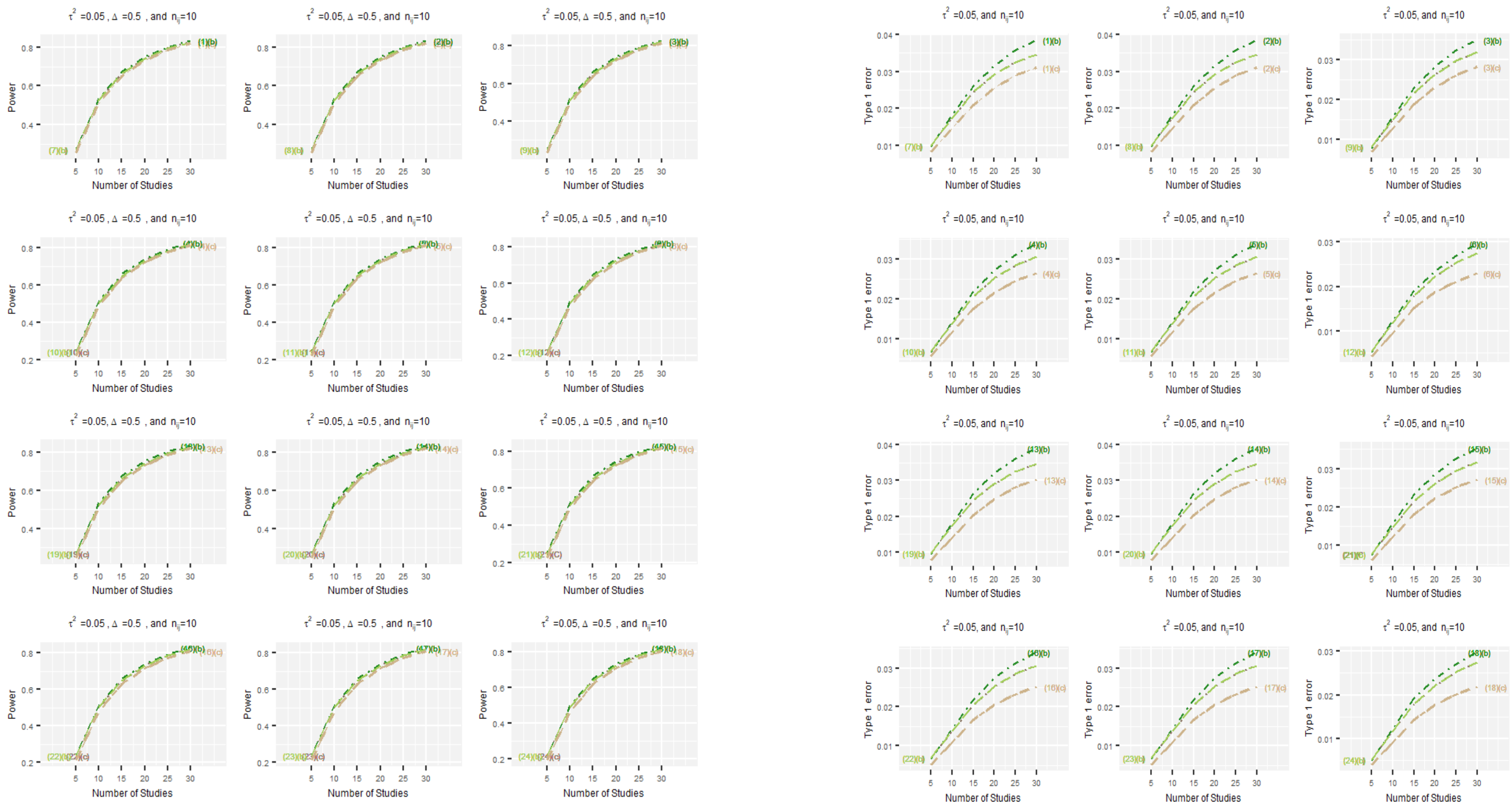


(a) Bias

(b) MES

Figure B.12: *REML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes* (n_{1j} *not necessary equal to* n_{2j})*, medium effect size, and large heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: REML Method

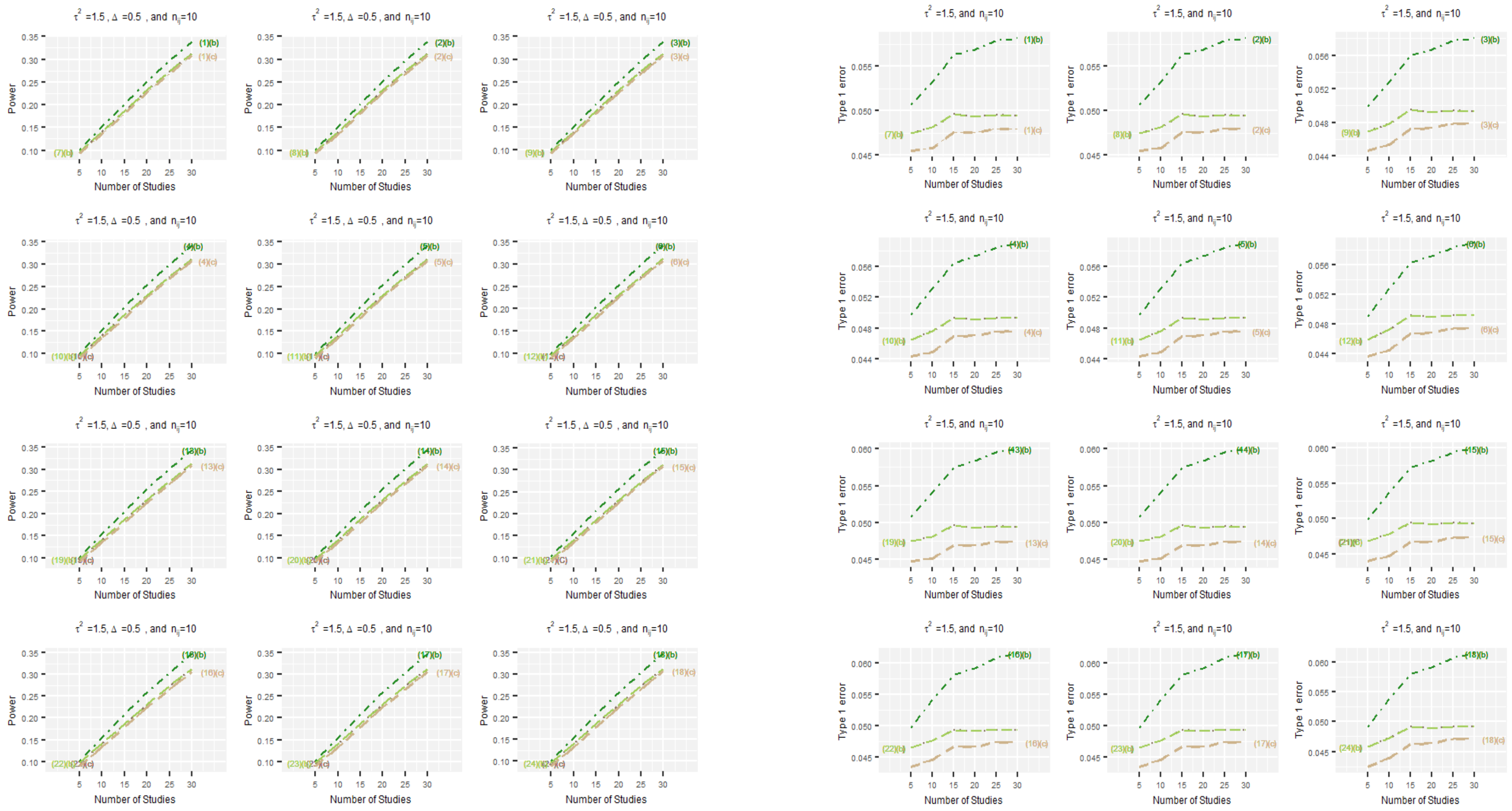


(a) Power

(b) Type 1 Error

Figure B.13: *REML method: Comparing 48 Power and Type I error by t distribution, with small sample sizes* ($n_{1j} = n_{2j}$), *and very low heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: REML Method

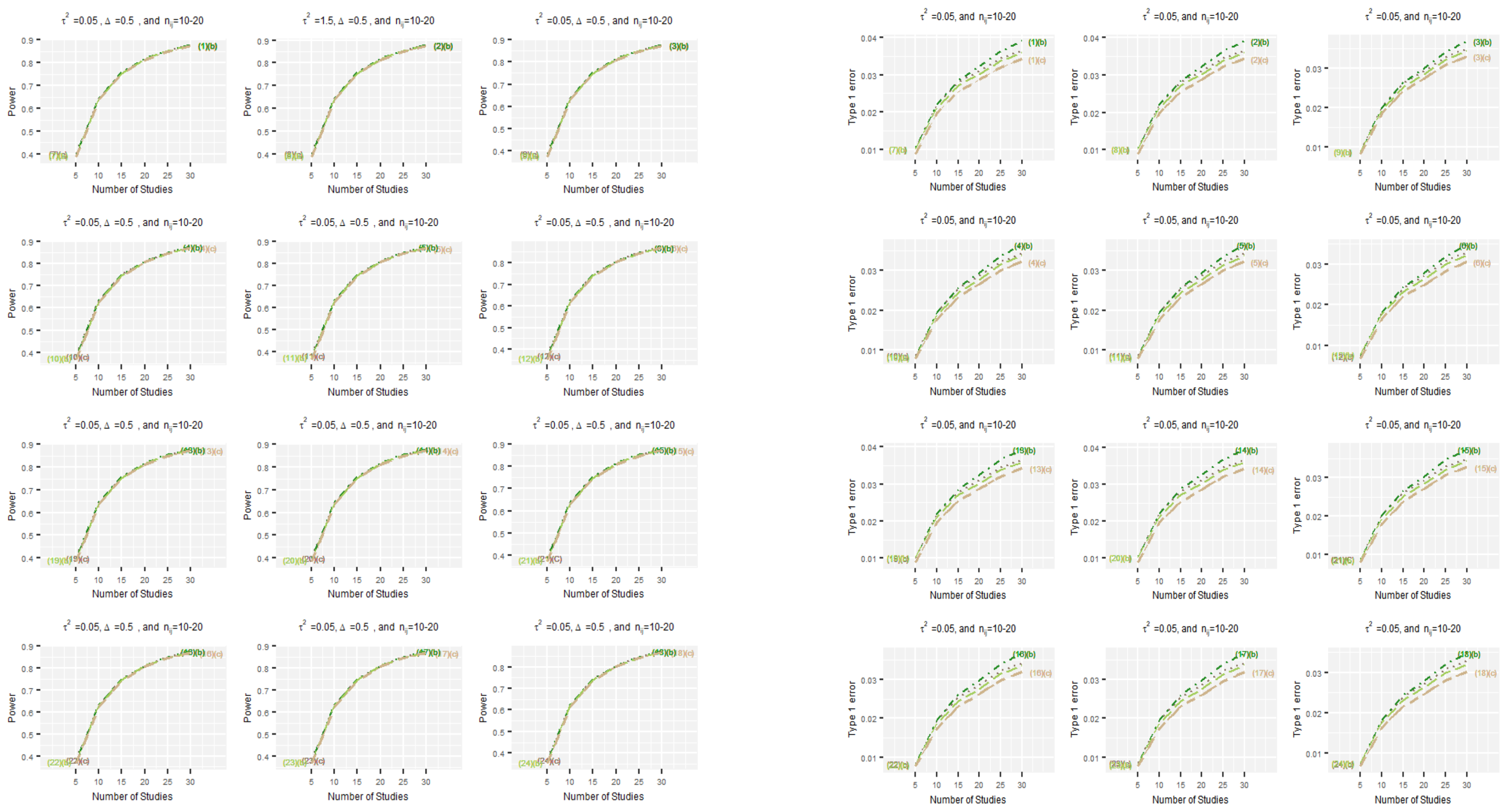


(a) Power

(b) Type 1 Error

Figure B.14: *REML method: Comparing 48 Power and Type I error by t distribution, with small sample sizes* $(n_{1j} = n_{2j})$*, and large heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: REML Method

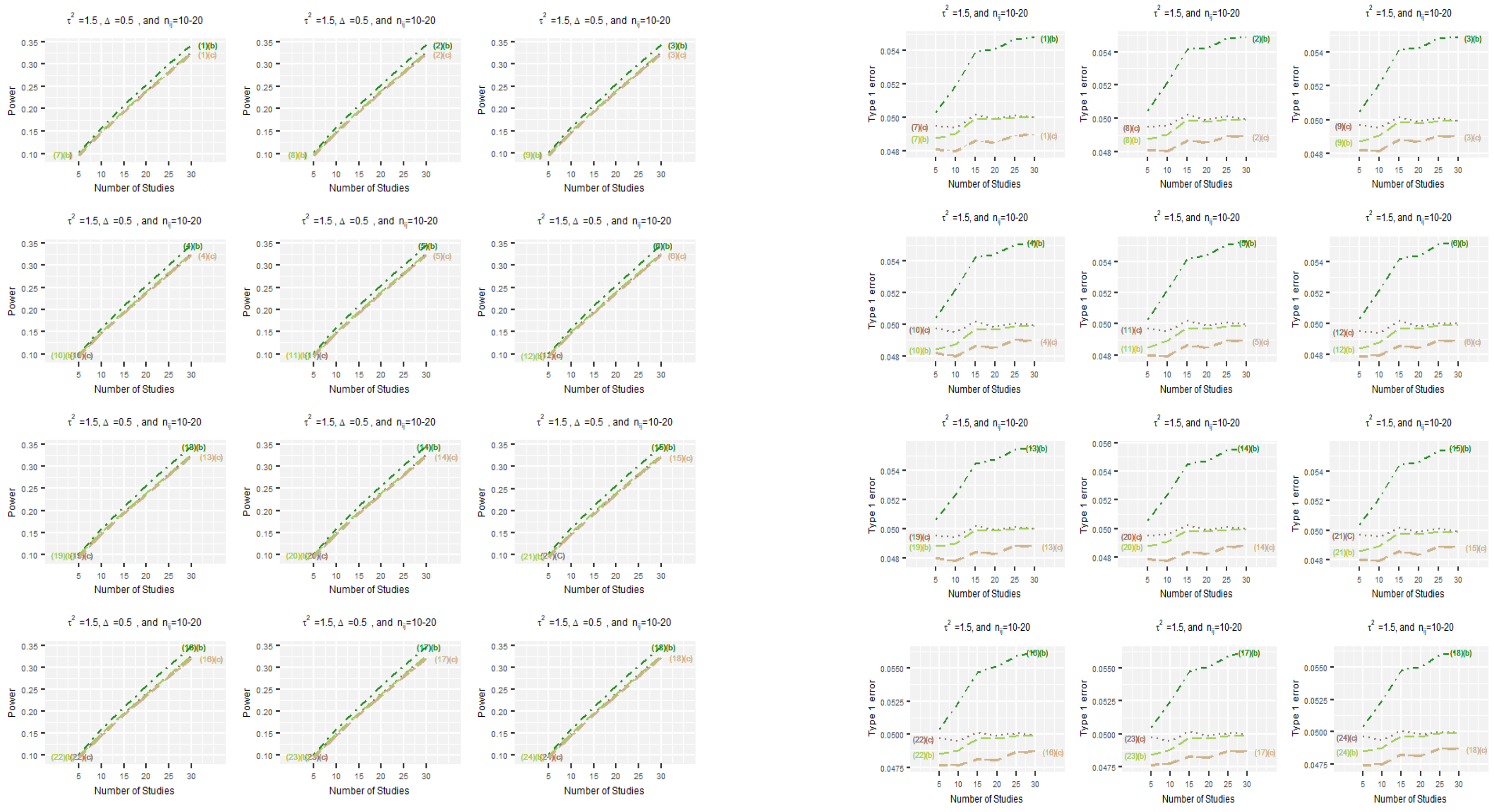


(a) Power

(b) Type 1 Error

Figure B.15: *REML method: Comparing 48 Power and Type I error by t distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), , and very low heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: REML Method



(a) Power

(b) Type 1 Error

Figure B.16: *REML method: Comparing 48 Power and Type I error by t distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), and large heterogeneity.*

Type I Error by Hartung-Knapp t-distribution: REML Method

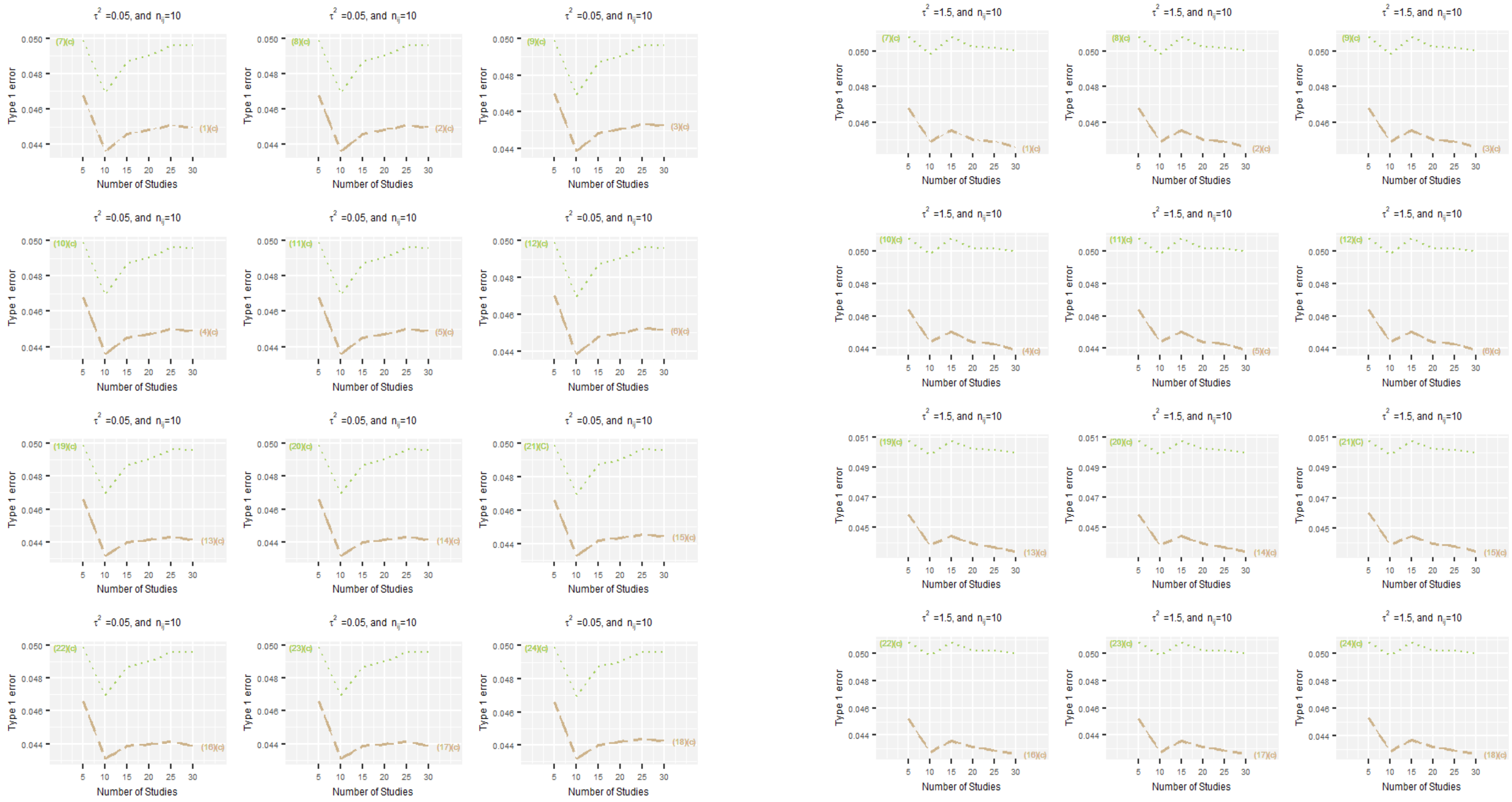


Figure B.17: *REML method: Comparing 24 Type I error by Hartung-Knapp t-distribution, with small sample sizes* $(n_{1j} = n_{2j})$, *and very low and large heterogeneity, for further details see Section B.1.1.*

Type I Error by Hartung-Knapp t-distribution: REML Method

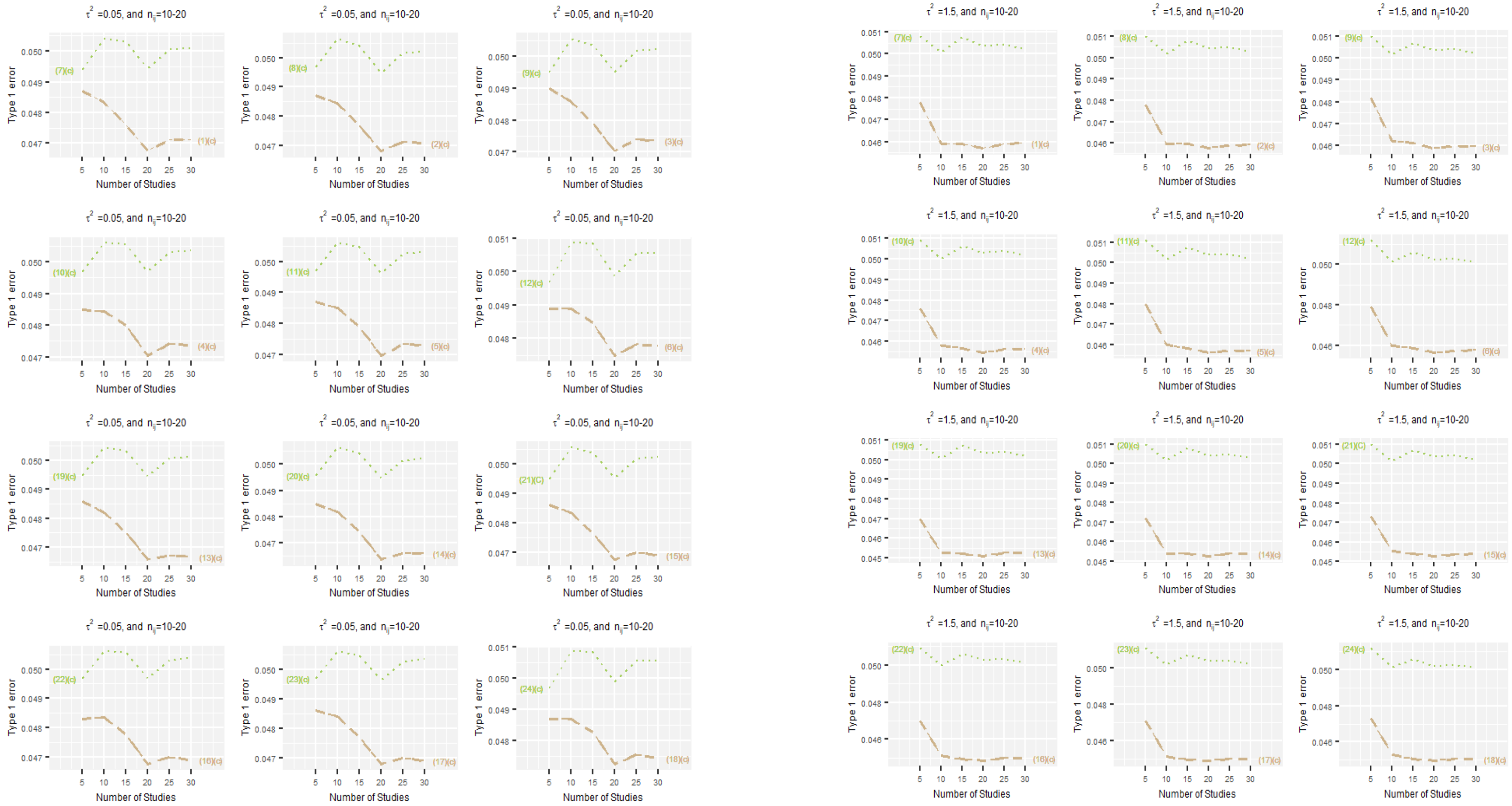
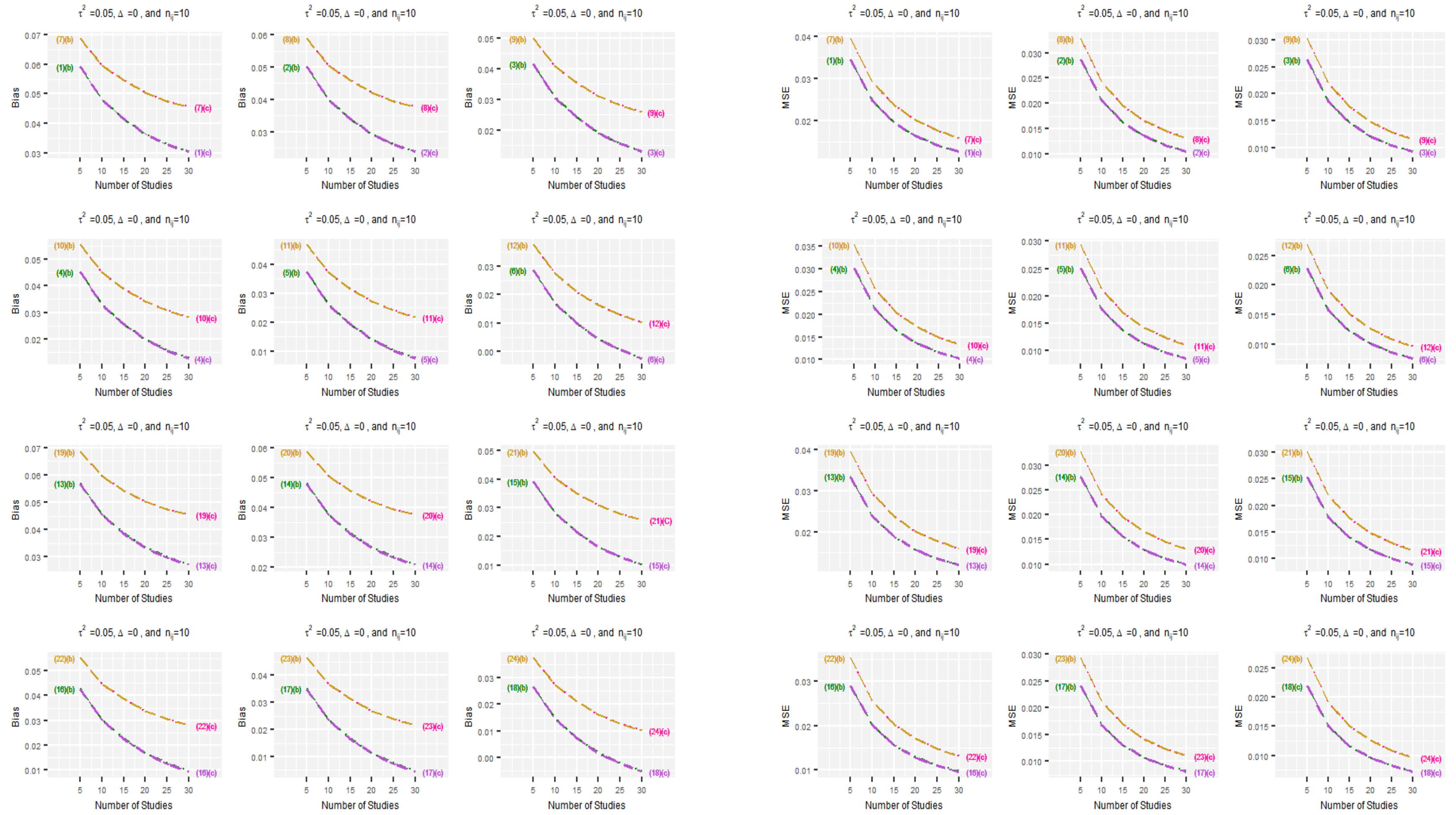


Figure B.18: *REML method: Comparing 24 Type I error by Hartung-Knapp t-distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), and very low and large heterogeneity, for further details see Section B.1.1.*

Between-Study Variance: PM Method

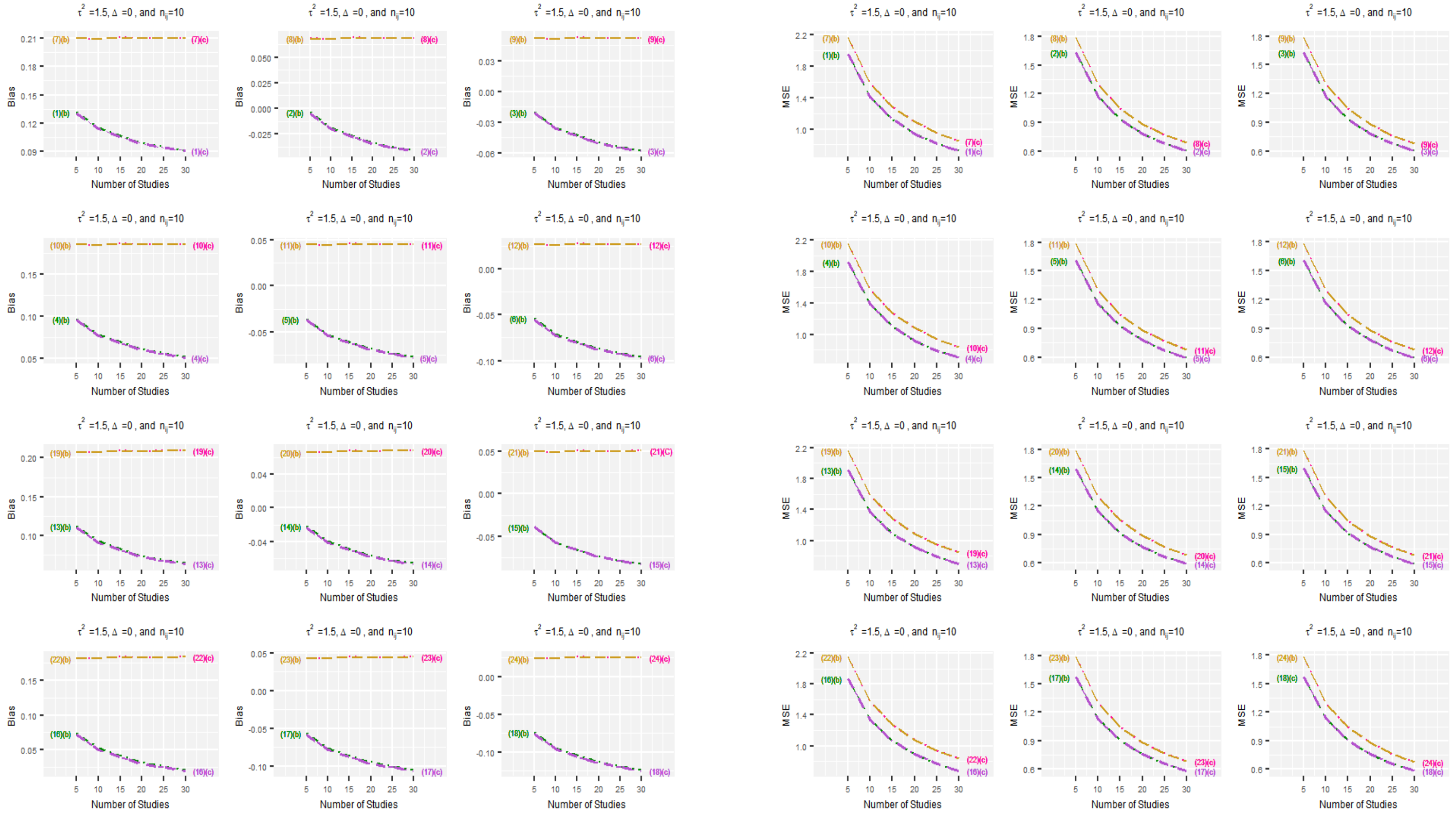


(a) Bias

(b) MES

Figure B.19: *PM Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes ($n_{1j} = n_{2j}$), no effect size, and very low heterogeneity, for further details see Section B.1.1.*

Between-Study Variance: PM Method

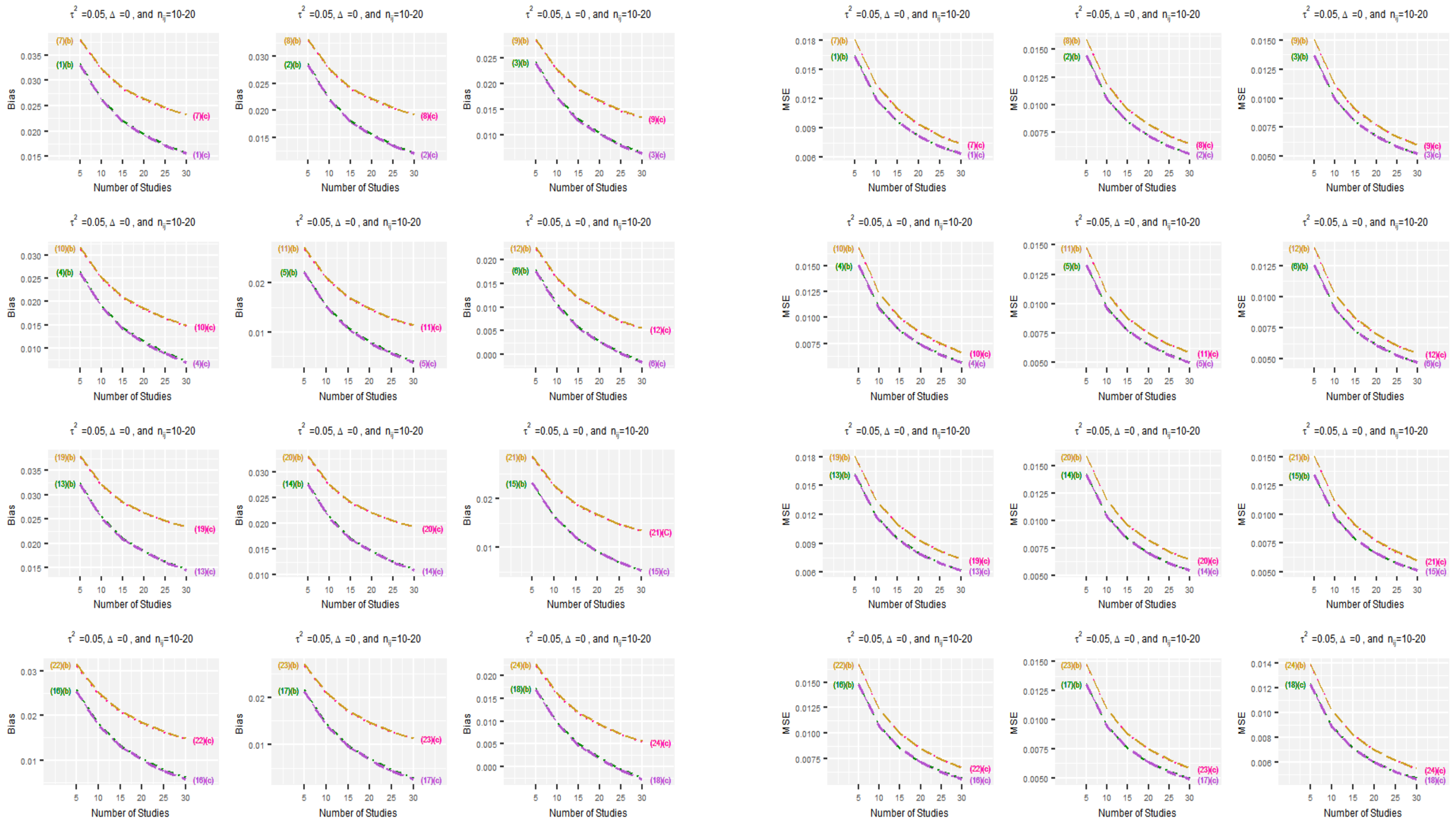


(a) Bias

(b) MES

Figure B.20: *PM Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and large heterogeneity, for further details see Section B.1.1.*

Between-Study Variance: PM Method

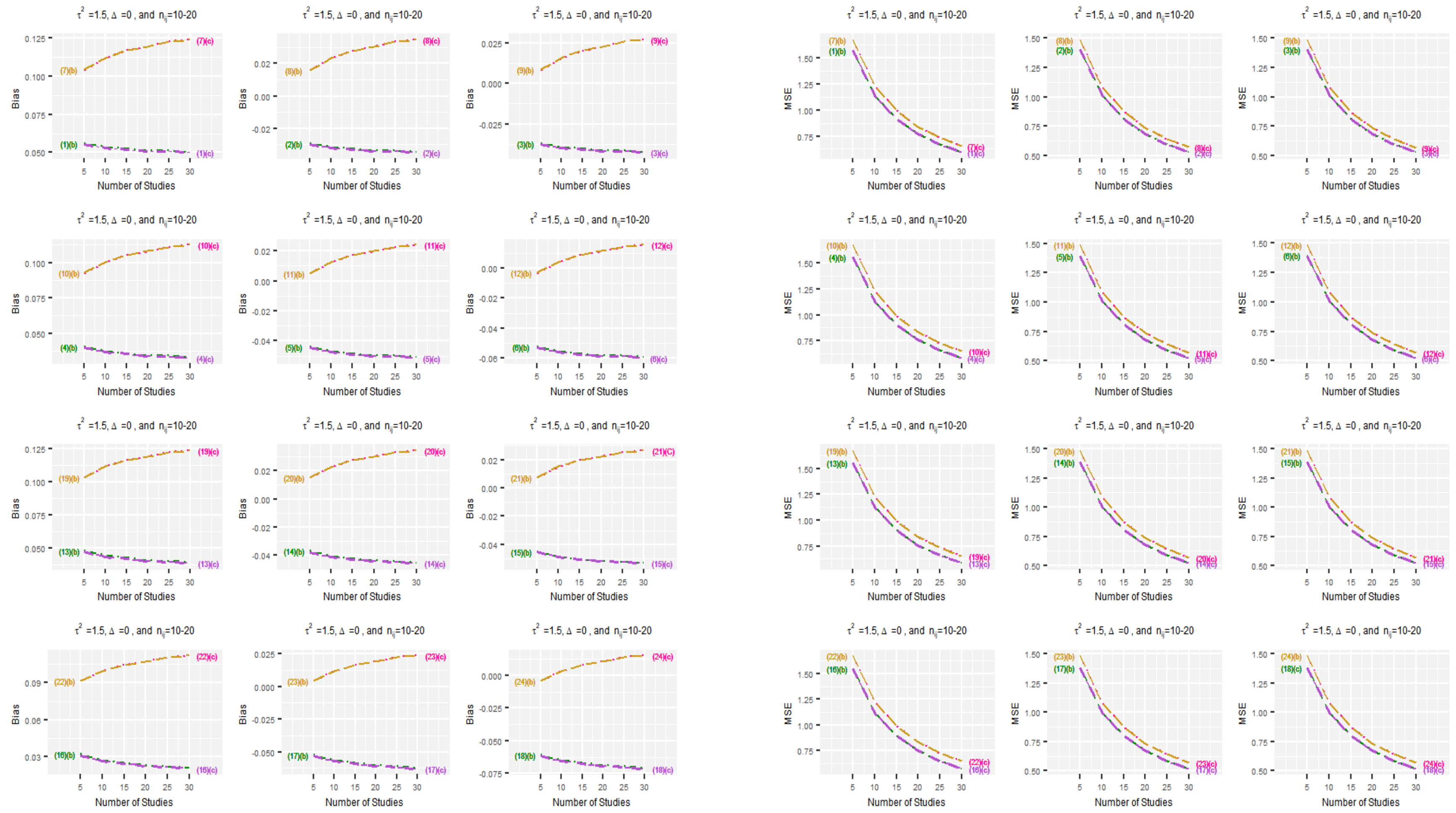


(a) Bias

(b) MES

Figure B.21: *PM Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes (n_{1j} not necessary equal to n_{2j}), no effect size, and very low heterogeneity, for further details see Section B.1.1.*

Between-Study Variance: PM Method



(a) Bias

(b) MES

Figure B.22: *PM Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes (n_{1j} not necessary equal to n_{2j}), no effect size, and large heterogeneity, for further details see Section B.1.1.*

Proportion of zero heterogeneity variance estimates: PM Method

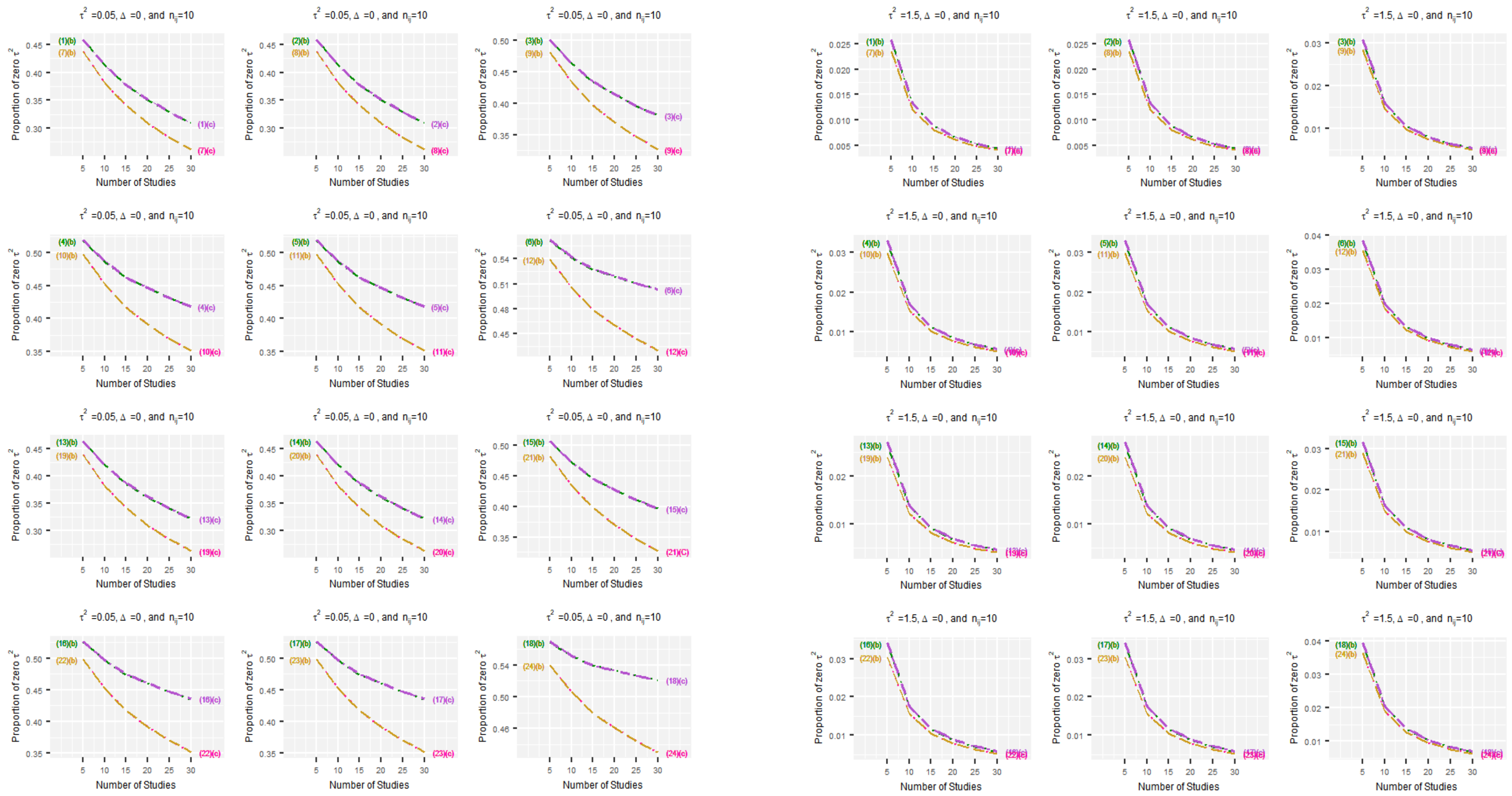


Figure B.23: *PM Method: 48 between-study variance estimators are compared through their proportion of zero with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and very low and large heterogeneity, for further details see Section B.1.1.*

Proportion of zero heterogeneity variance estimates: PM Method

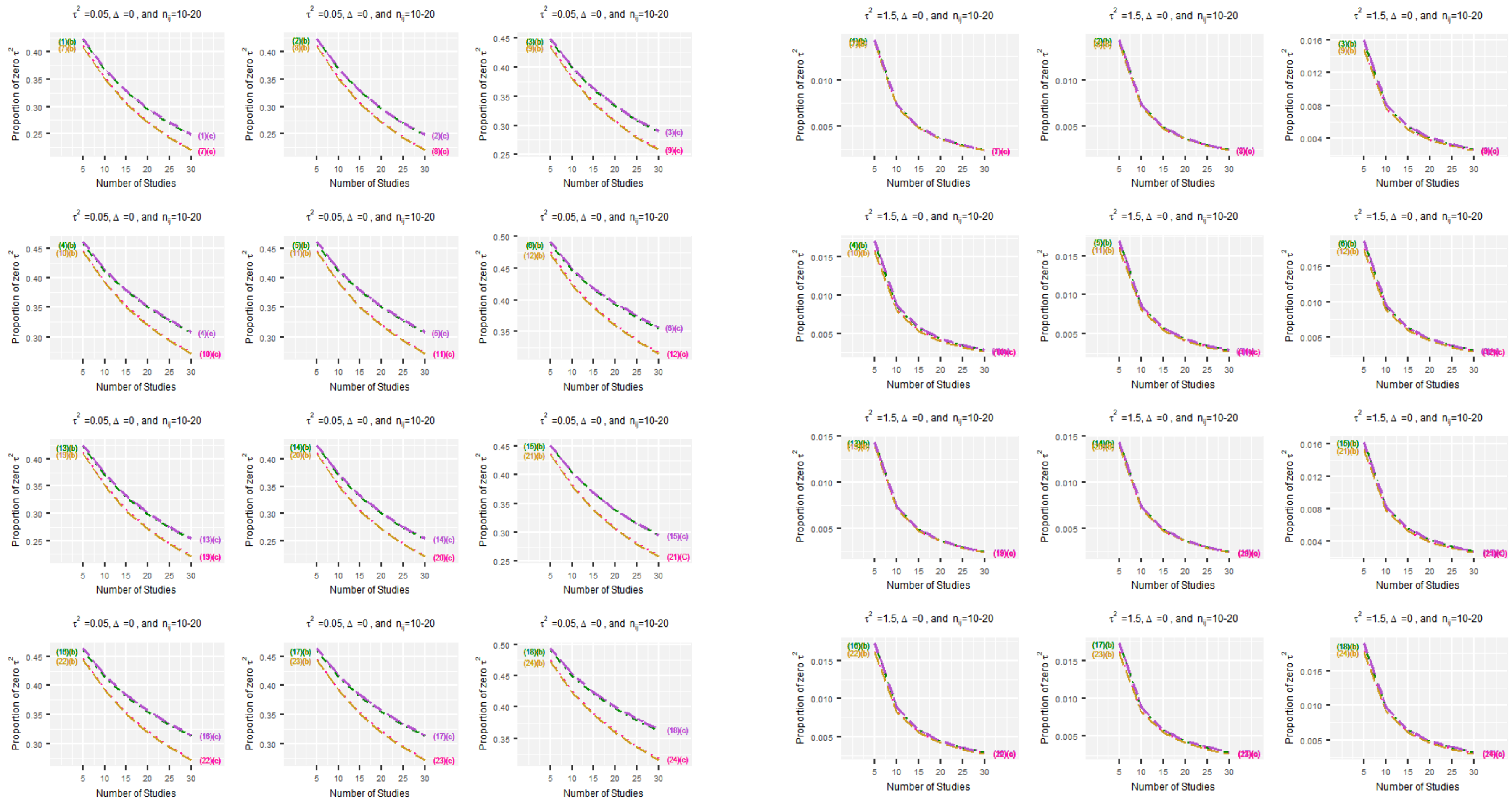
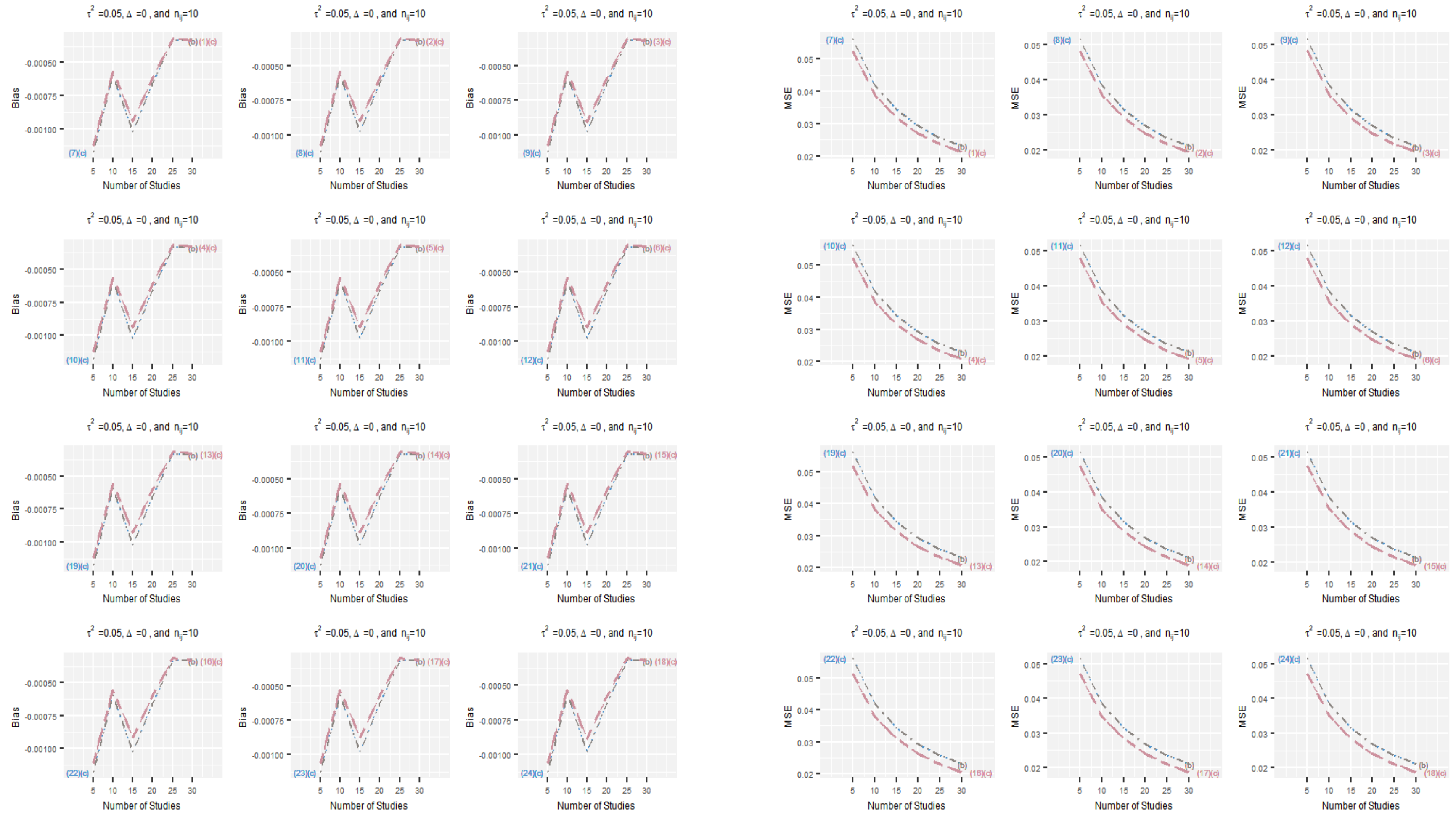


Figure B.24: *PM Method: 48 between-study variance estimators are compared through their proportion of zero with small sample sizes (n_{1j} not necessary equal to n_{2j}), no effect size, and very low and large heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: PM Method

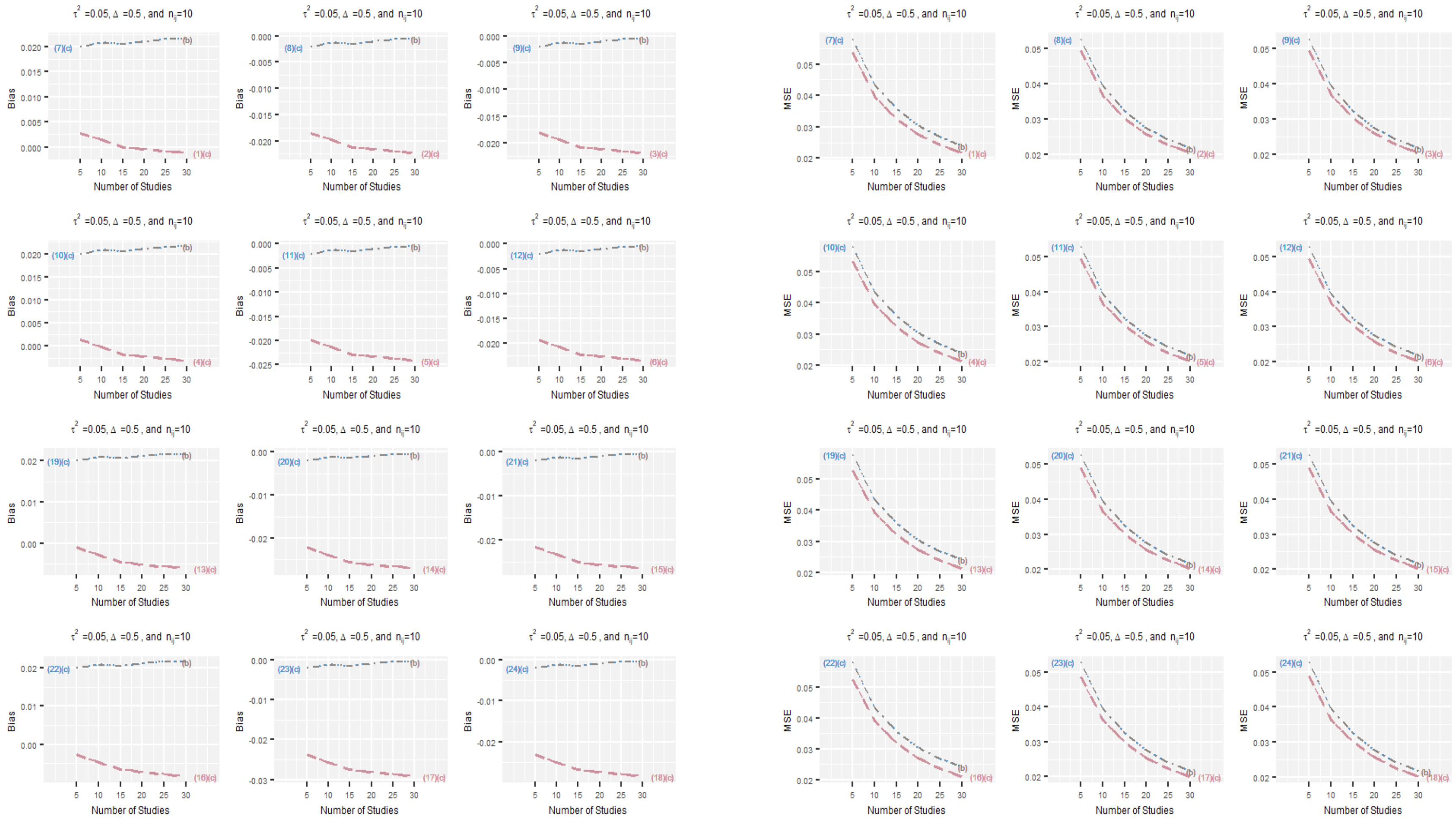


(a) Bias

(b) MES

Figure B.25: *PM method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect sizes , and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: PM Method

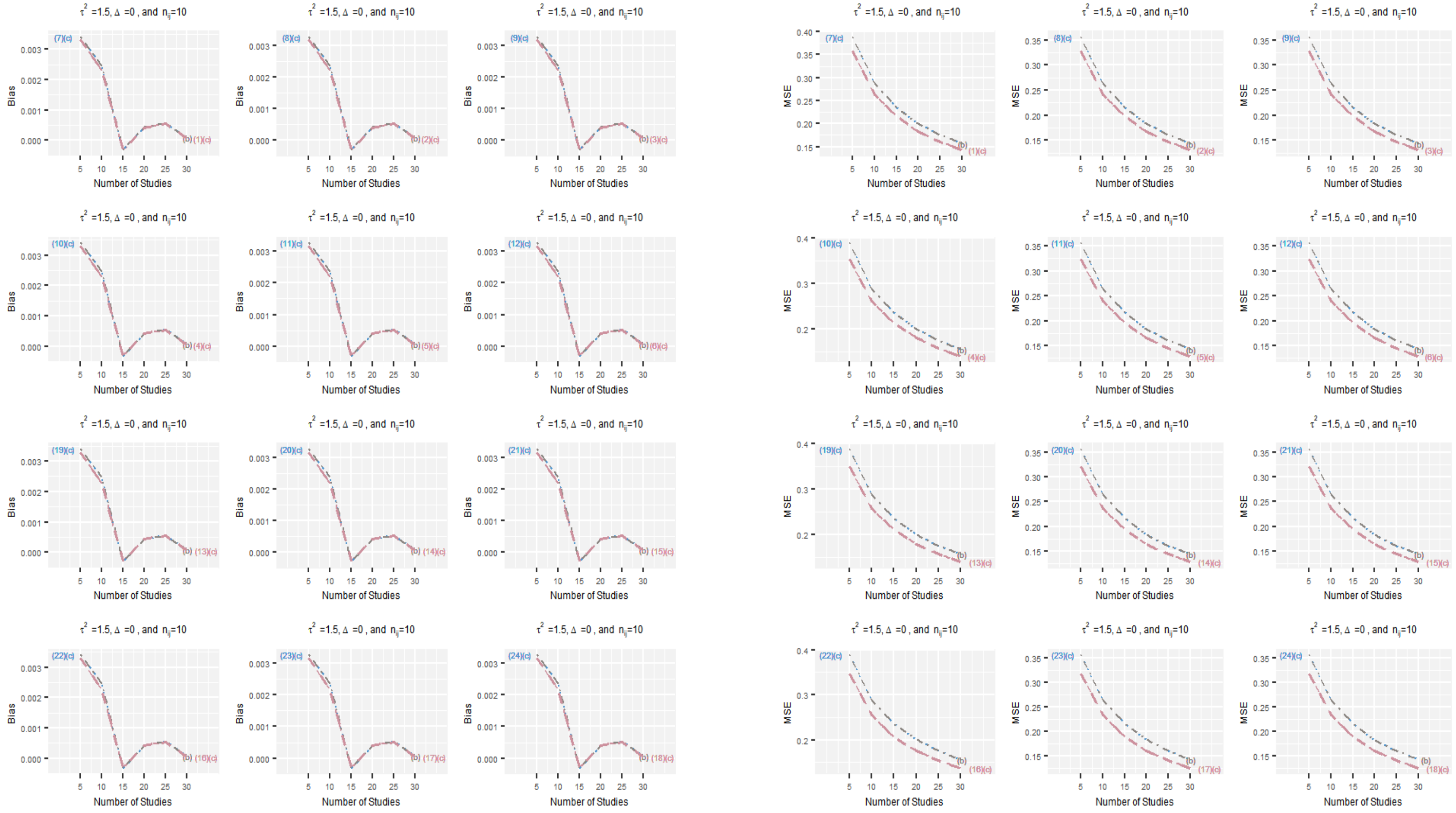


(a) Bias

(b) MES

Figure B.26: *PM method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes ($n_{1j} = n_{2j}$), medium effect sizes , and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: PM Method

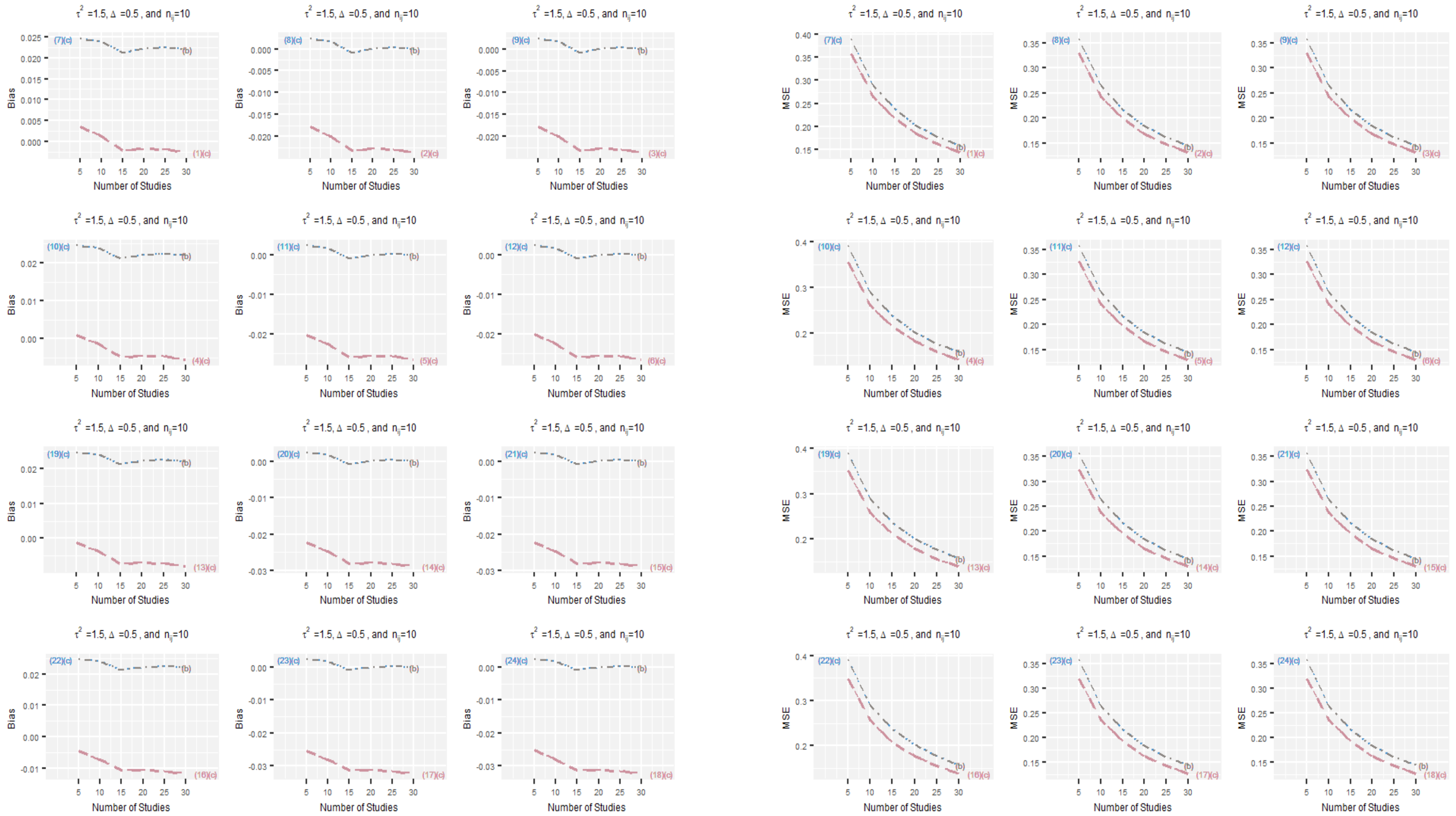


(a) Bias

(b) MES

Figure B.27: *PM method: 25 overall effect size estimators are compared through their bias and mean squared error by using REML method, with small sample sizes* ($n_{1j} = n_{2j}$), *no effect sizes , and large heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: PM Method

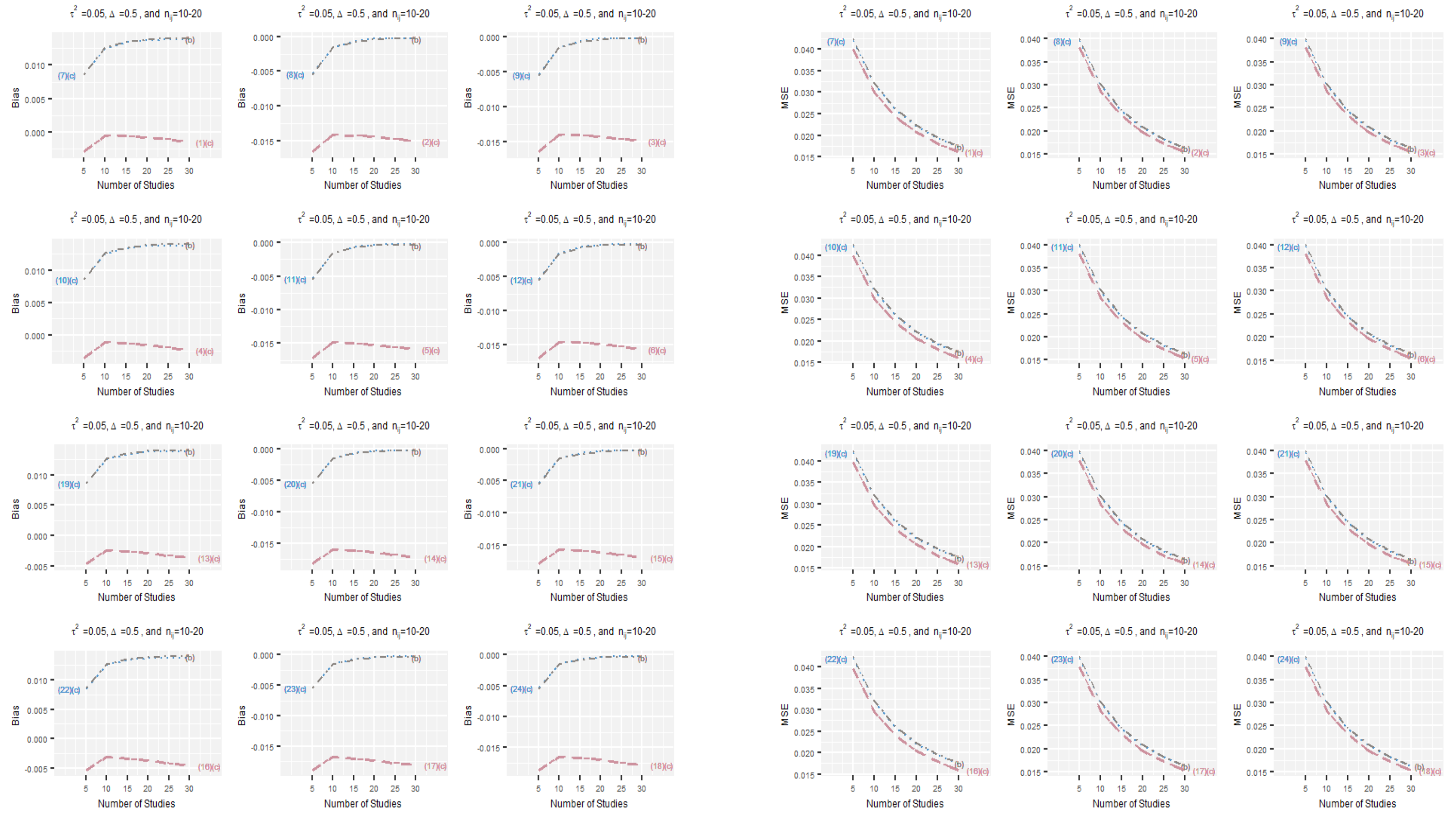


(a) Bias

(b) MES

Figure B.28: *PM method: 25 overall effect size estimators are compared through their bias and mean squared error by using REML method, with small sample sizes* ($n_{1j} = n_{2j}$)*, medium effect sizes , and large heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: PM Method

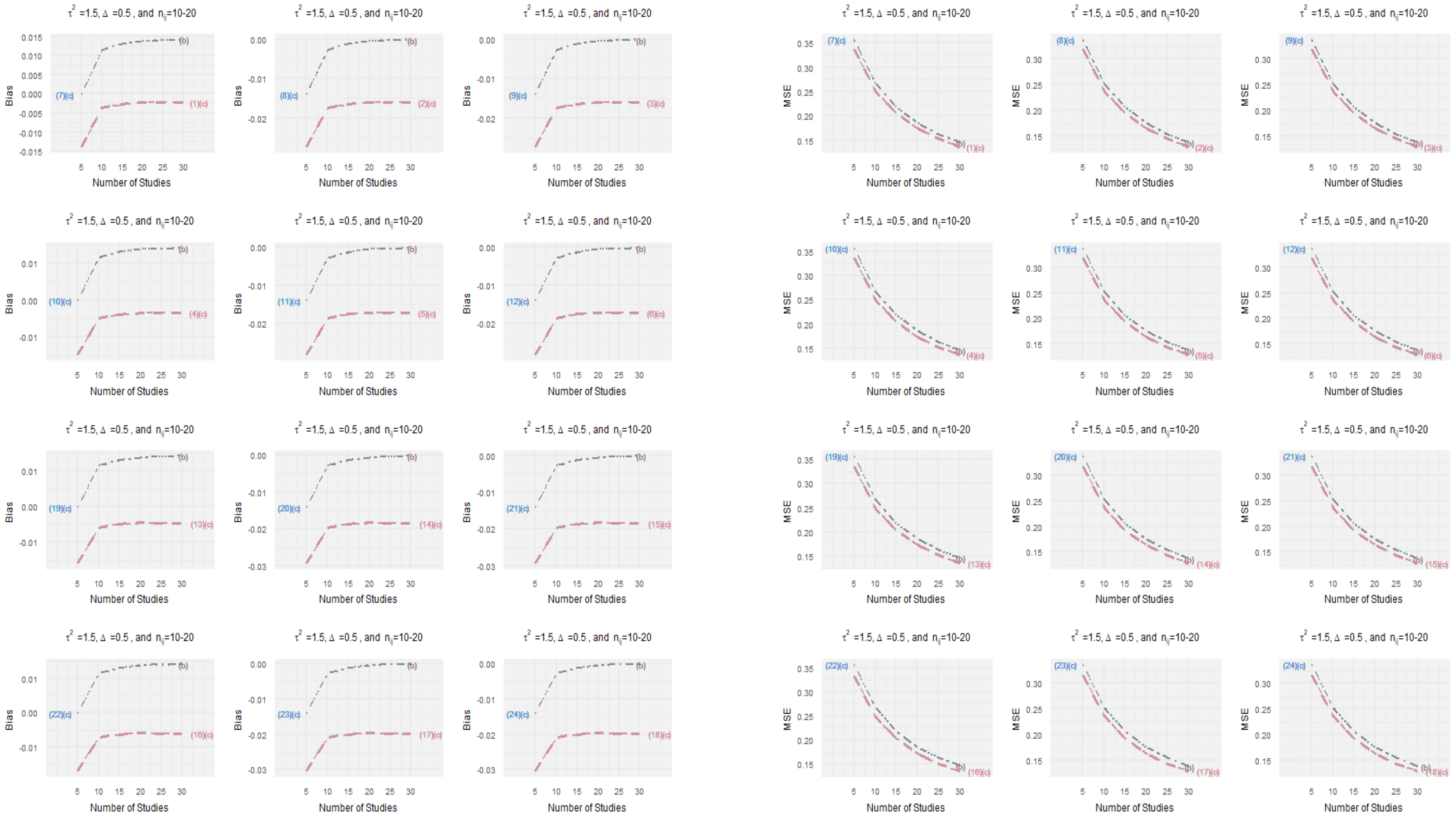


(a) Bias

(b) MES

Figure B.29: *PM method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes (n_{1j} not necessary equal to n_{2j}), medium effect size, and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: PM Method

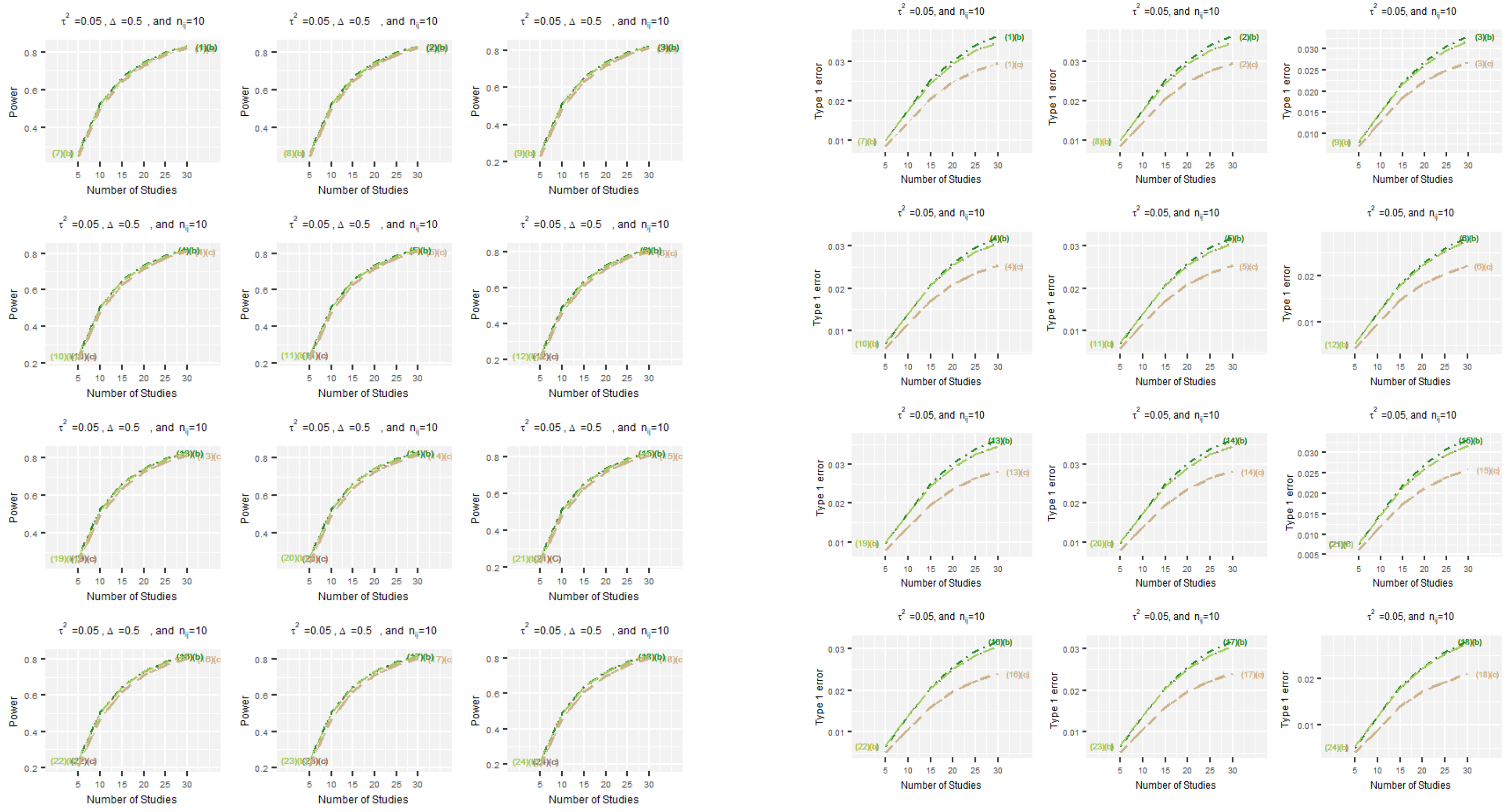


(a) Bias

(b) MES

Figure B.30: *PM method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes (n_{1j} not necessary equal to n_{2j}), medium effect size, and large heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: PM Method

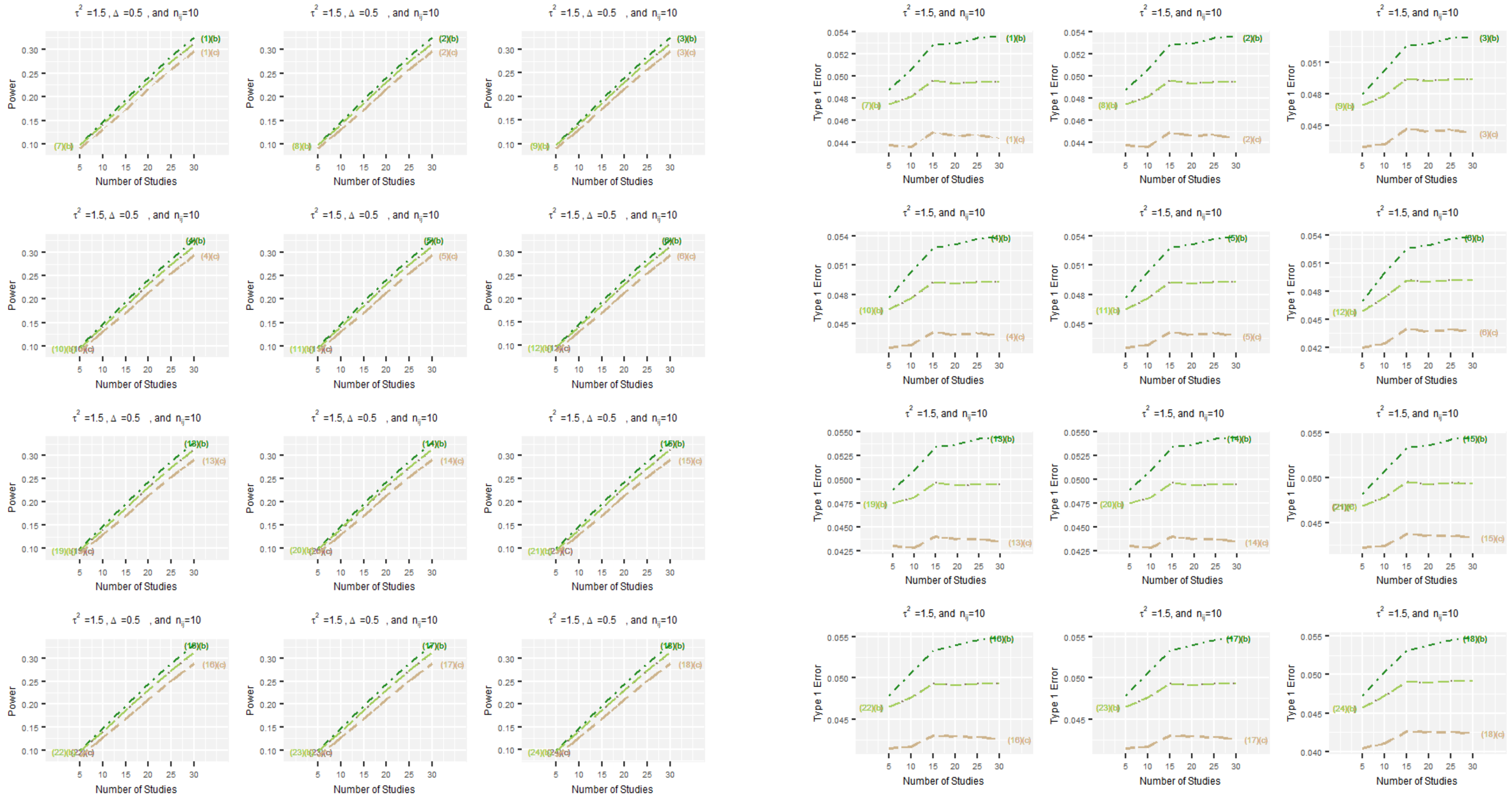


(a) Power

(b) Type 1 Error

Figure B.31: *PM method: Comparing 48 Power and Type I error by t distribution, with small sample sizes ($n_{1j} = n_{2j}$), and very low heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: PM Method

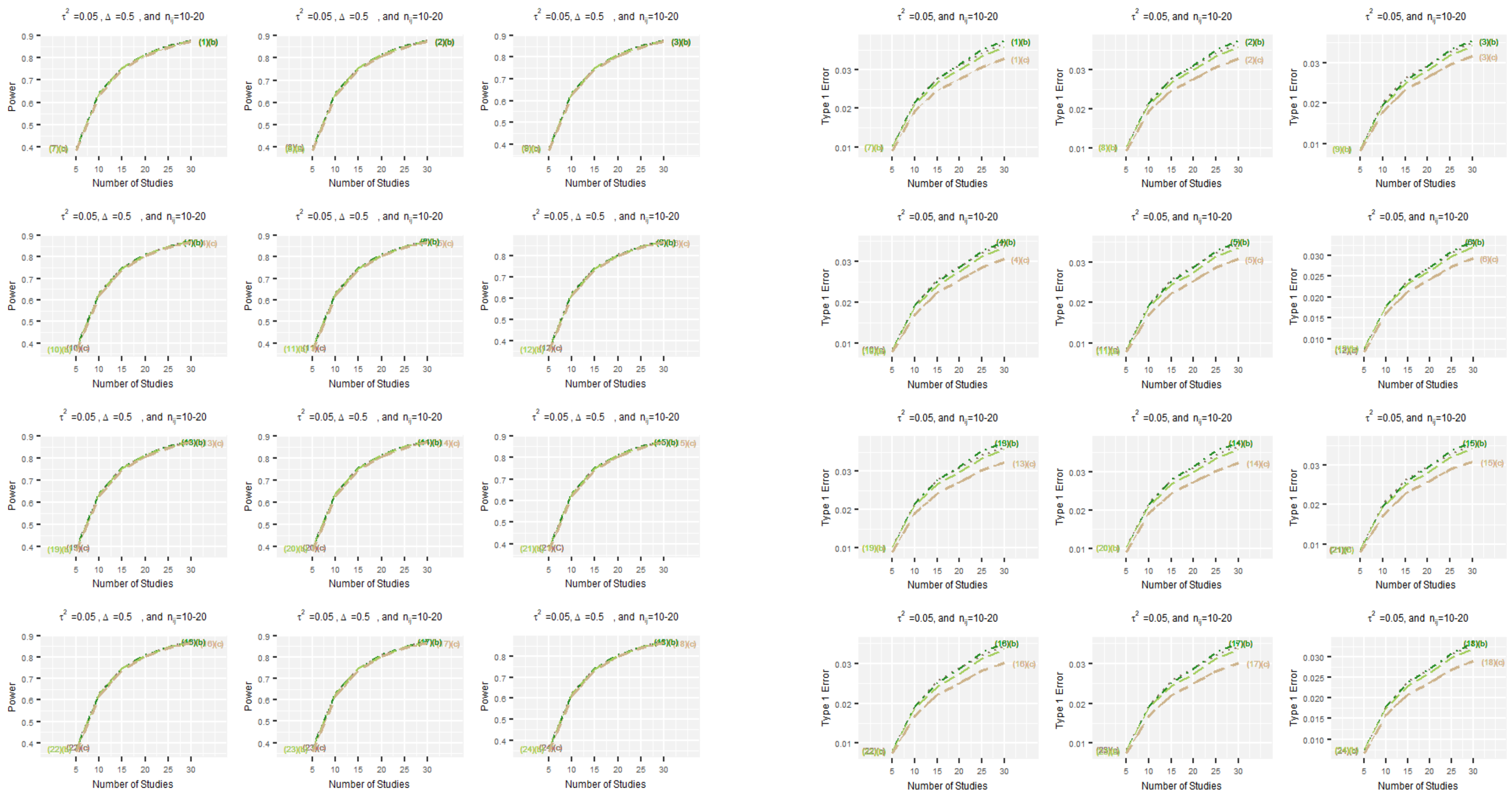


(a) Power

(b) Type 1 Error

Figure B.32: *PM method: Comparing 48 Power and Type I error by t distribution, with small sample sizes ($n_{1j} = n_{2j}$), and large heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: PM Method

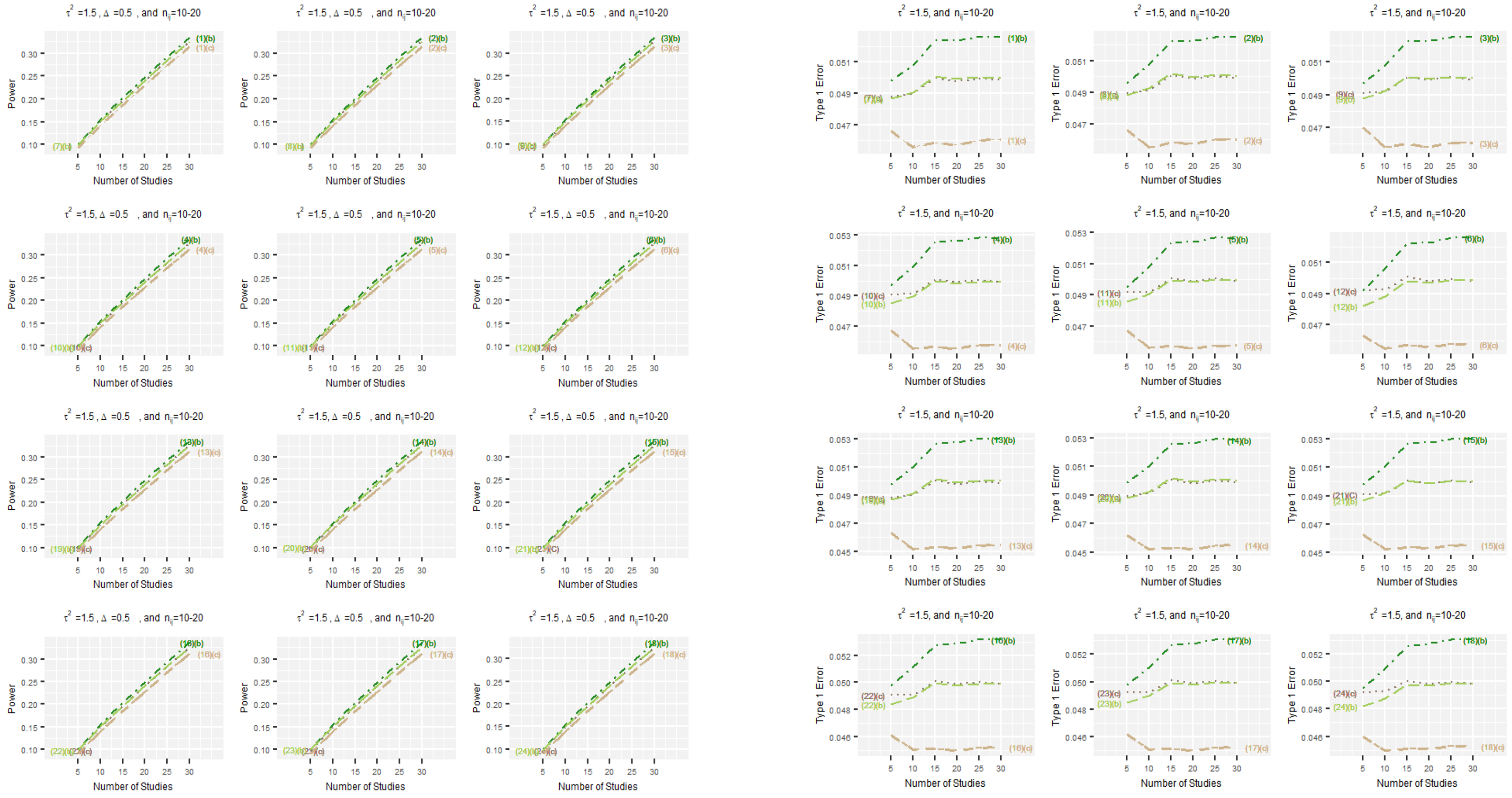


(a) Power

(b) Type 1 Error

Figure B.33: *PM method: Comparing 48 Power and Type I error by t distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), and very low heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: PM Method



(a) Power

(b) Type 1 Error

Figure B.34: *PM method: Comparing 48 Power and Type I error by t distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), and large heterogeneity, for further details see Section B.1.1.*

Type I Error by Hartung-Knapp t-distribution: PM Method

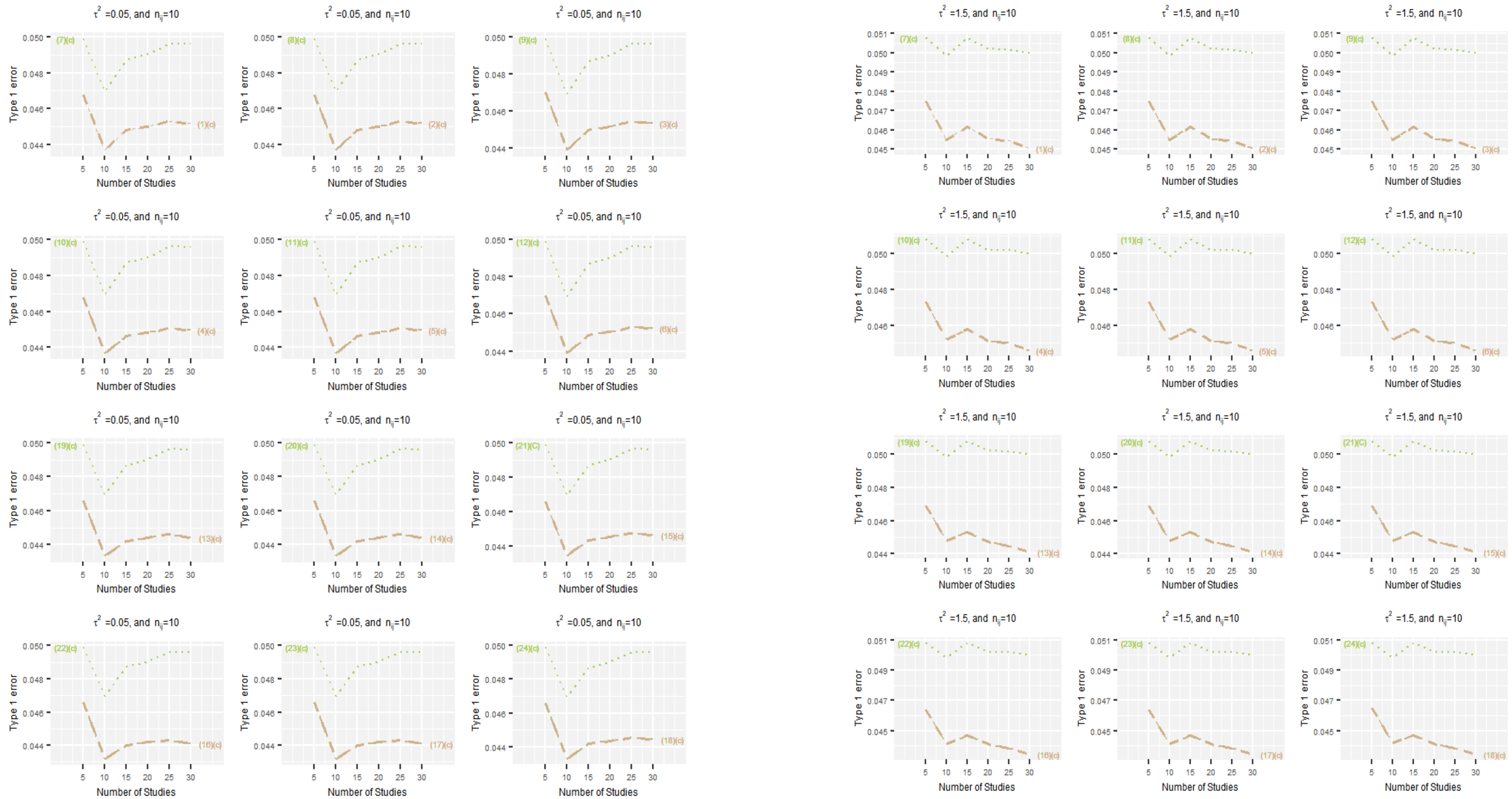


Figure B.35: *PM method: Comparing 24 Type I error by Hartung-Knapp t-distribution, with small sample sizes* $(n_{1j} = n_{2j})$*, and very low and large heterogeneity, for further details see Section B.1.1.*

Type I Error by Hartung-Knapp t-distribution: PM Method

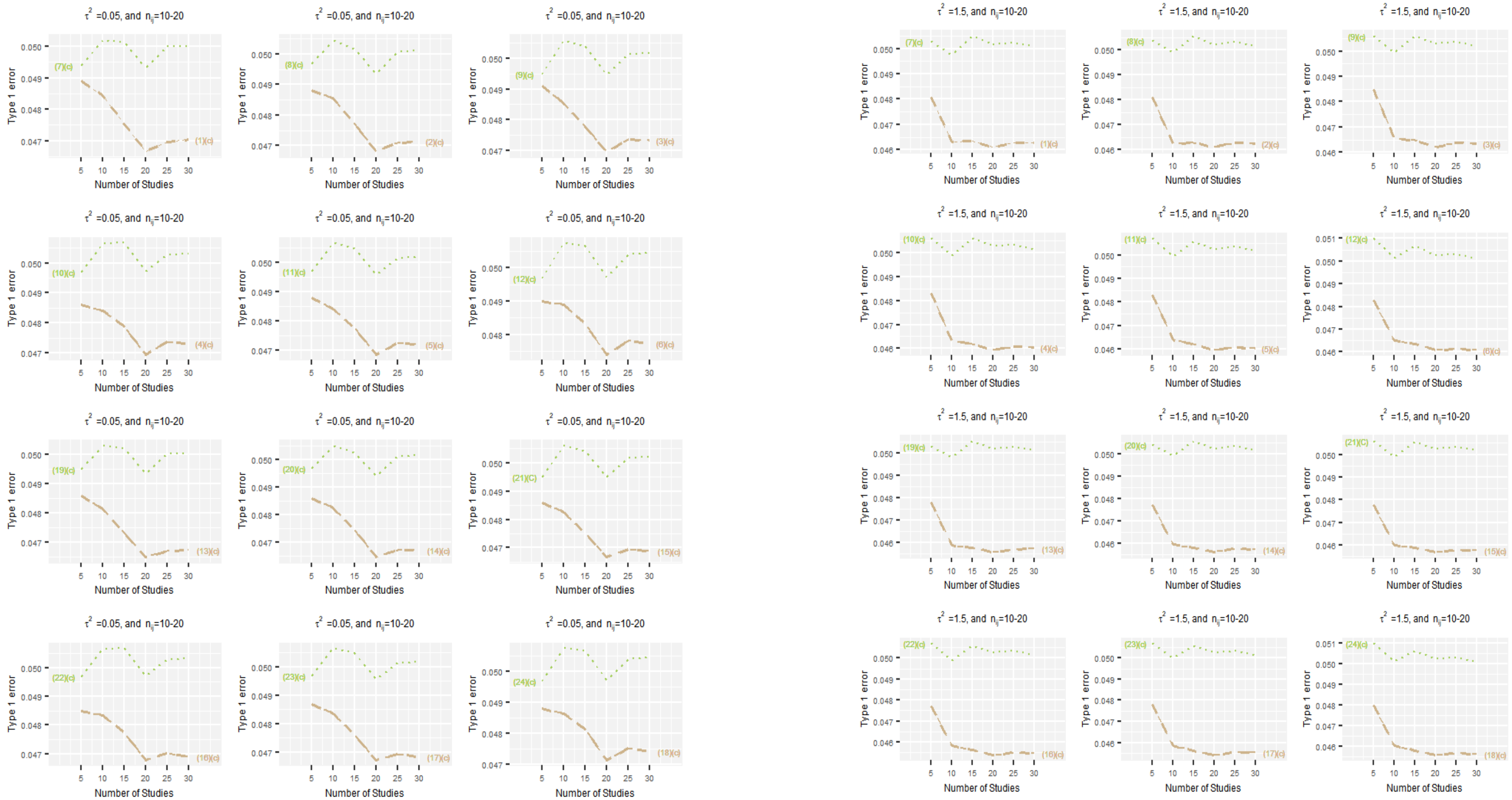
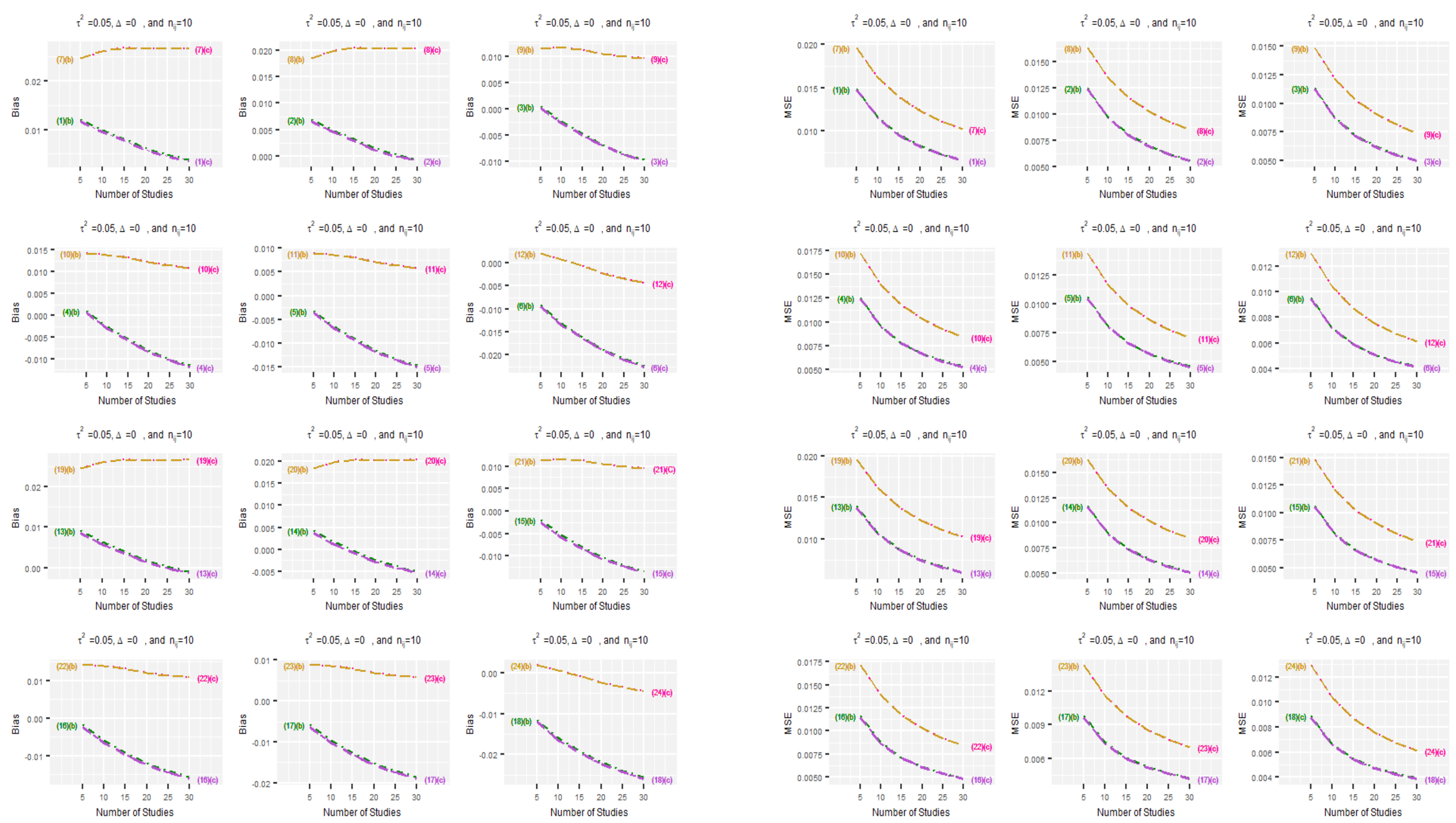


Figure B.36: *PM method: Comparing 24 Type I error by Hartung-Knapp t-distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), and very low and large heterogeneity, for further details see Section B.1.1.*

Between-Study Variance: ML Method

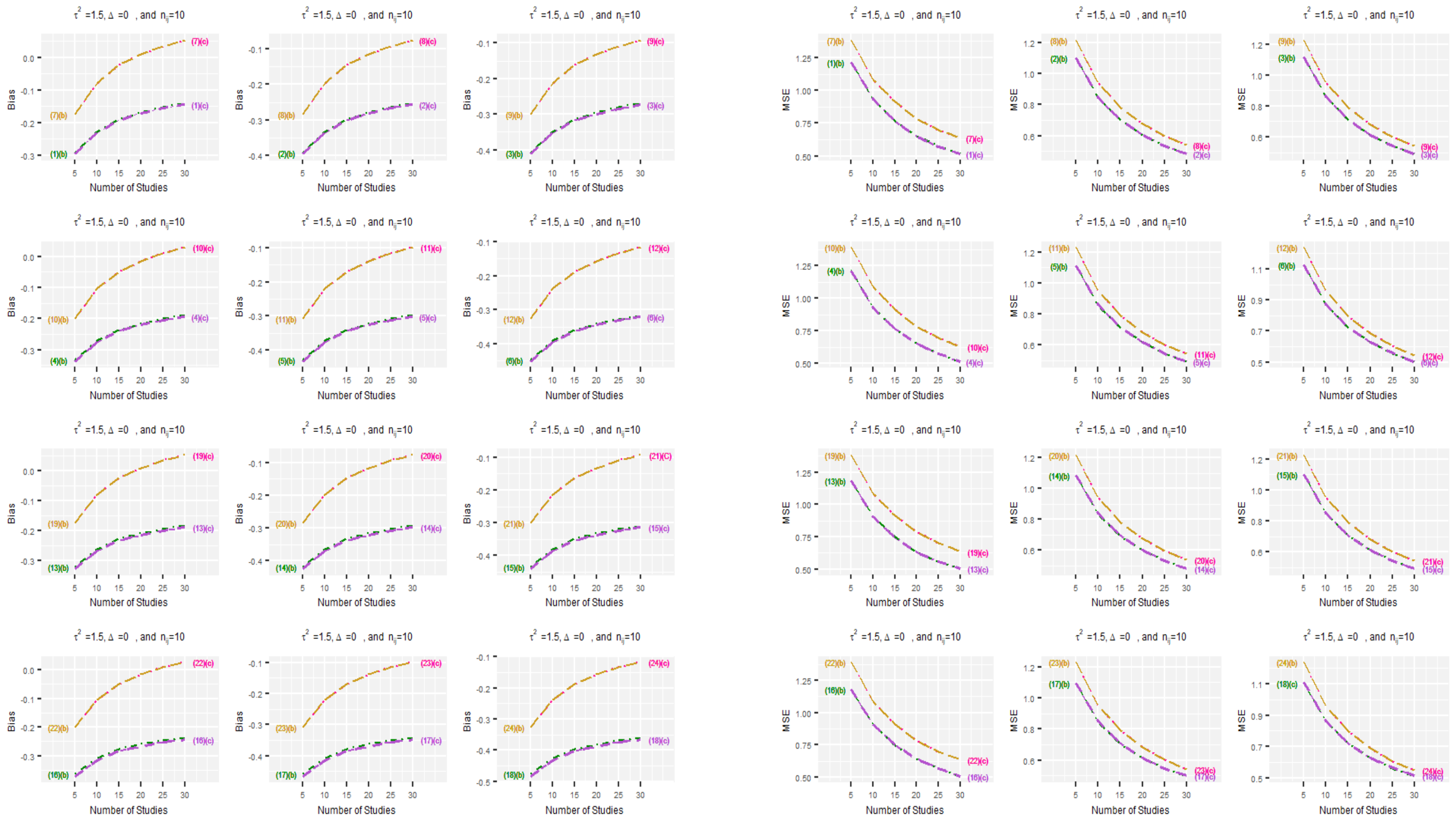


(a) Bias

(b) MES

Figure B.37: *ML Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and very low heterogeneity, for further details see Section B.1.1.*

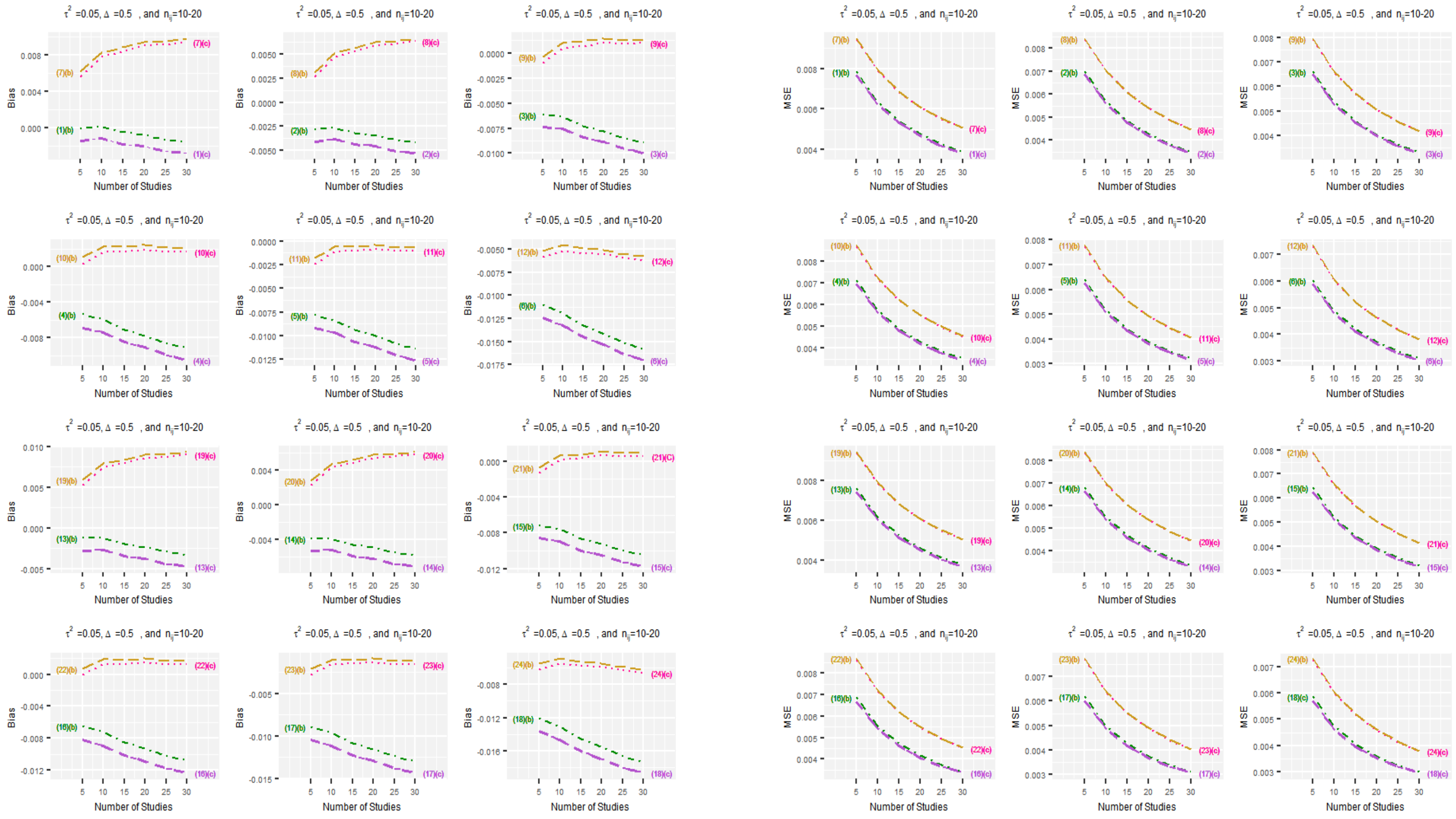
Between-Study Variance: ML Method



(a) Bias

(b) MES

Figure B.38: *ML Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and large heterogeneity, for further details see Section B.1.1.*

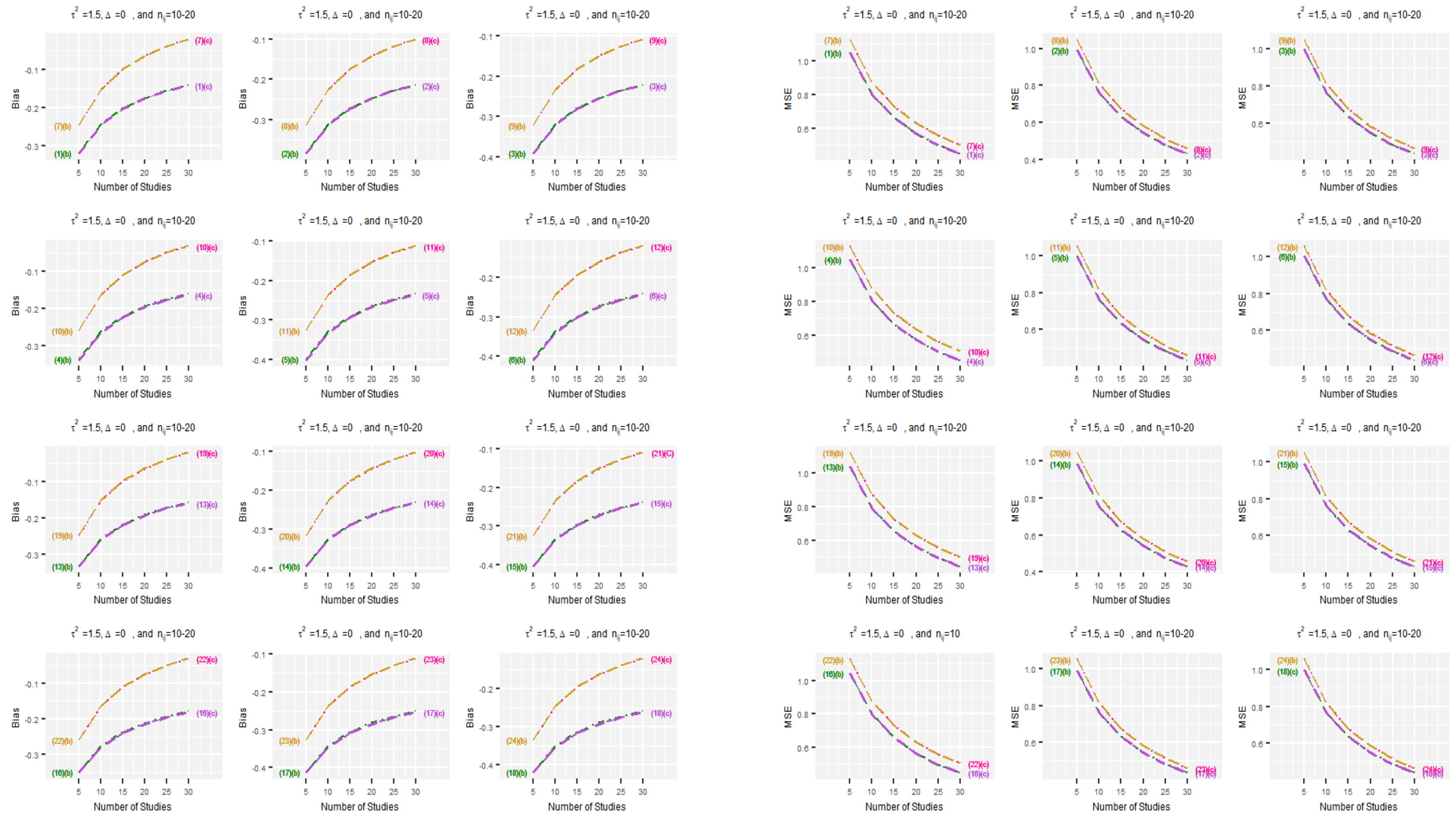


(a) Bias

(b) MES

Figure B.39: *ML Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes (n_{1j} not necessary equal to n_{2j}), no effect size, and very low heterogeneity, for further details see Section B.1.1.*

Between-Study Variance: ML Method



(a) Bias

(b) MES

Figure B.40: *ML Method: 48 between-study variance estimators are compared through their bias and mean squared error with small sample sizes* (n_{1j} *not necessary equal to* n_{2j}), *no effect size, and large heterogeneity, for further details see Section* B.1.1.

Proportion of zero heterogeneity variance estimates: ML Method

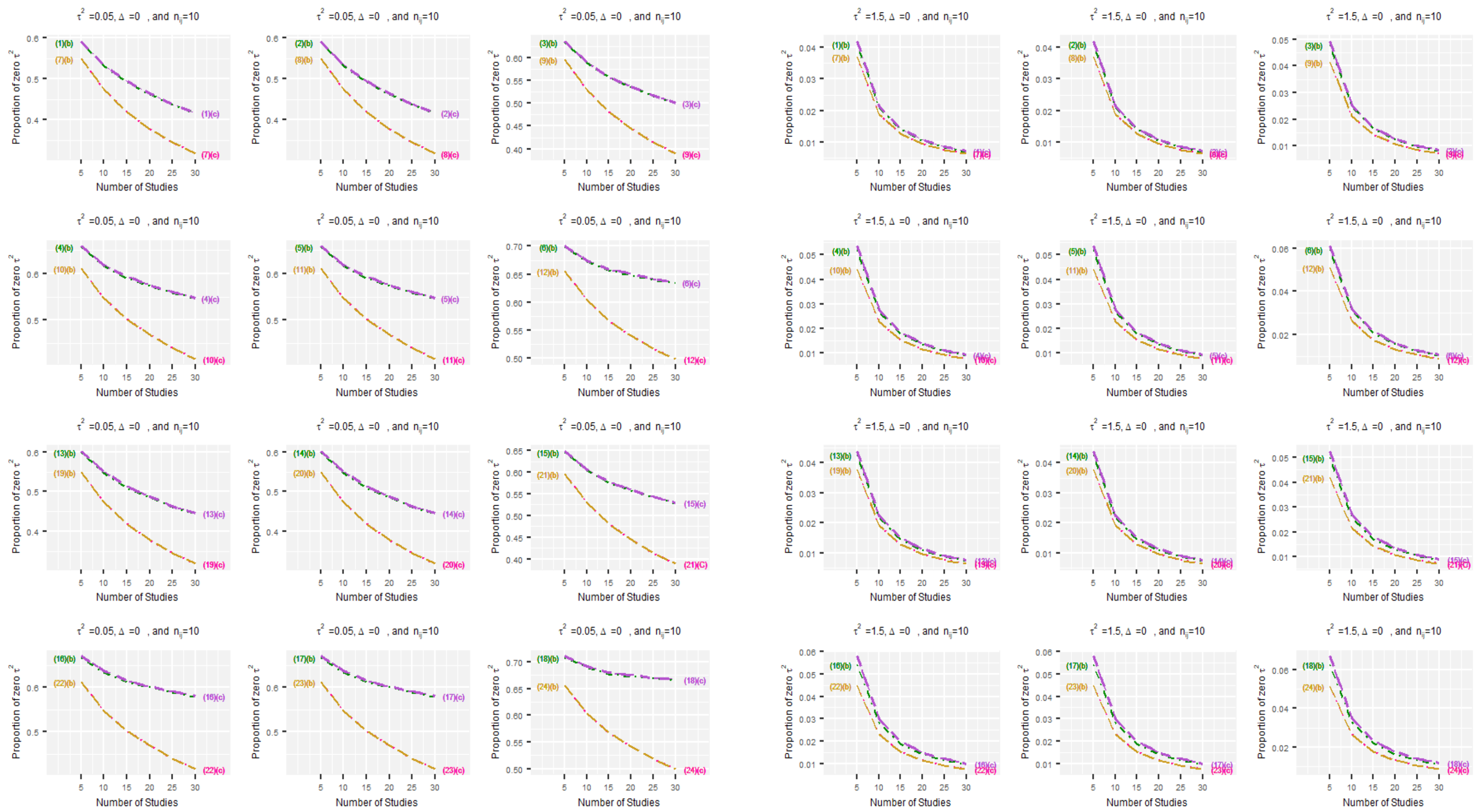


Figure B.41: *ML Method: 48 between-study variance estimators are compared through their proportion of zero with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and very low and large heterogeneity, for further details see Section B.1.1.*

Proportion of zero heterogeneity variance estimates: ML Method

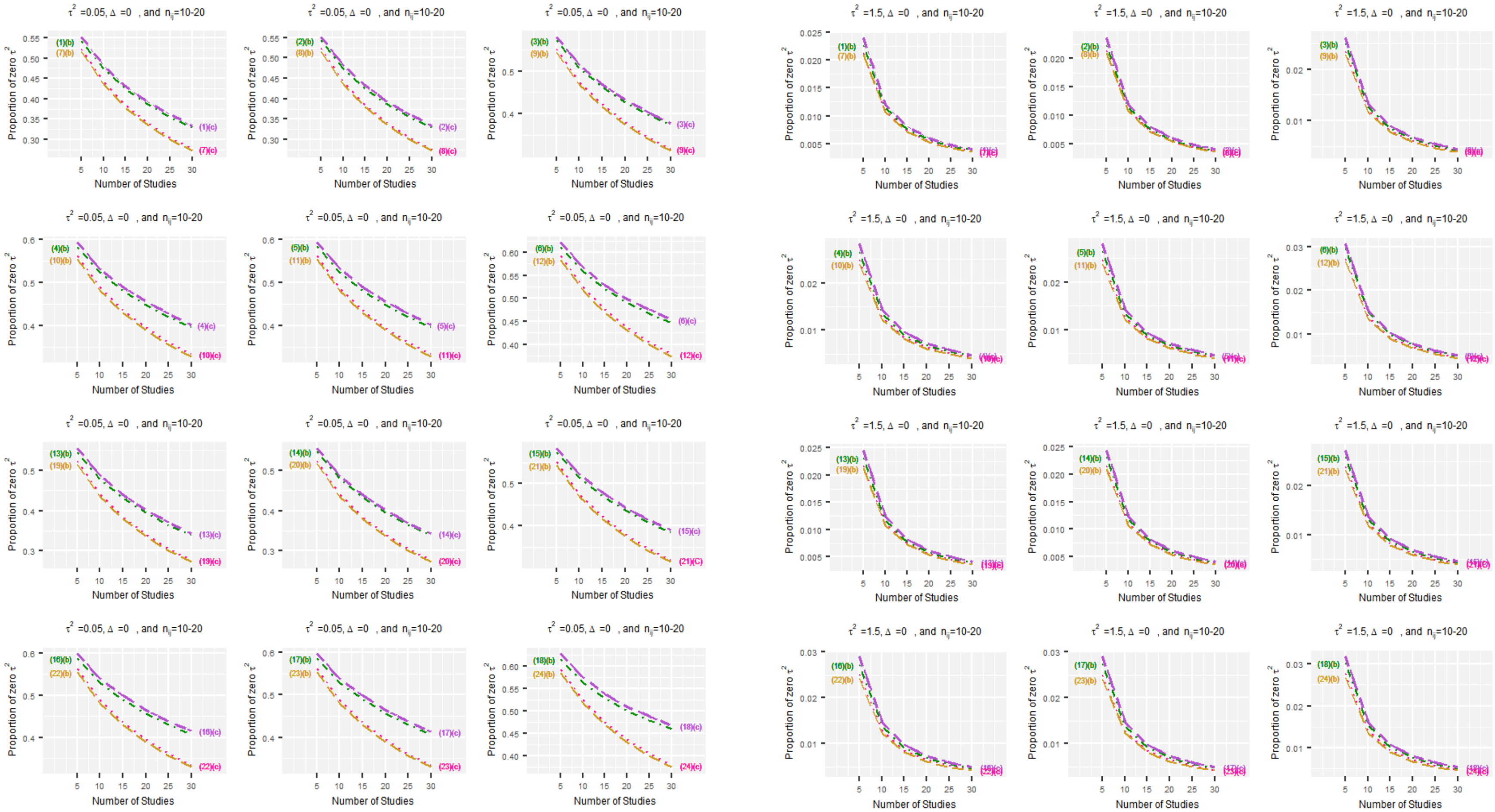
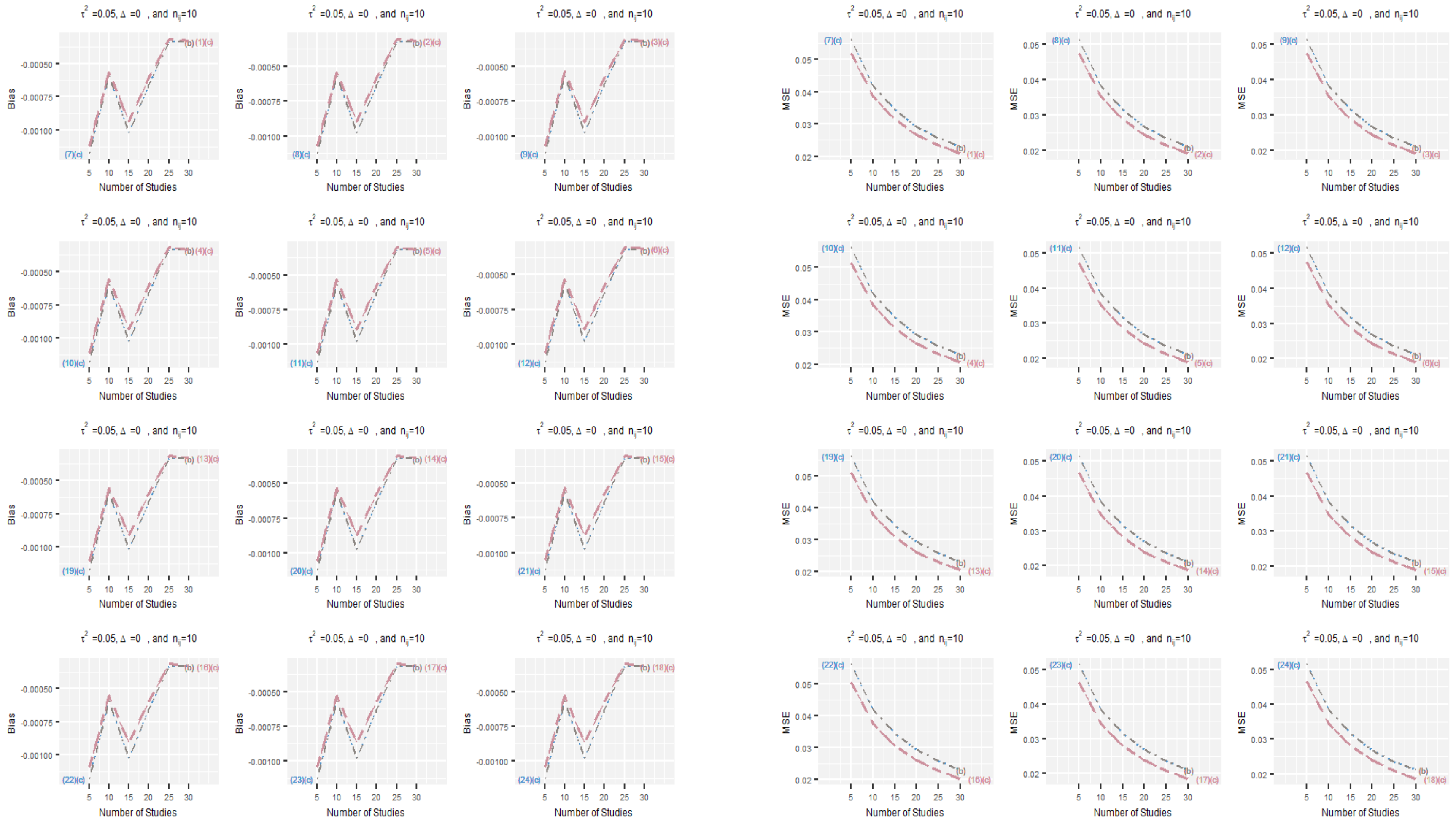


Figure B.42: *ML Method: 48 between-study variance estimators are compared through their proportion of zero with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect size, and very low and large heterogeneity, for further details see Section* B.1.1.

Overall Effect Size: ML Method

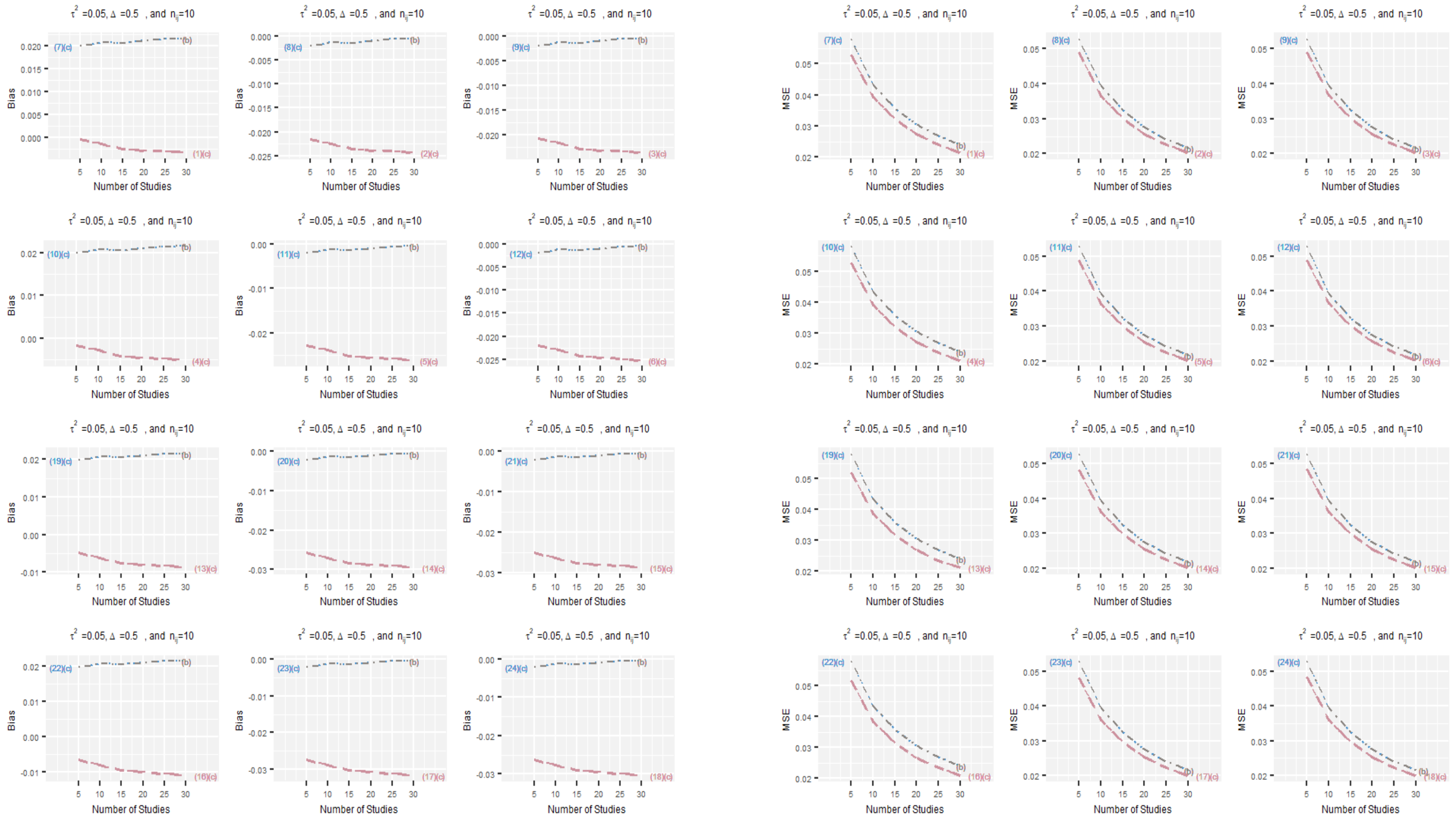


(a) Bias

(b) MES

Figure B.43: *ML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect sizes , and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: ML Method

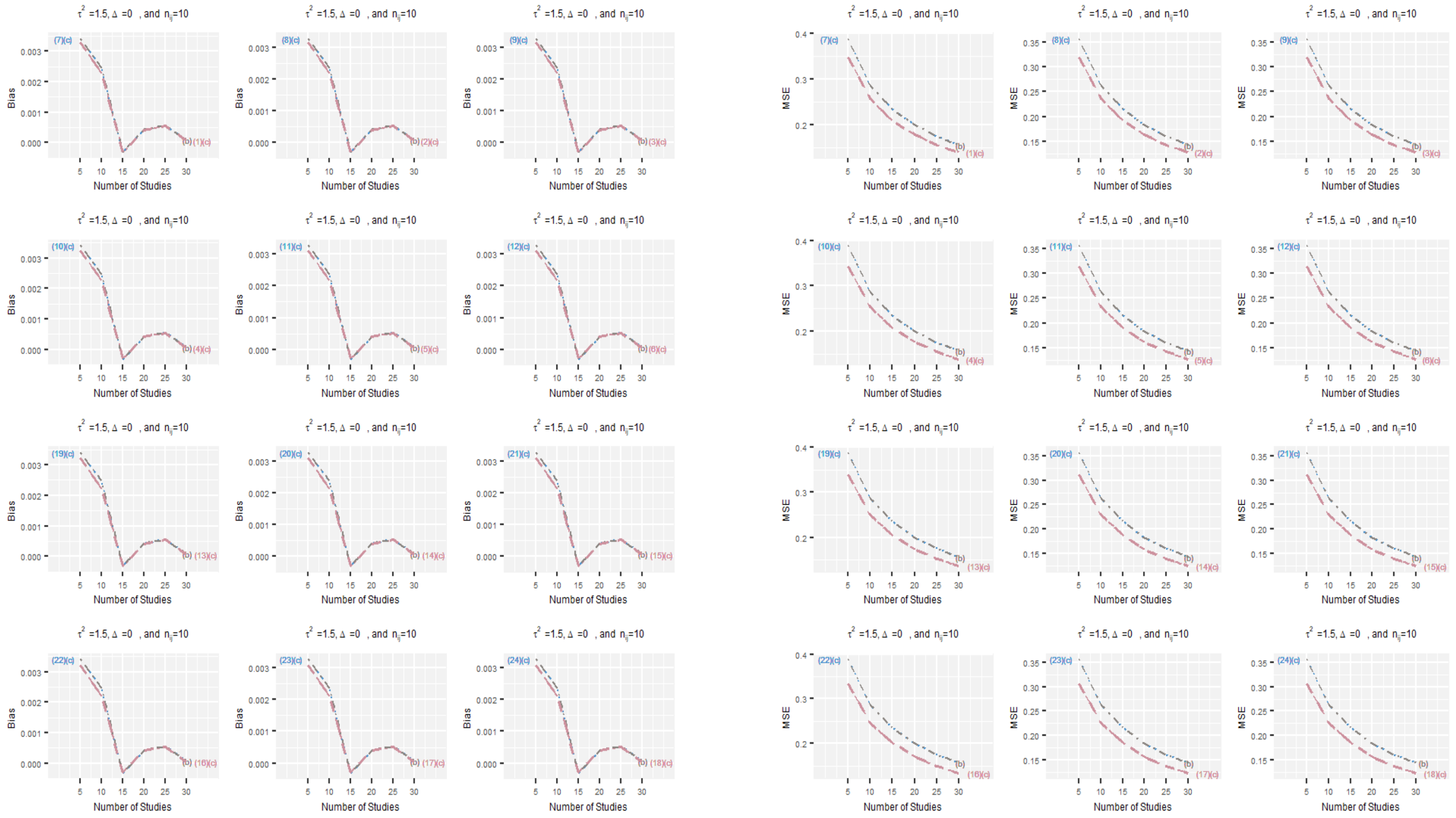


(a) Bias

(b) MES

Figure B.44: *ML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes* ($n_{1j} = n_{2j}$)*, medium effect sizes , and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: ML Method

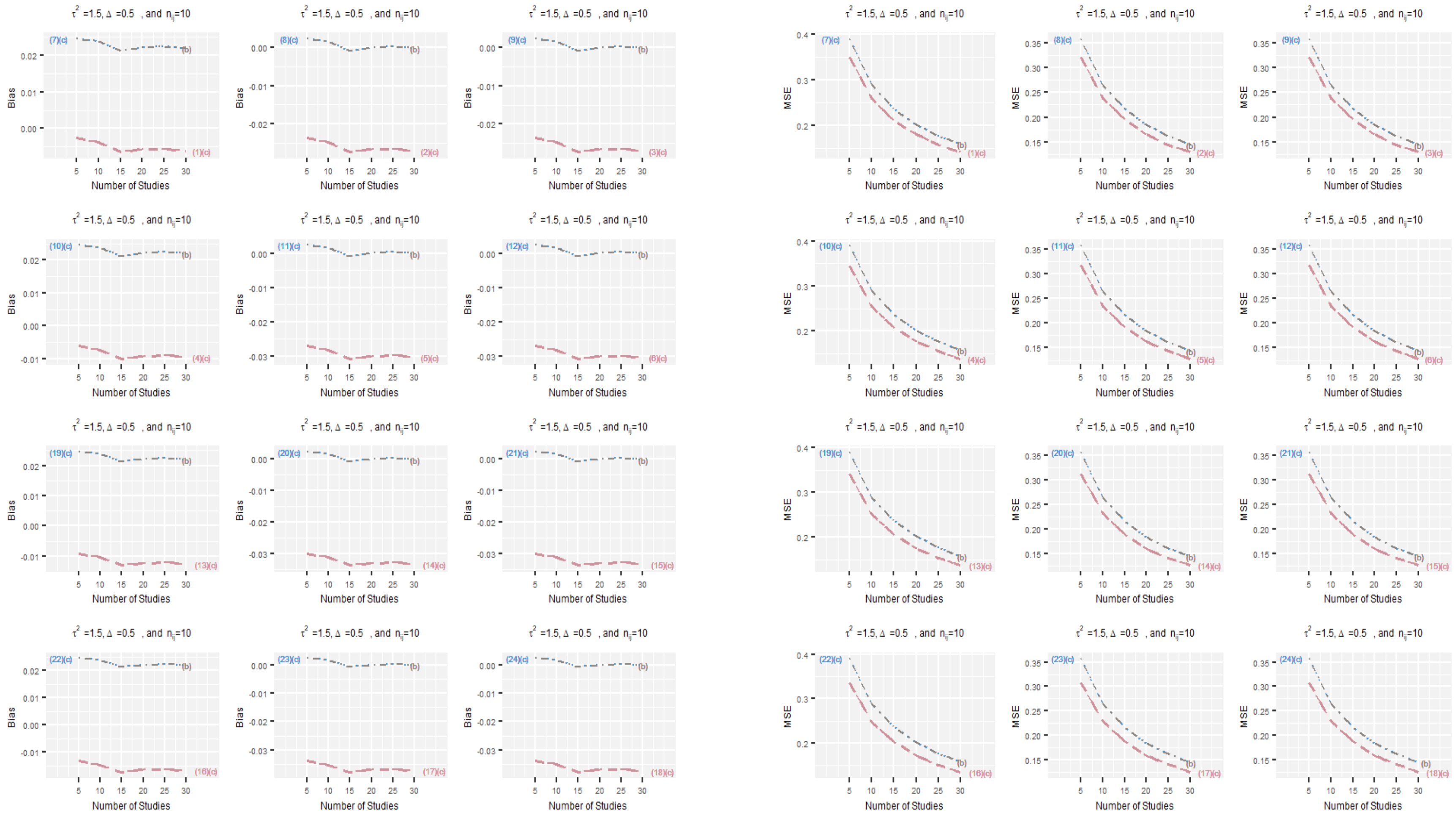


(a) Bias

(b) MES

Figure B.45: *ML method: 25 overall effect size estimators are compared through their bias and mean squared error by using REML method, with small sample sizes* ($n_{1j} = n_{2j}$)*, no effect sizes , and large heterogeneity, for further details see Section* B.1.1.

Overall Effect Size: ML Method

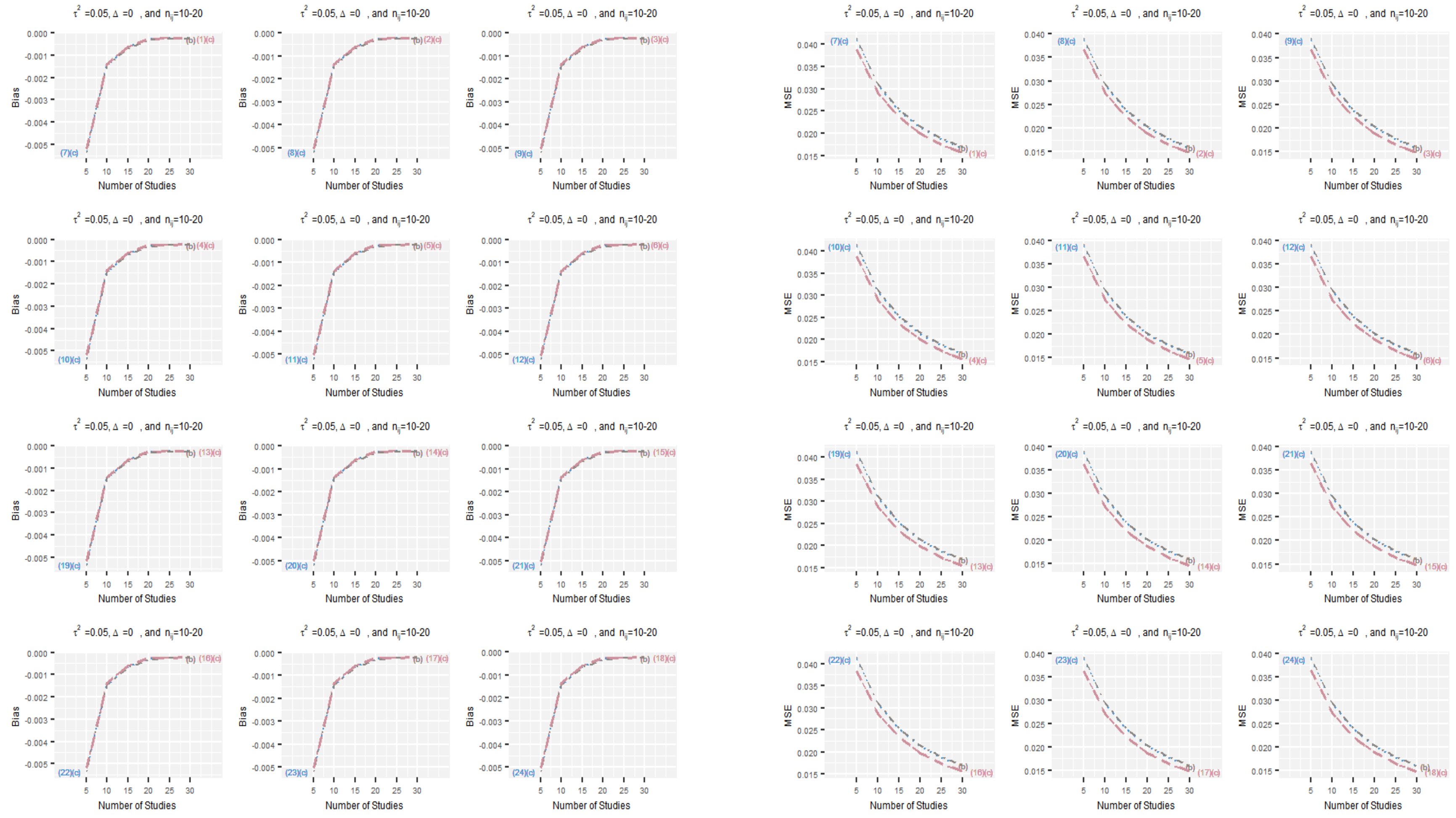


(a) Bias

(b) MES

Figure B.46: *ML method: 25 overall effect size estimators are compared through their bias and mean squared error by using REML method, with small sample sizes* ($n_{1j} = n_{2j}$)*, medium effect sizes , and large heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: ML Method

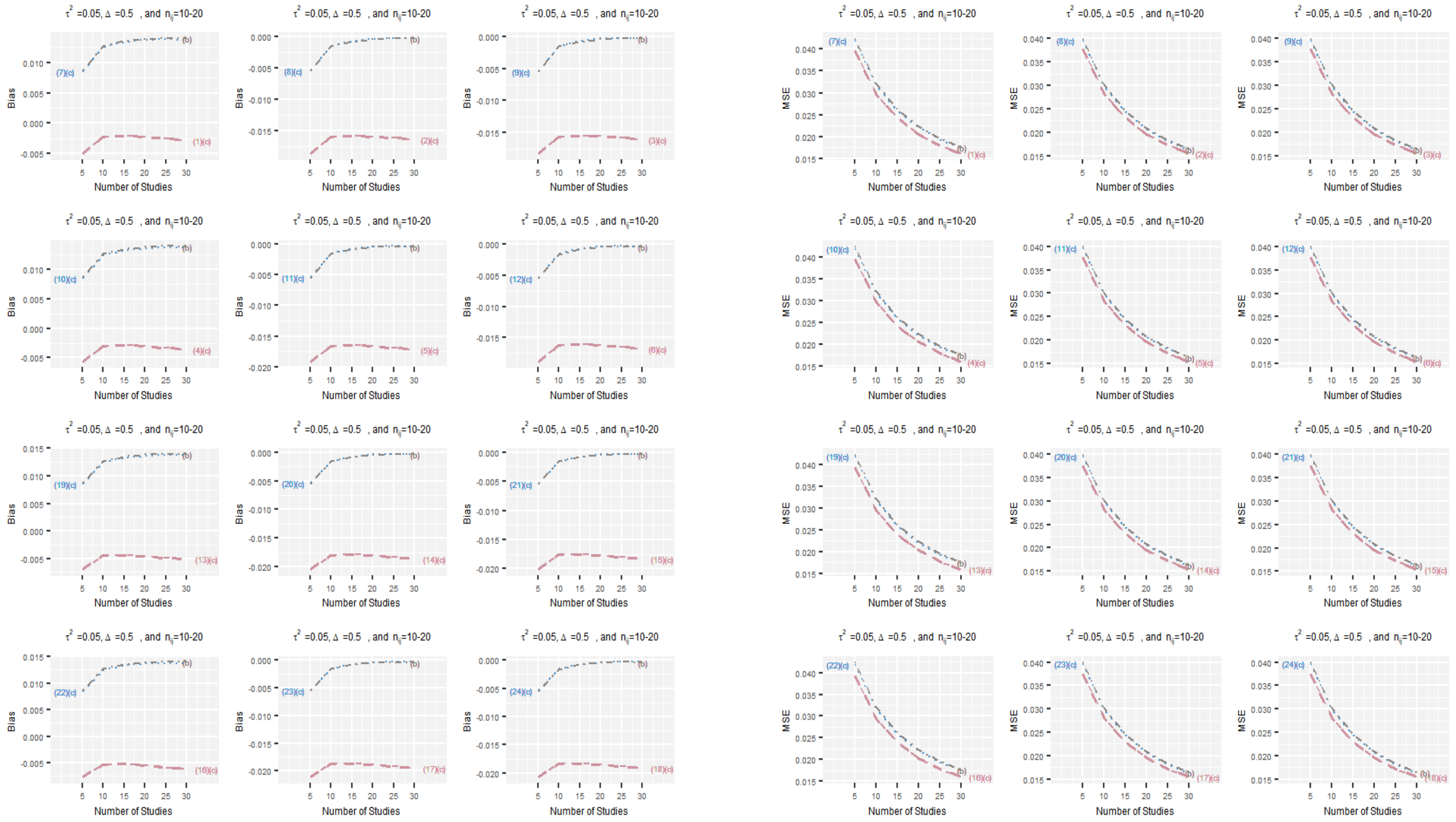


(a) Bias

(b) MES

Figure B.47: *ML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes* (n_{1j} *not necessary equal to* n_{2j})*, no effect size, and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: ML Method

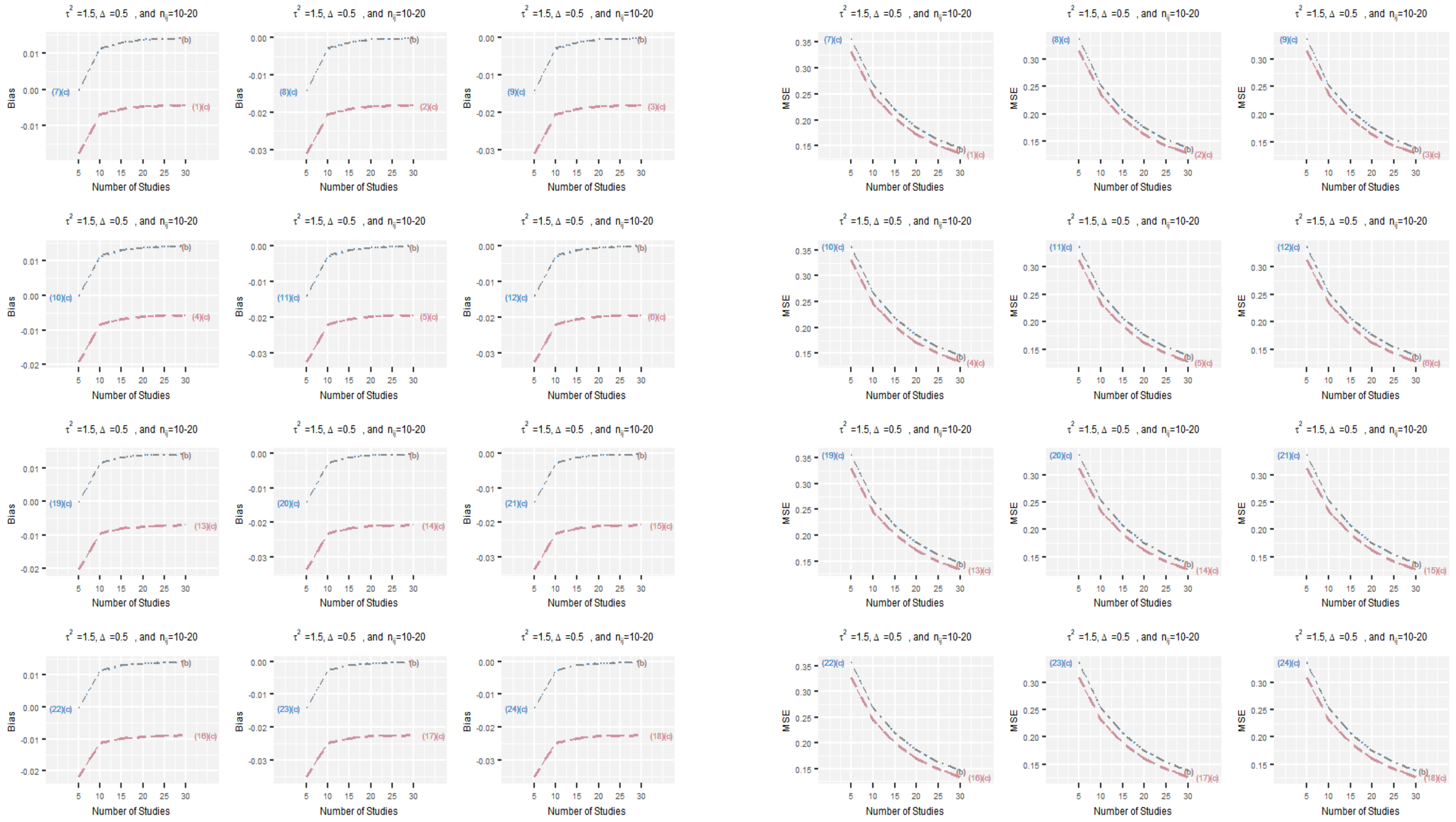


(a) Bias

(b) MES

Figure B.48: *ML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes* (n_{1j} *not necessary equal to* n_{2j}), *medium effect size, and very low heterogeneity, for further details see Section B.1.1.*

Overall Effect Size: ML Method

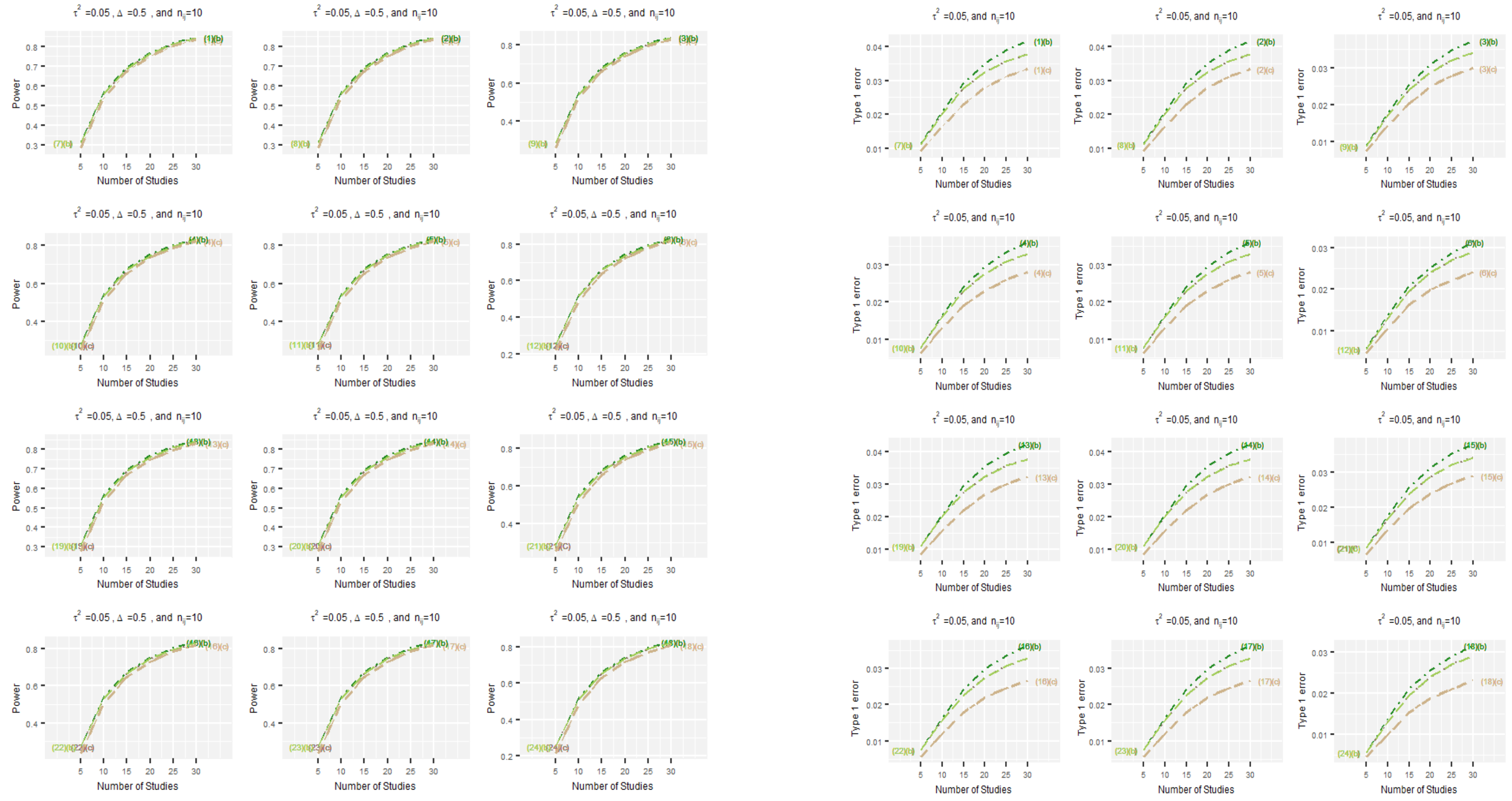


(a) Bias

(b) MES

Figure B.49: *ML method: 25 overall effect size estimators are compared through their bias and mean squared error, with small sample sizes (n_{1j} not necessary equal to n_{2j}), medium effect size, and large heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: ML Method

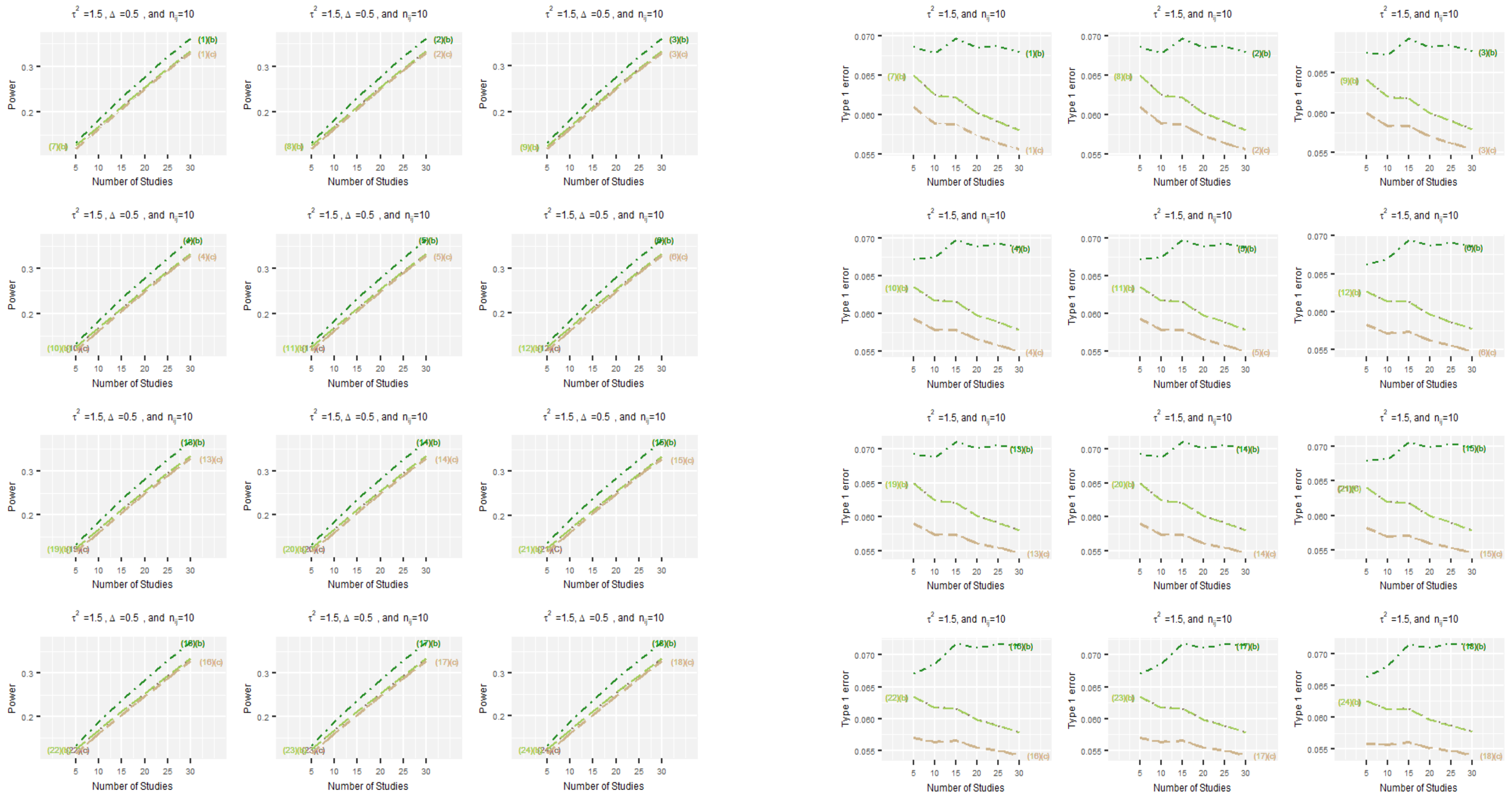


(a) Power

(b) Type 1 Error

Figure B.50: *ML method: Comparing 48 Power and Type I error by t distribution, with small sample sizes* ($n_{1j} = n_{2j}$)*, and very low heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: ML Method

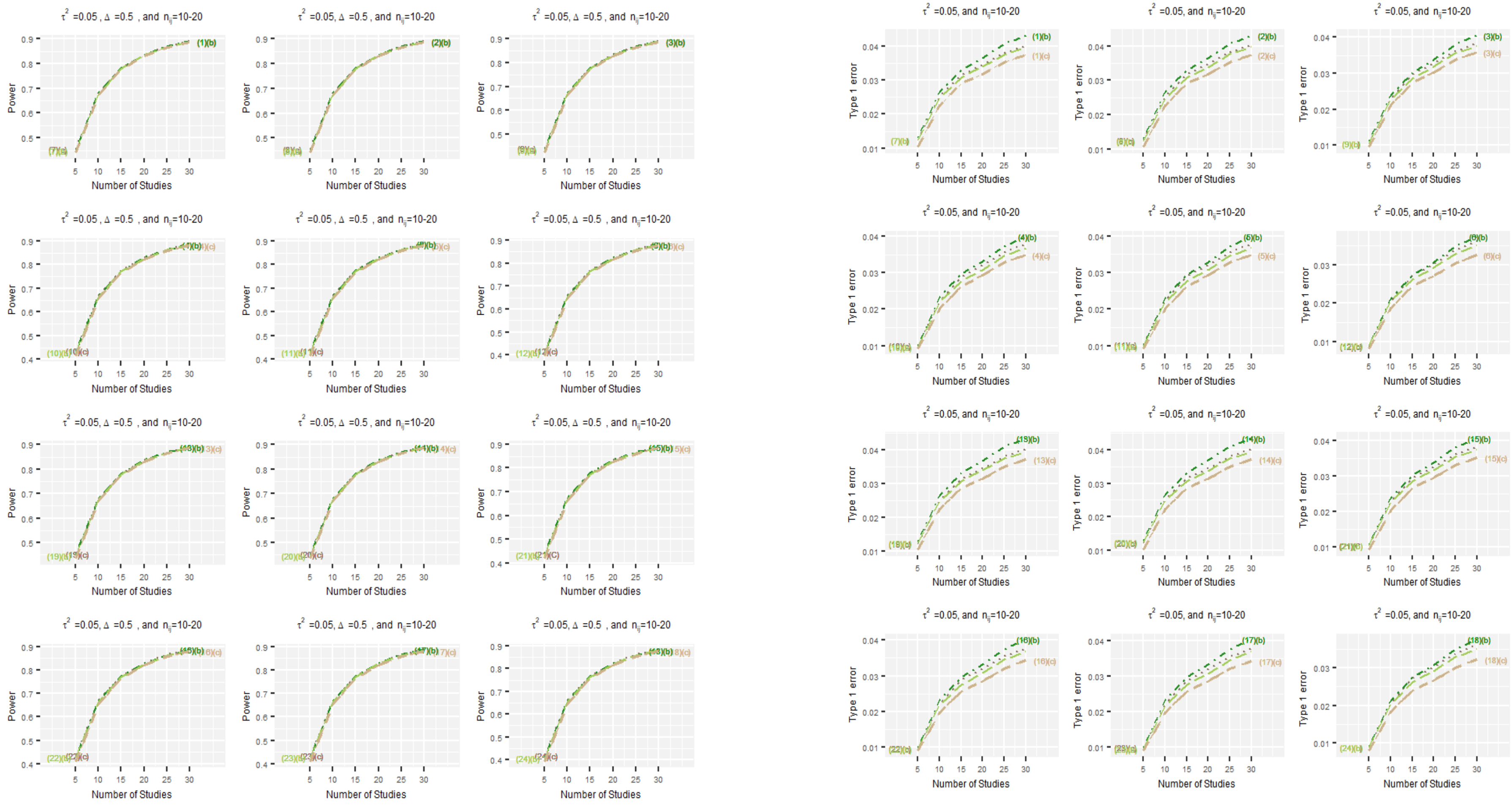


(a) Power

(b) Type 1 Error

Figure B.51: *ML method: Comparing 48 Power and Type I error by t distribution, with small sample sizes ($n_{1j} = n_{2j}$), and large heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: ML Method

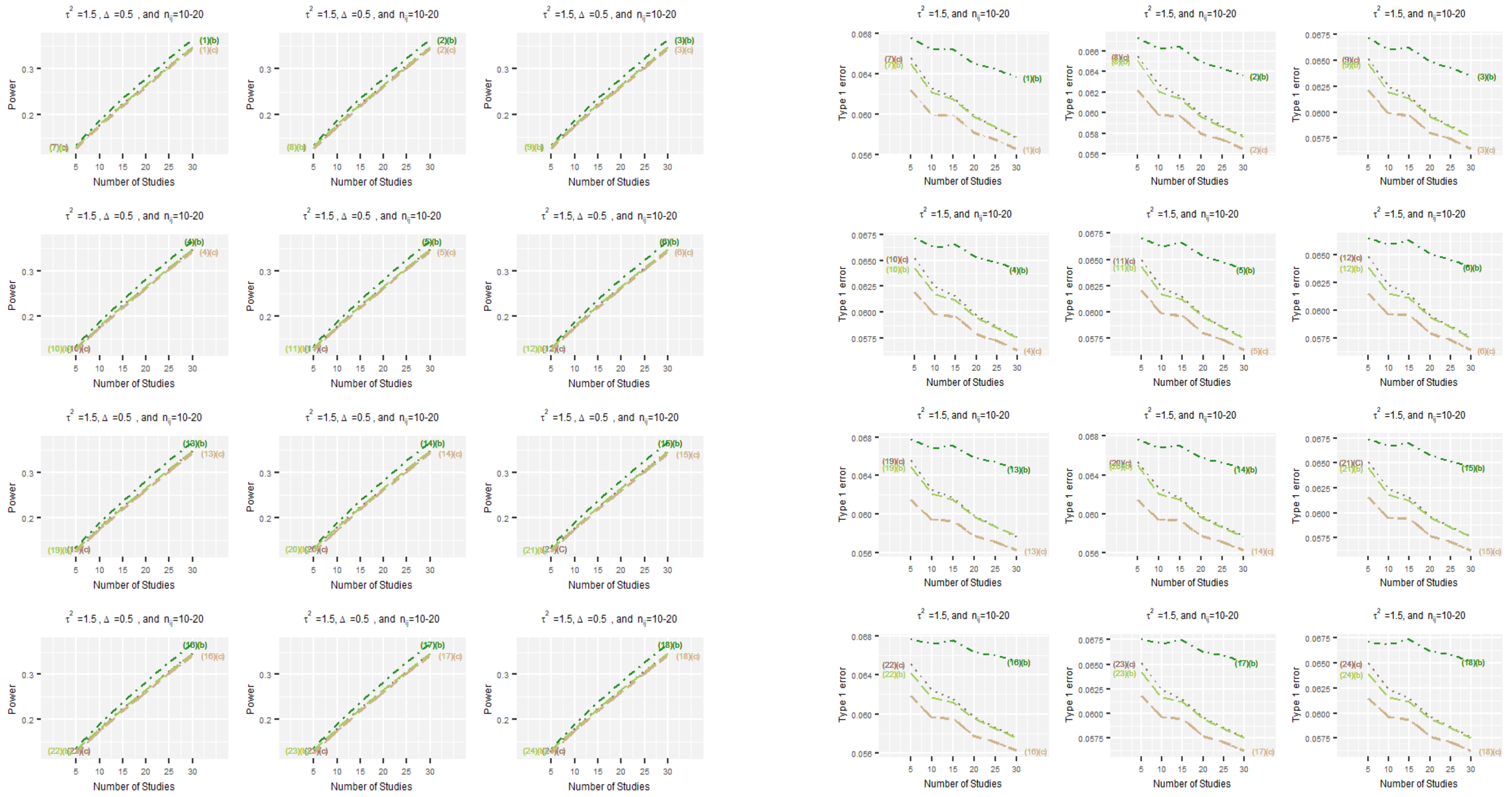


(a) Power

(b) Type 1 Error

Figure B.52: *ML method: Comparing 48 Power and Type I error by t distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), and very low heterogeneity, for further details see Section B.1.1.*

Power and Type I Error by t-distribution: ML Method



(a) Power

(b) Type 1 Error

Figure B.53: *ML method: Comparing 48 Power and Type I error by t distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), and large heterogeneity, for further details see Section B.1.1.*

Type I Error by Hartung-Knapp t-distribution: ML Method

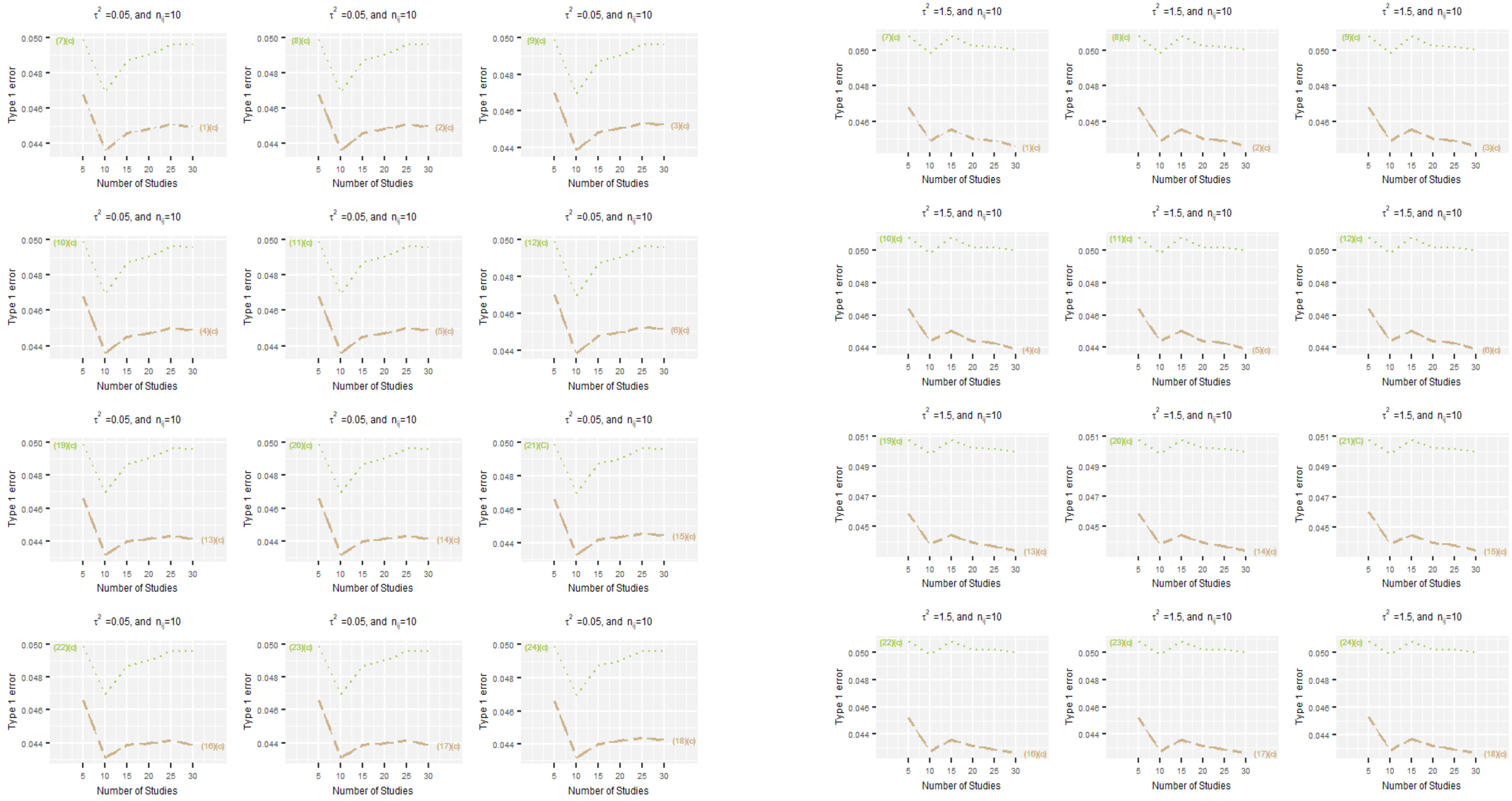


Figure B.54: *ML method: Comparing 24 Type I error by Hartung-Knapp t-distribution, with small sample sizes* ($n_{1j} = n_{2j}$)*, and very low and large heterogeneity, for further details see Section B.1.1.*

Type I Error by Hartung-Knapp t-distribution: ML Method

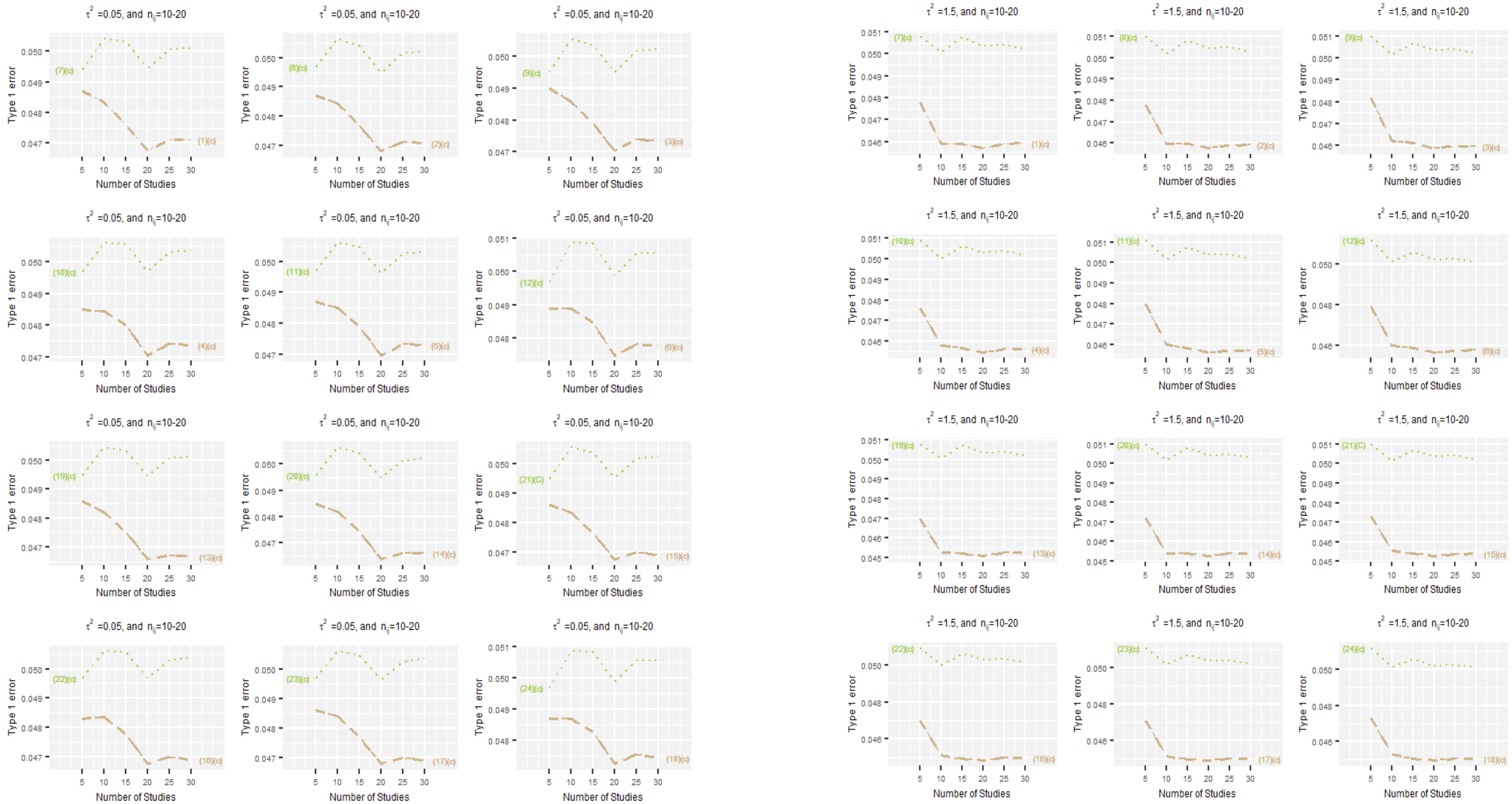


Figure B.55: *ML method: Comparing 24 Type I error by Hartung-Knapp t-distribution, with small sample sizes (n_{1j} not necessary equal to n_{2j}), and very low and large heterogeneity, for further details see Section B.1.1.*

B.1.2 Summarizing Tables of all Simulations

In Section B.1.1, we explained how the estimators were compared. None of the 48 estimators of $\hat{\tau}^2$, or 26 overall effect sizes, give optimal results in all scenarios. Thus, we will select the best quarter of the total estimators for each scenario. Because there are 48 estimators of $\hat{\tau}^2$, we will choose 12 in terms of bias, MSE, and proportion of zero, and because there are 26 overall effect sizes, we will select 6 in terms of bias and MSE (see Figure 2.1). If more than a quarter of the estimators have similar results, we will select them. Because there are a lot of different notations in different Tables, which confuse the reader, it is important to define them first.

Table B.1: *Notations*

Abbreviations	**Definitions**
(c)	Average effect size by inverse variance method
(b)	Average effect sizes by unweighted method
(1)-(24)	number of equations of within-study variance
✓	The estimator that we are looking for *
$\checkmark_c$	Using (c) method gives better result comparing to (b)
$\checkmark_{b,c}$	Using either (b) or (c) method, the estimator gives the same result.
($c_{(1)} - c_{(24)}$)	Using inverse variance method to average effect sizes by using within study variance (1) - (24), respectively.
$(b)_H$ ($(b)_C$)	Average effect sizes by unweighted method by Hedges estimator (Cohen d estimator).

* *If we look for less bias, the estimator that gives less bias compared to others will be checked, similarly, with less MSE, less proportion of zero, and so on.*

Because the heterogeneity variance estimators ($\hat{\tau}^2$) and the overall effect size ($\hat{\Delta}$) rely on s_j^2, we decide to make the comparison based on s_j^2, read Sections 2.5.3 and B.1.1 first to understand the concept.

Explanation of the Table:

Each row indicates a new simulation, so there are 40 different simulations for each Table under both study variance and overall effect size. Regarding between study variance, in each row, we compare 48 estimators. The Table's design used for selecting less bias, less MSE, and less proportion of zero is defined.

s_j^2 ⟵ 24 estimators

Table for 48 comparisons

n_{ij}	τ^2	Δ	(1)	(2)	. . . . . .	(24)
10	0.05	0				
		0.5		✓		
⋮ ⋮	⋮ ⋮	⋮ ⋮				
40-60	1.5	0.5				

↑

40 rows $\left(4(n_{ij}) \times 5(\tau^2) \times 2(\Delta) \right)$

⟵ For each row,
under each s_j^2, which method
for averaging effect size gives better
result (b), (c) or both, indicate
it by $\left(✓_b,\ ✓_c, \text{ or } ✓_{b,c} \right)$.
Select 12 best estimators.

Our simulations are based on Figure 2.1. Sometimes more than 12 estimators are selected because they are in the same range. After running all simulations for each parameter combination, the 12 columns with the most check marks will be selected as the best estimators. Similar starting we follow for flinging acceptable Type 1 error and power since there are 48 comparisons. However, the tables are given below regarding the summary effect of bias and MSE.

$\hat{\Delta}$ ⟵ inverse variance method
($c_{(1)} - c_{(24)}$), and unweighted
method ($(b)_H, (b)_C$)

Table for 26 comparisons

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	. . . .	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0						
		0.5		✓				
⋮ ⋮	⋮ ⋮	⋮ ⋮						
40-60	1.5	0.5						

For each row,
⟵ select the best six estimators
by indicating ✓

↑
40 rows ($4(n_{ij}) \times 5(\tau^2) \times 2(\Delta)$)

Under each method (REML, ML, and PM), we will have seven Tables as defined below; in total we will have 21 Tables.

Tables

Between-Study Variance

Table: Bias

n_{ij}	τ^2	Δ	(1)	(2)	. . . . . .	(24)
10	0.05	0				
		0.5				
⋮ ⋮	⋮ ⋮	⋮ ⋮				
40-60	1.5	0.5				

Table: MSE

n_{ij}	τ^2	Δ	(1)	(2)	. . . . . .	(24)
10	0.05	0				
		0.5				
⋮ ⋮	⋮ ⋮	⋮ ⋮				
40-60	1.5	0.5				

Table: Proportion of Zero

n_{ij}	τ^2	Δ	(1)	(2)	. . . . . .	(24)
10	0.05	0				
		0.5				
⋮ ⋮	⋮ ⋮	⋮ ⋮				
40-60	1.5	0.5				

Overall Effect Size

Table: Bias

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	. . . .	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0						
		0.5						
⋮ ⋮	⋮ ⋮	⋮ ⋮						
40-60	1.5	0.5						

Table: MSE

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	. . . .	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0						
		0.5						
⋮ ⋮	⋮ ⋮	⋮ ⋮						
40-60	1.5	0.5						

Table: Type 1 error by Hurting t-distribution

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	. . . .	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0						
		0.5						
⋮ ⋮	⋮ ⋮	⋮ ⋮						
40-60	1.5	0.5						

Table: Type 1 error by t-distribution

n_{ij}	τ^2	Δ	(1)	(2)	. . . . . .	(24)
10	0.05	0				
		0.5				
⋮ ⋮	⋮ ⋮	⋮ ⋮				
40-60	1.5	0.5				

Our selection for each parameter combination is based on the overall behavior of the estimator through a different number of studies. For example, if the first estimator gives 0 bias with a small number of studies and gives 0.04 to a large number of studies, the range for m=5 to 30 is [0,0.04]. However, the second estimator gives 0.02 bias with a small number of studies and gives -0.02 to a large number of studies; the range for m=5 to 30 is [0.02,-0.02]. We select the second estimator.

Table B.2: *Less MSE of Between-study variance by REML method*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0					$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5					$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0					$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5					$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
10-20	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
50	0.05	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
40-60	0.05	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$, *we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the Table can be found in Section B.1.2. Note, for unequal sample size, using the (b) method is slightly less zero produced, but since the difference is tiny, we ignore it.*

Table B.3: *Less bias of Between-study variance by REML method*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0			$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
	0.3	0		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6	0				$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5				$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5	0				$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5				$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
10-20	0.05	0			$\checkmark_{b,c}$	$\checkmark_{c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5			$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
	0.1	0			$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$
		0.5			$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$
	0.3	0				$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5				$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6	0				$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5				$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5	0	$\checkmark_{b,c}$			$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5	$\checkmark_{b,c}$			$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$
50	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.3	0	$\checkmark_{b,c}$			$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
		0.5	$\checkmark_{b,c}$			$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6	0				$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	
														$\checkmark_{b,c}$												
		0.5				$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	
														$\checkmark_{b,c}$												
	1.5	0				$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5				$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
40-60	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.3	0		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6	0	$\checkmark_{b,c}$			$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$			$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5	$\checkmark_{b,c}$			$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$			$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5	0	$\checkmark_{b,c}$			$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$			$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5	$\checkmark_{b,c}$			$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$			$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* ($n_{1j} = n_{2j}$), *we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the Table can be found in Section B.1.2.*

Table B.4: *Less proportion of zero heterogeneity variance estimates by REML method*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
10-20	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
50	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
40-60	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the table can be found in Section B.1.2.*

Table B.5: *Less MSE of overall effect size by REML method*

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
10-20	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
50	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								

(To be continued)

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
40-60	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								

The symbol $\left(c_{(1)} - c_{(24)}\right)$ indicates the averaging effect size by inverse variance method under different within-study variance estimators $\left((1) - (24)\right)$, which are defined in Section 2.2. However, $(b)_H$ $((b)_C)$ means that averaging effect size by using the unweighted method by Hedges g estimator (Cohen d estimator). The n_{ij} indicates the sample sizes of each group. With equal sample size $(n_{1j} = n_{2j})$, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60. τ^2 indicates the between-study variance, and Δ indicates the effect size. An explanation of the table can be found in Section B.1.2.

Table B.6: *Less MSE of overall effect size by REML method*

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0	✓	✓	✓	✓	✓	✓						✓	✓	✓	✓	✓	✓									
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
10-20	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
50	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		

(To be continued)

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		✓
40-60	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		✓

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$, *we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the table can be found in Section B.1.2.* ***Note:*** *when* $\Delta = 0$, *the differences between all estimators under small equal sample sizes are very small. Thus, the selecting estimators are slightly better than others. However, when* $\Delta = 0.5$, *the differences are high.*

Table B.7: *Acceptable Type I error and power by t distribution of overall effect size by REML method*

n_{ij}	τ^2	(1)	(2)	(3)	(4)	(5)	(6)	(13)	(14)	(15)	(16)	(17)	(18)	(7)	(8)	(9)	(10)	(11)	(12)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
	1.5													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
10-20	0.05													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	0.1													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	0.3													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$		$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	
	0.6													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$		$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	
	1.5													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$		$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	
50	0.05													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.1													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.3													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
	0.6													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
	1.5													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
40-60	0.05													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$
	0.1													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$
	0.3													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance. An explanation of the Table can be found in Section B.1.2.*

Table B.8: *Acceptable Type I error by Hartung-Knapp t-distribution of overall effect size by REML method*

n_{ij}	τ^2	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$
10	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10-20	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
50	0.05	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												
	0.1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40-60	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The symbol $\left(c_{(1)} - c_{(24)}\right)$ *indicates the averaging effect size by inverse variance method under different within-study variance estimators* $\left((1) - (24)\right)$*, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance. An explanation of the Table can be found in Section B.1.2.*

Table B.9: *Less MSE of Between-study variance by PM*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
10-20	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
50	0.05	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
40-60	0.05	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						

The numbers $\big((1)-(24)\big)$ indicate the within-study variance estimators, which are defined in Section 2.2. The n_{ij} indicates the sample sizes of each group. With equal sample size ($n_{1j} = n_{2j}$), we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60. τ^2 indicates the between-study variance, and Δ indicates the effect size. An explanation of the Table can be found in Section B.1.2.

Table B.10: *Less bias of Between-study variance by PM method*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0					$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5					$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
	0.1	0					$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5					$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$
	1.5	0		$\checkmark_{b,c}$							$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$							$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
10-20	0.05	0					$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5					$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
	0.1	0					$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5																								
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5	0									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$					$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$
50	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
	0.1	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5																								
40-60	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						$\checkmark_{b,c}$
	0.1	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$							$\checkmark_{b,c}$
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$						$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$						$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j}=n_{2j})$, *we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the table can be found in Section B.1.2.*

Table B.11: *Less proportion of zero heterogeneity variance estimates by PM method*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
10-20	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
50	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
40-60	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the Table can be found in Section B.1.2.*

Table B.12: *Less MSE of overall effect size by PM method*

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
10-20	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
50	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								

(To be continued)

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
40-60	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								

The symbol $\left(c_{(1)} - c_{(24)}\right)$ *indicates the averaging effect size by inverse variance method under different within-study variance estimators* $\left((1) - (24)\right)$*, which are defined in Section 2.2. However,* $(b)_H$ $((b)_C)$ *means that averaging effect size by using the unweighted method by Hedges g estimator (Cohen d estimator). The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the Table can be found in Section B.1.2.*

Table B.13: *Less bias of overall effect size by PM method*

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5	✓							✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5	✓							✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.6	0		✓	✓		✓	✓		✓	✓		✓	✓		✓	✓		✓	✓			✓	✓			✓	✓
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	1.5	0		✓	✓		✓	✓		✓	✓		✓	✓		✓	✓		✓	✓			✓	✓			✓	✓
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
10-20	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5	✓							✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5	✓							✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
50	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5	✓							✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

(To be continued)

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
		0.5	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓
40-60	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		0.5	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓

The symbol $\left(c_{(1)} - c_{(24)}\right)$ *indicates the averaging effect size by inverse variance method under different within-study variance estimators* $\left((1) - (24)\right)$*, which are defined in Section 2.2. However,* $(b)_H$ $((b)_C)$ *means that averaging effect size by using the unweighted method by Hedges g estimator (Cohen d estimator). The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the Table can be found in Section B.1.2.* ***Note:*** *when* $\Delta = 0$*, the differences between all estimators under small equal sample sizes are very small. Thus, the selecting estimators are slightly better than others. However, when* $\Delta = 0.5$*, the differences are high.*

Table B.14: *Acceptable Type I error by t distribution of overall effect size by PM method*

n_{ij}	τ^2	(1)	(2)	(3)	(4)	(5)	(6)	(13)	(14)	(15)	(16)	(17)	(18)	(7)	(8)	(9)	(10)	(11)	(12)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
10-20	0.05													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	0.1													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	0.3													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	0.3													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	1.5													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$
50	0.05													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.1													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.3													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
40-60	0.05													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.1													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.3													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* ($n_{1j} = n_{2j}$)*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance. An explanation of the table can be found in Section B.1.2.*

Table B.15: *Acceptable Type I error by Hartung-Knapp t-distribution of overall effect size by PM method*

n_{ij}	τ^2	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$
10	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10-20	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
50	0.05	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												
	0.1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40-60	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The symbol $\left(c_{(1)} - c_{(24)}\right)$ *indicates the averaging effect size by inverse variance method under different within-study variance estimators* $\left((1) - (24)\right)$*, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance. An explanation of the table can be found in Section B.1.2.*

Table B.16: *Less MSE of Between-study variance by ML method*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
10-20	0.05	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5			$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
50	0.05	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
40-60	0.05	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.1	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.3	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	0.6	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
	1.5	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$	$\checkmark_{b,c}$						

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$, *we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the table can be found in Section B.1.2.*

Table B.17: *Less bias of Between-study variance by ML method*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5		$\checkmark_{b,c}$	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.1	0	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
		0.5	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.6	0																								
		0.5																								
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
10-20	0.05	0	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
		0.5	$\checkmark_{b,c}$								$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.6	0																								
		0.5																								
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
50	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
40-60	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$									$\checkmark_{b,c}$	$\checkmark_{b,c}$		$\checkmark_{b,c}$		
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the table can be found in Section B.1.2.*

Table B.18: *Less proportion of zero heterogeneity variance estimates by ML method*

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
10-20	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$										$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
50	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$

(To be continued)

n_{ij}	τ^2	Δ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
40-60	0.05	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.1	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.3	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.6	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	1.5	0							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
		0.5							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$							$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$, *we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the table can be found in Section B.1.2.*

Table B.19: *Less MSE of overall effect size by ML method*

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
10-20	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
50	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								

(To be continued)

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
40-60	0.05	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.1	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.3	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	0.6	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								
	1.5	0		✓	✓		✓	✓								✓	✓		✓	✓								
		0.5		✓	✓		✓	✓								✓	✓		✓	✓								

The symbol $\left(c_{(1)} - c_{(24)}\right)$ *indicates the averaging effect size by inverse variance method under different within-study variance estimators* $\left((1) - (24)\right)$*, which are defined in Section 2.2. However,* $(b)_H$ $((b)_C)$ *means that averaging effect size by using the unweighted method by Hedges g estimator (Cohen d estimator). The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the table can be found in Section B.1.2.*

Table B.20: *Less bias of overall effect size by ML method*

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
10	0.05	0	✓	✓	✓	✓	✓	✓							✓	✓	✓	✓	✓	✓								
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓							✓	✓	✓	✓	✓	✓								
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
10-20	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
50	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		

(To be continued)

n_{ij}	τ^2	Δ	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$	$(b)_C$	$(b)_H$
		0.5							✓	✓		✓	✓								✓	✓		✓	✓			✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		✓
40-60	0.05	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.1	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.3	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	0.6	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5								✓	✓		✓	✓								✓	✓		✓	✓		✓
	1.5	0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
		0.5	✓			✓				✓	✓		✓	✓	✓			✓				✓	✓		✓	✓		✓

The symbol $\left(c_{(1)} - c_{(24)}\right)$ *indicates the averaging effect size by inverse variance method under different within-study variance estimators* $\left((1) - (24)\right)$*, which are defined in Section 2.2. However,* $(b)_H$ $((b)_C)$ *means that averaging effect size by using the unweighted method by Hedges g estimator (Cohen d estimator). The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance, and* Δ *indicates the effect size. An explanation of the table can be found in Section B.1.2.* ***Note:*** *when* $\Delta = 0$*, the differences between all estimators under small equal sample sizes are very small. Thus, the selecting estimators are slightly better than others. However, when* $\Delta = 0.5$*, the differences are high.*

Table B.21: *Acceptable Type I error by t distribution of overall effect size by ML method*

n_{ij}	τ^2	(1)	(2)	(3)	(4)	(5)	(6)	(13)	(14)	(15)	(16)	(17)	(18)	(7)	(8)	(9)	(10)	(11)	(12)	(19)	(20)	(21)	(22)	(23)	(24)
10	0.05													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.1													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.3													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$				$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$			
	0.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	1.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
10-20	0.05													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	0.1													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	0.3													$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$				$\checkmark_{c}$	$\checkmark_{c}$	$\checkmark_{c}$			
	0.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	1.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
50	0.05													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.1													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	0.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	1.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
40-60	0.05													$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$	$\checkmark_{b,c}$
	0.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	0.3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	0.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	1.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

The numbers $\big((1)-(24)\big)$ *indicate the within-study variance estimators, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$, *we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance. (–) indicates they are inflated. An explanation of the table can be found in Section B.1.2.*

Table B.22: *Acceptable Type I error by Hartung-Knapp t-distribution of overall effect size by ML method*

n_{ij}	τ^2	$c_{(1)}$	$c_{(2)}$	$c_{(3)}$	$c_{(4)}$	$c_{(5)}$	$c_{(6)}$	$c_{(7)}$	$c_{(8)}$	$c_{(9)}$	$c_{(10)}$	$c_{(11)}$	$c_{(12)}$	$c_{(13)}$	$c_{(14)}$	$c_{(15)}$	$c_{(16)}$	$c_{(17)}$	$c_{(18)}$	$c_{(19)}$	$c_{(20)}$	$c_{(21)}$	$c_{(22)}$	$c_{(23)}$	$c_{(24)}$
10	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10-20	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
50	0.05	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												
	0.1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40-60	0.05													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.1													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.3													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	0.6													✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
	1.5													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The symbol $\left(c_{(1)} - c_{(24)}\right)$ *indicates the averaging effect size by inverse variance method under different within-study variance estimators* $\left((1) - (24)\right)$*, which are defined in Section 2.2. The* n_{ij} *indicates the sample sizes of each group. With equal sample size* $(n_{1j} = n_{2j})$*, we indicate the sample size by 10 or 50, but if the sample sizes are not necessarily equal, we define it as the range of two numbers 10-20 or 40-60.* τ^2 *indicates the between-study variance. An explanation of the table can be found in Section B.1.2.*

B.2 Mixed Model Approach

The same data sets were generated under meta-analysis (see Section 2.5.1) will be used in this section. Simulation studies are conducted under the REML method and ML method. For each simulation, we evaluate the performance of the overall effect size by using the *lmer* function. Only a subset of the simulations are provided. To read more about these methods, read section 2.4.

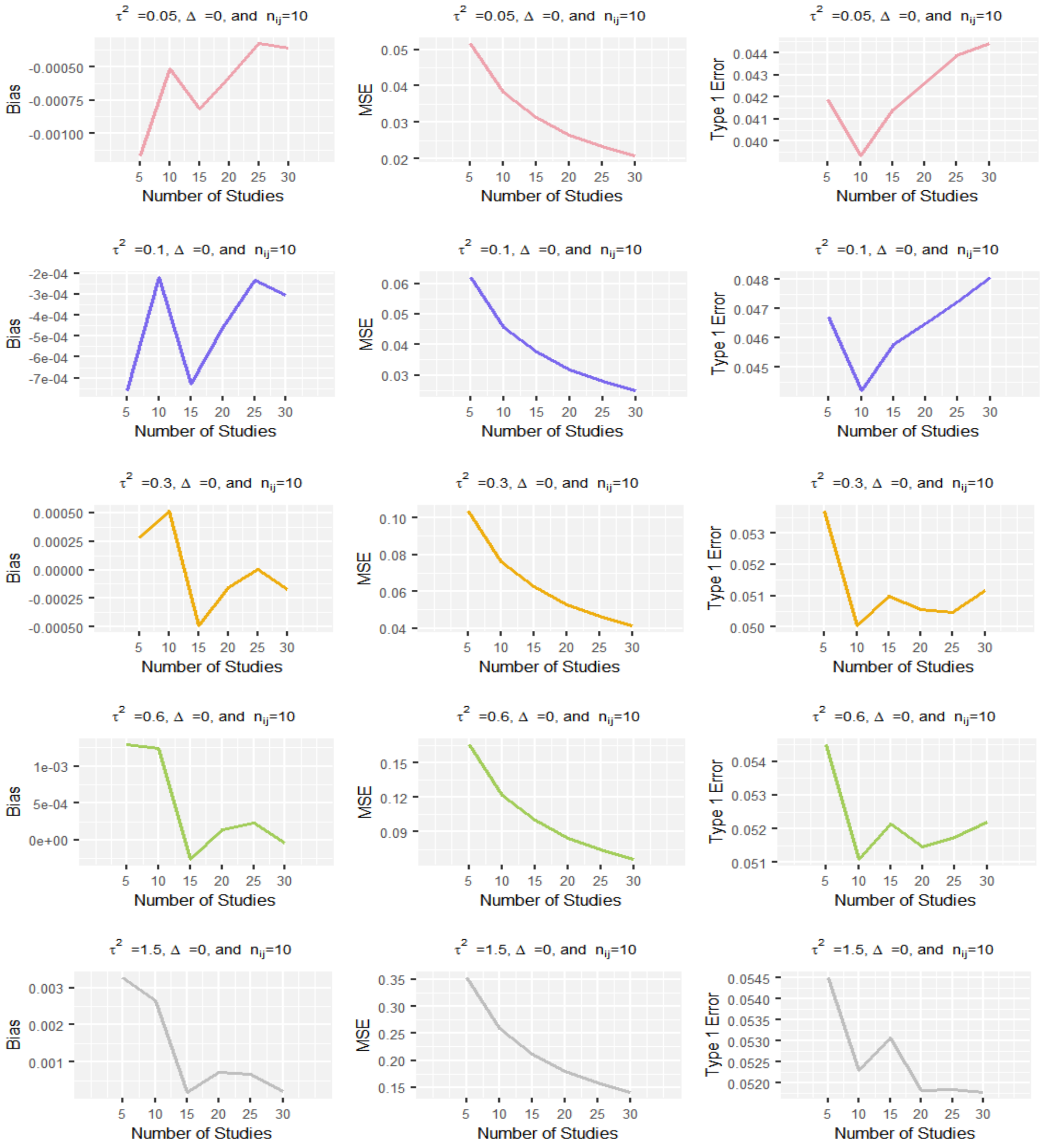


Figure B.56: *Properties of REML estimator under mixed model with equal small sample size and no effect*

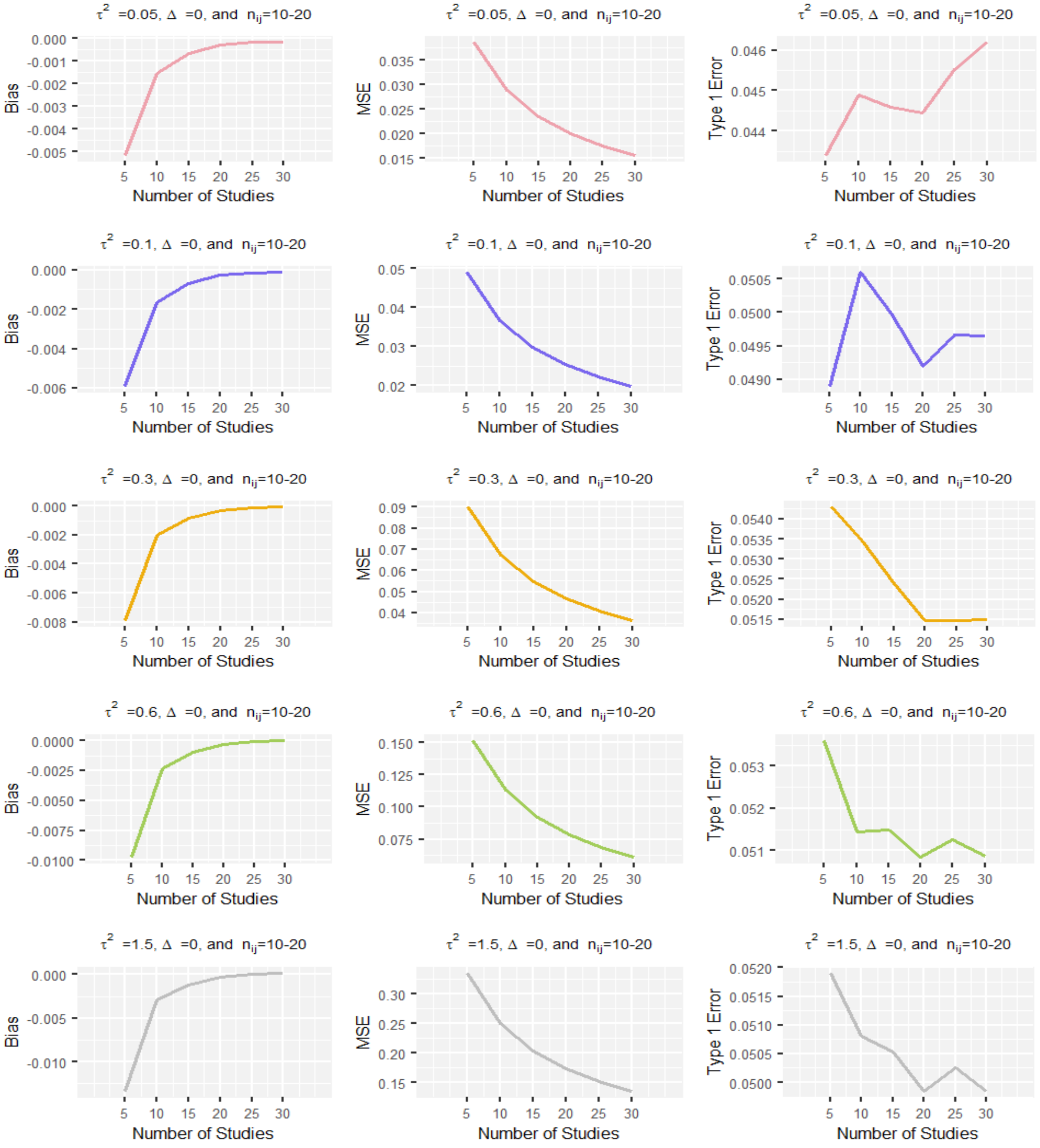


Figure B.57: *Properties of REML estimator under mixed model with not necessary equal small sample size and no effect*

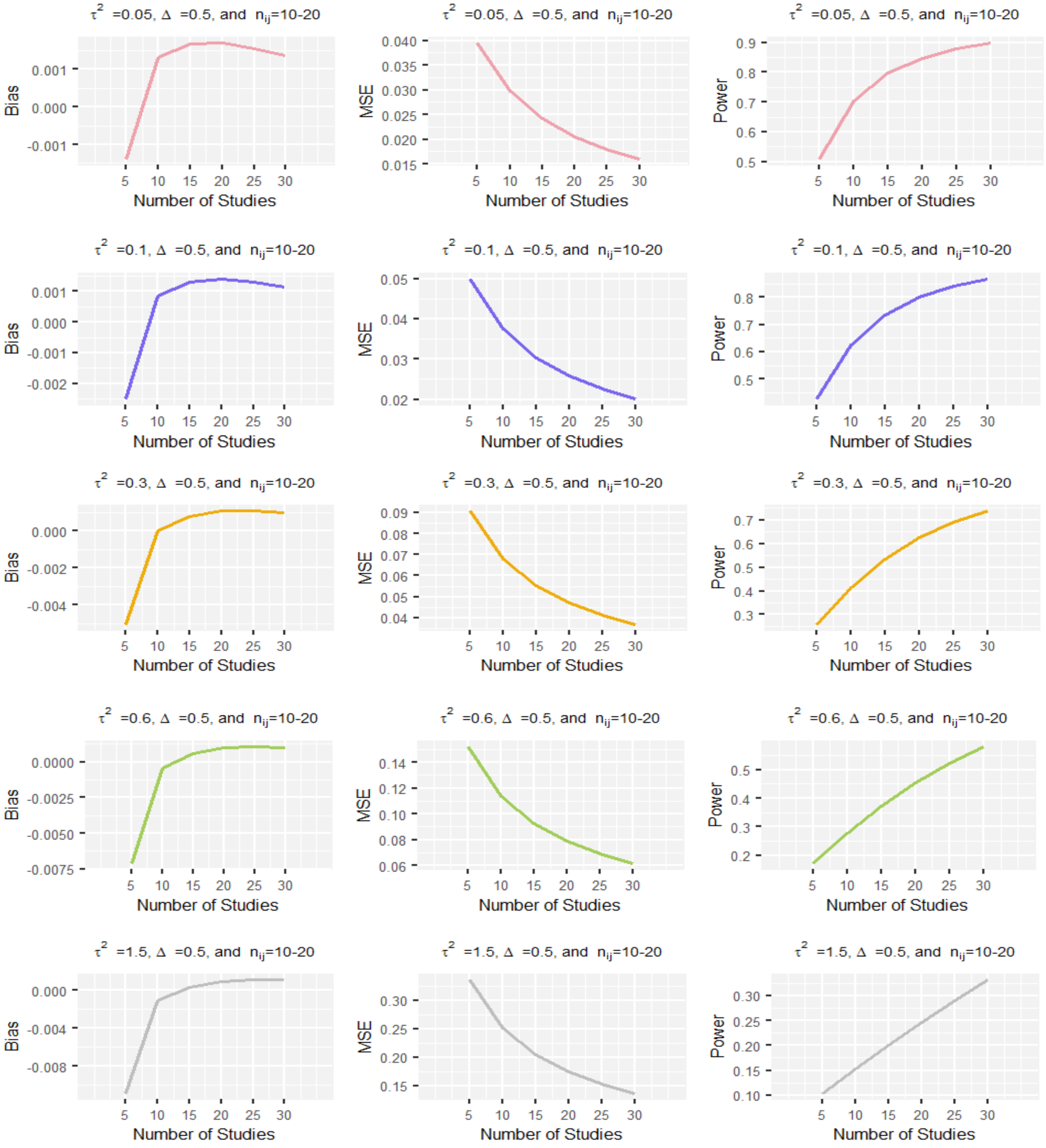


Figure B.58: *Properties of REML estimator under mixed model with not necessary equal small sample size and medium effect*

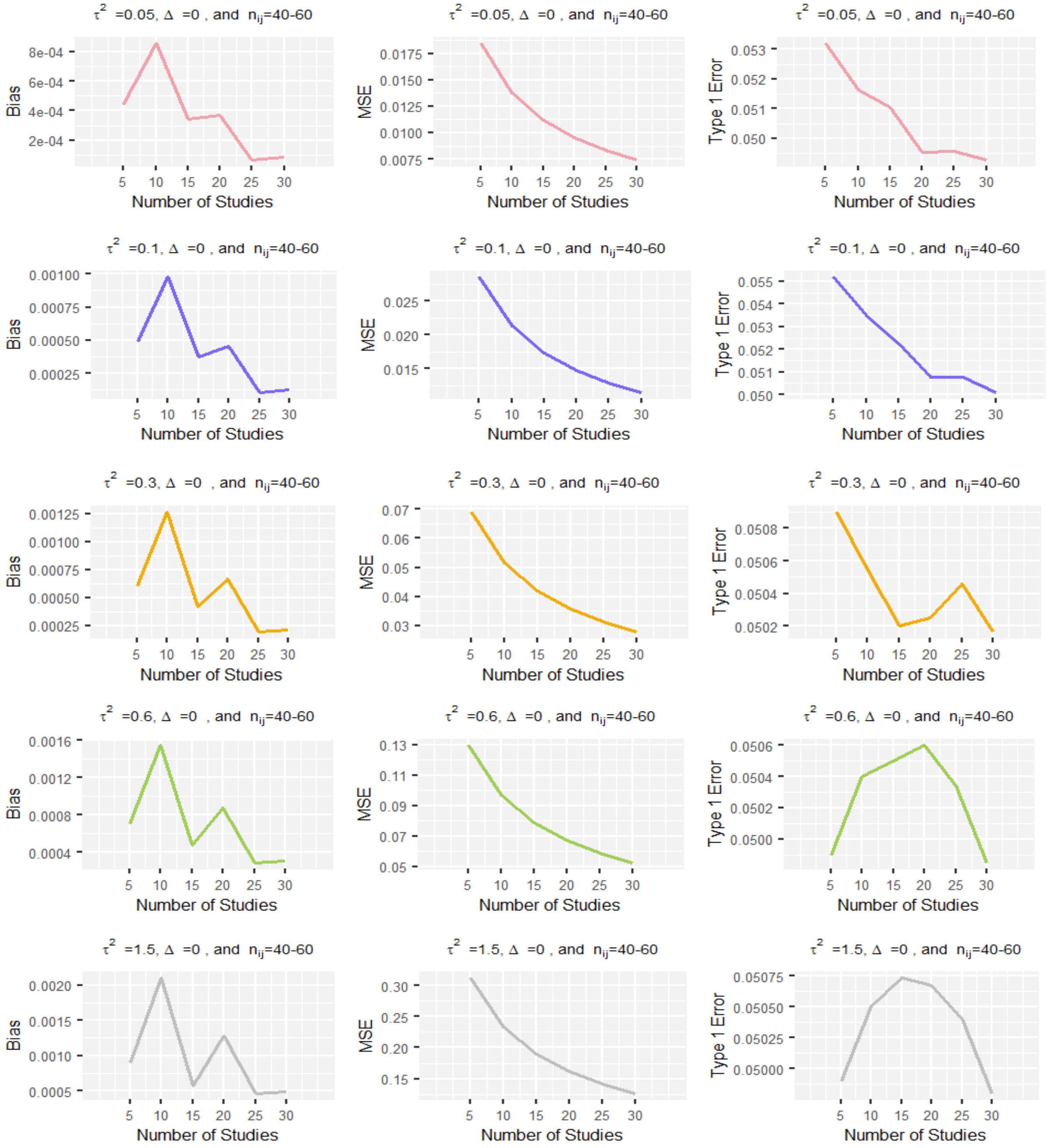


Figure B.59: *Properties of REML estimator under mixed model with not necessary equal large sample size and no effect*

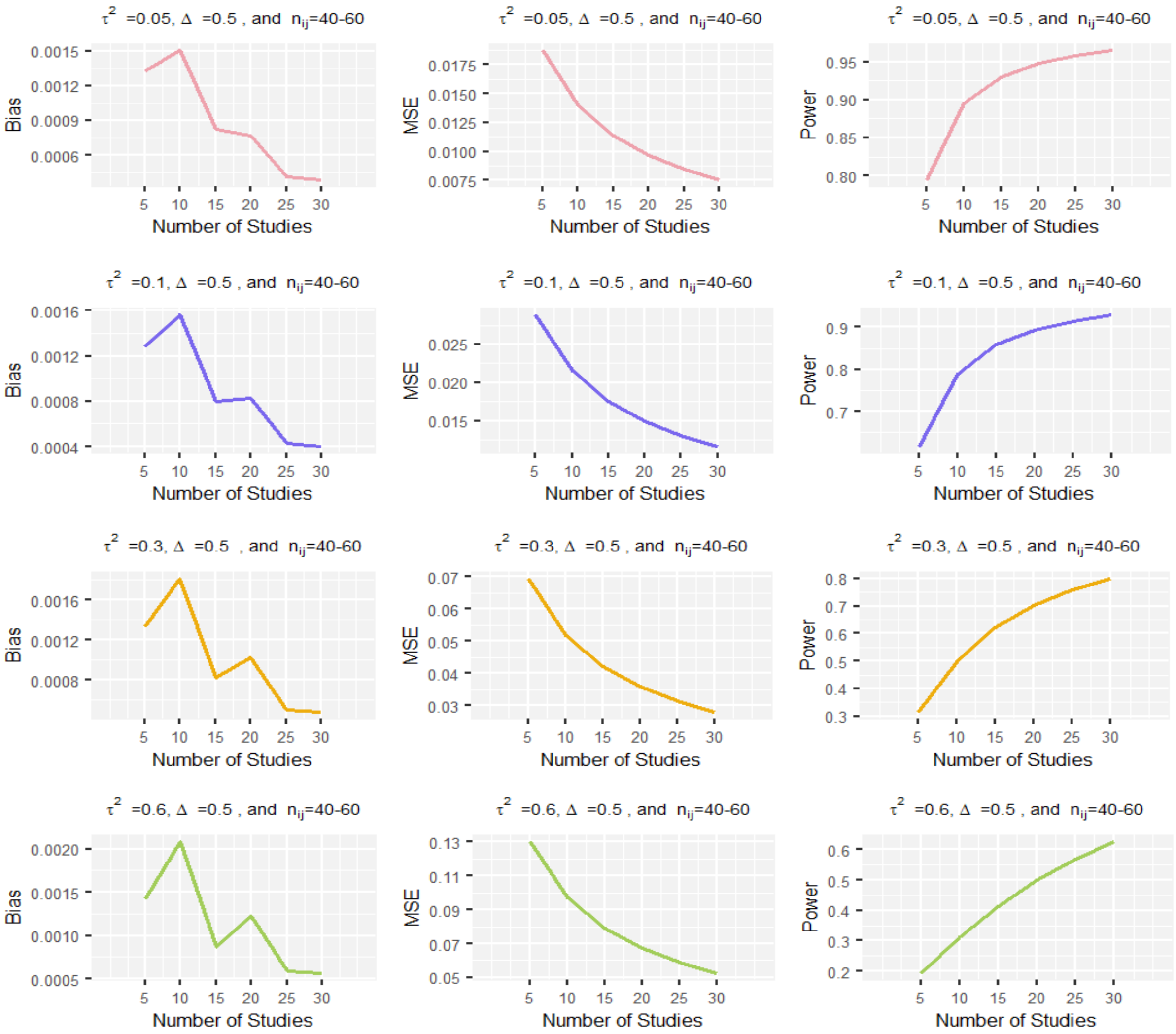


Figure B.60: *Properties of REML estimator under mixed model with not necessary equal large sample size and medium effect*

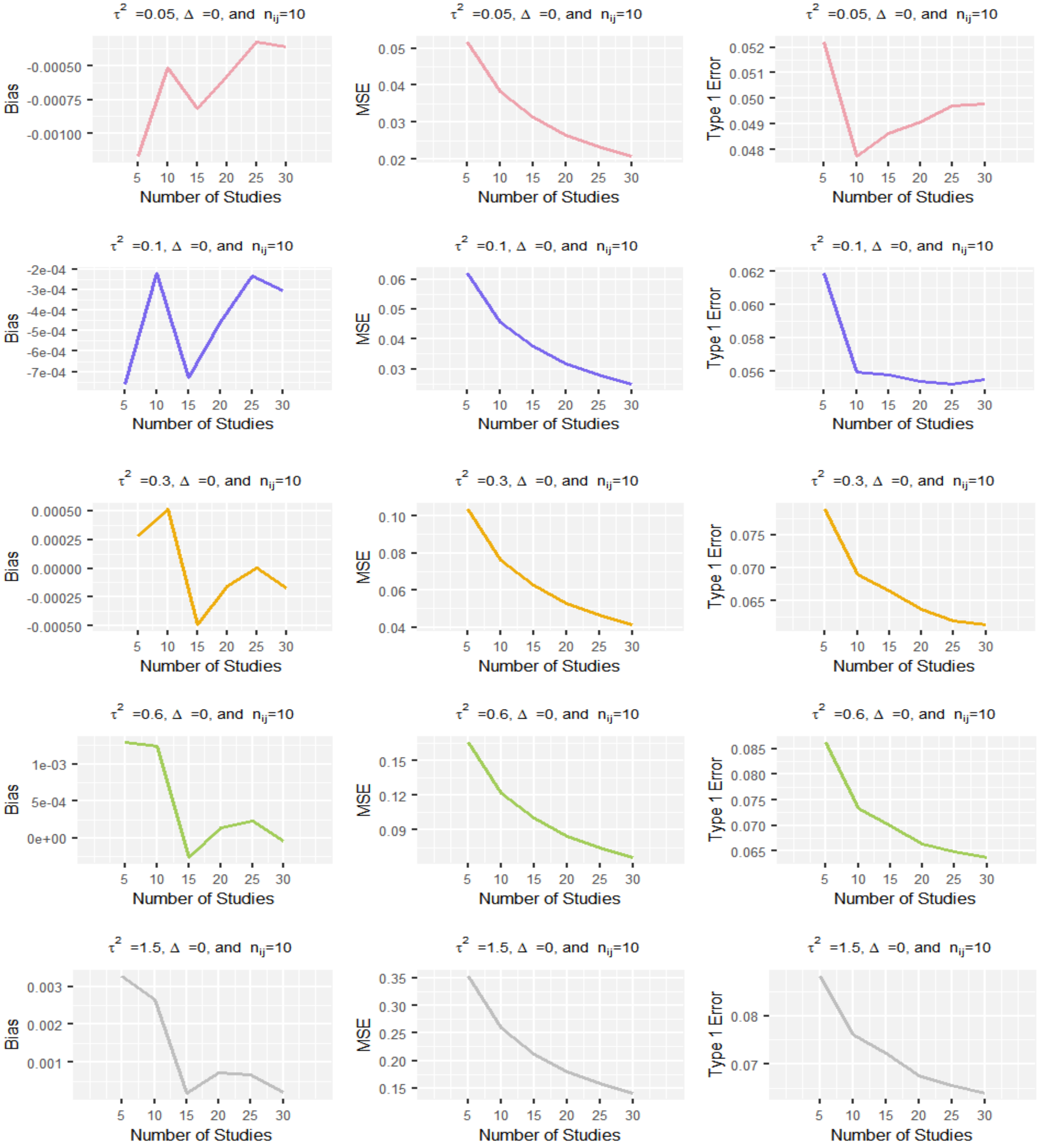


Figure B.61: *Properties of ML estimator under mixed model with equal small sample size and no effect*

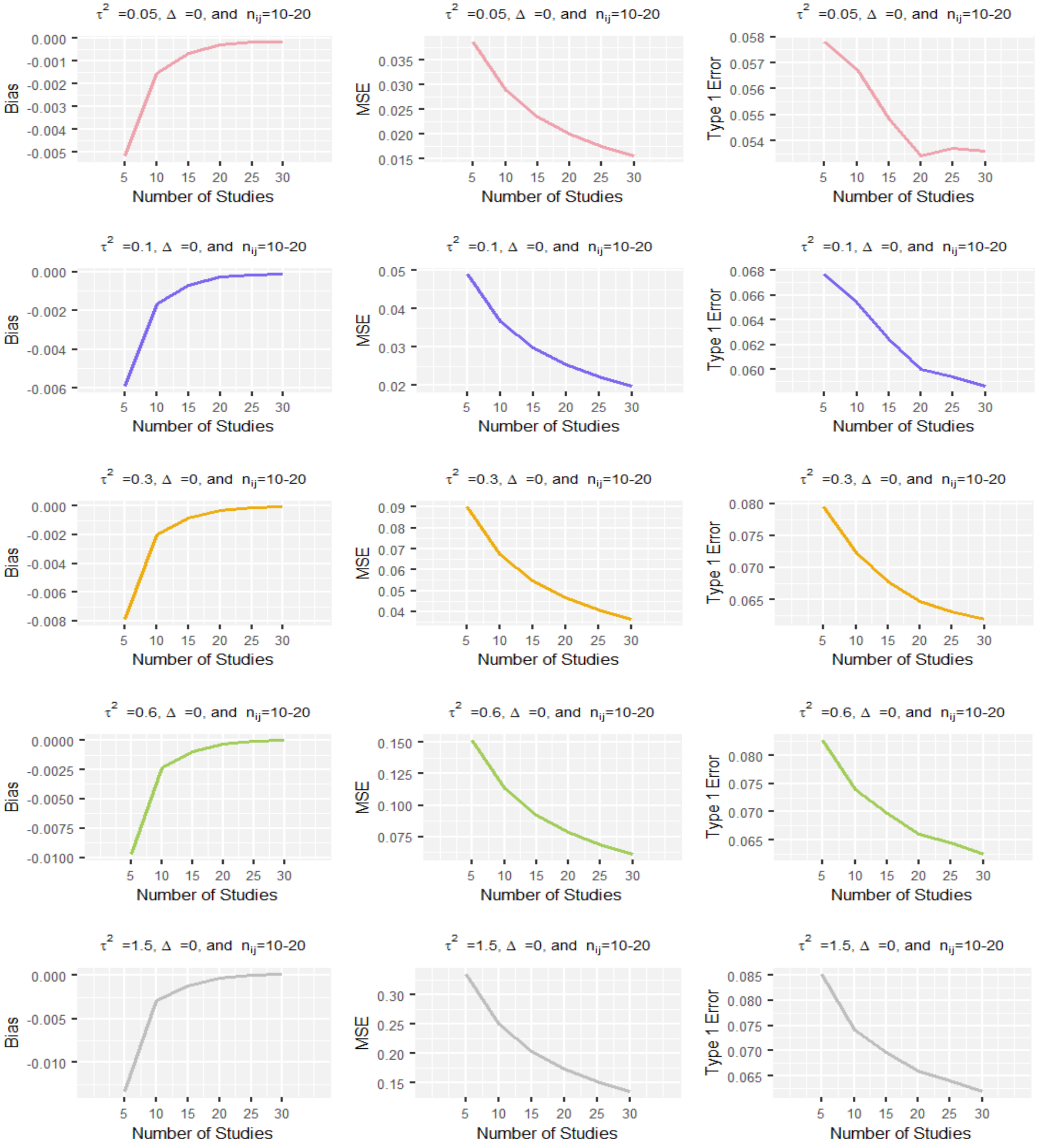


Figure B.62: *Properties of ML estimator under mixed model with not necessary equal small sample size and no effect*

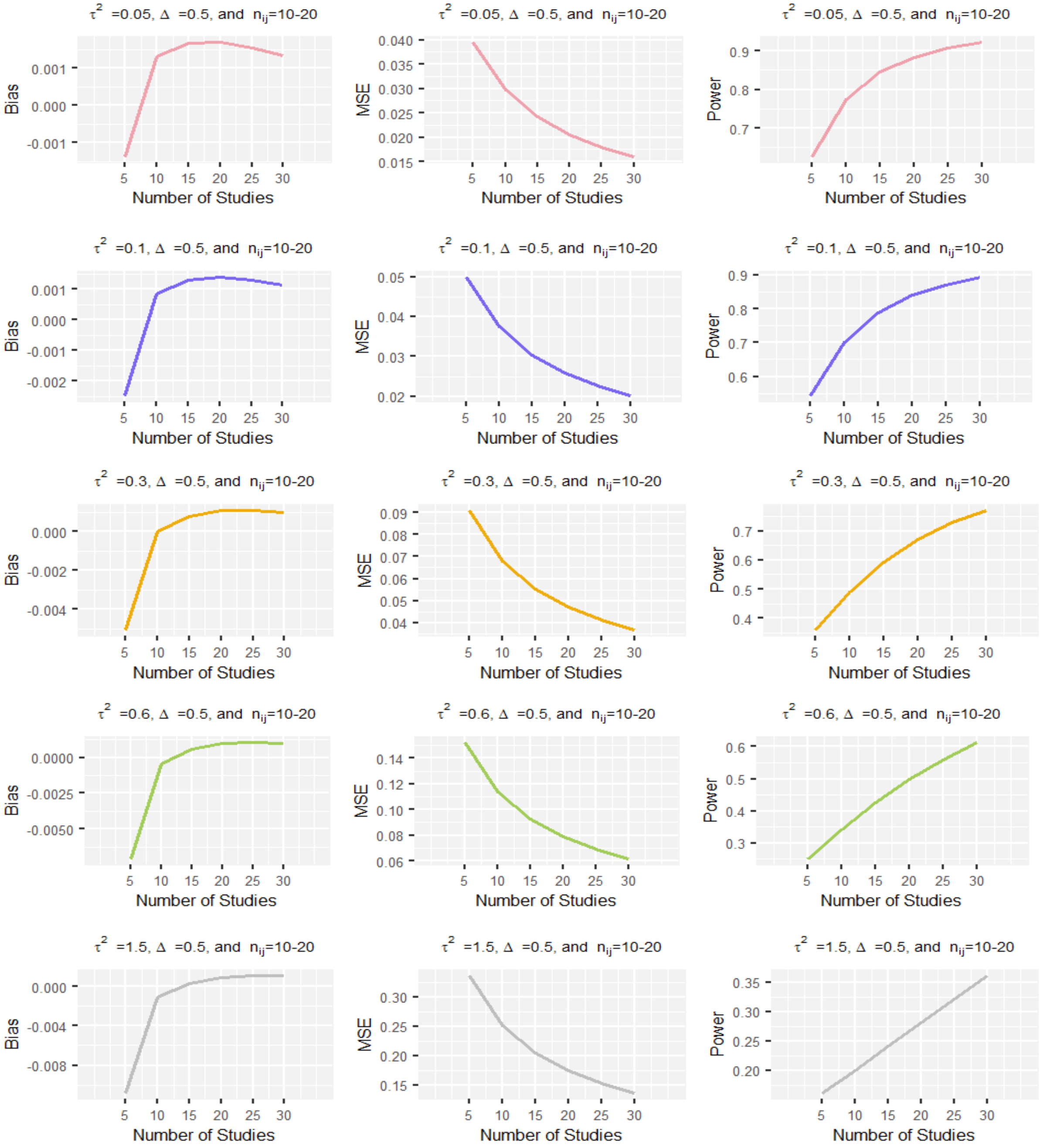


Figure B.63: *Properties of ML estimator under mixed model with not necessary equal small sample size and medium effect*

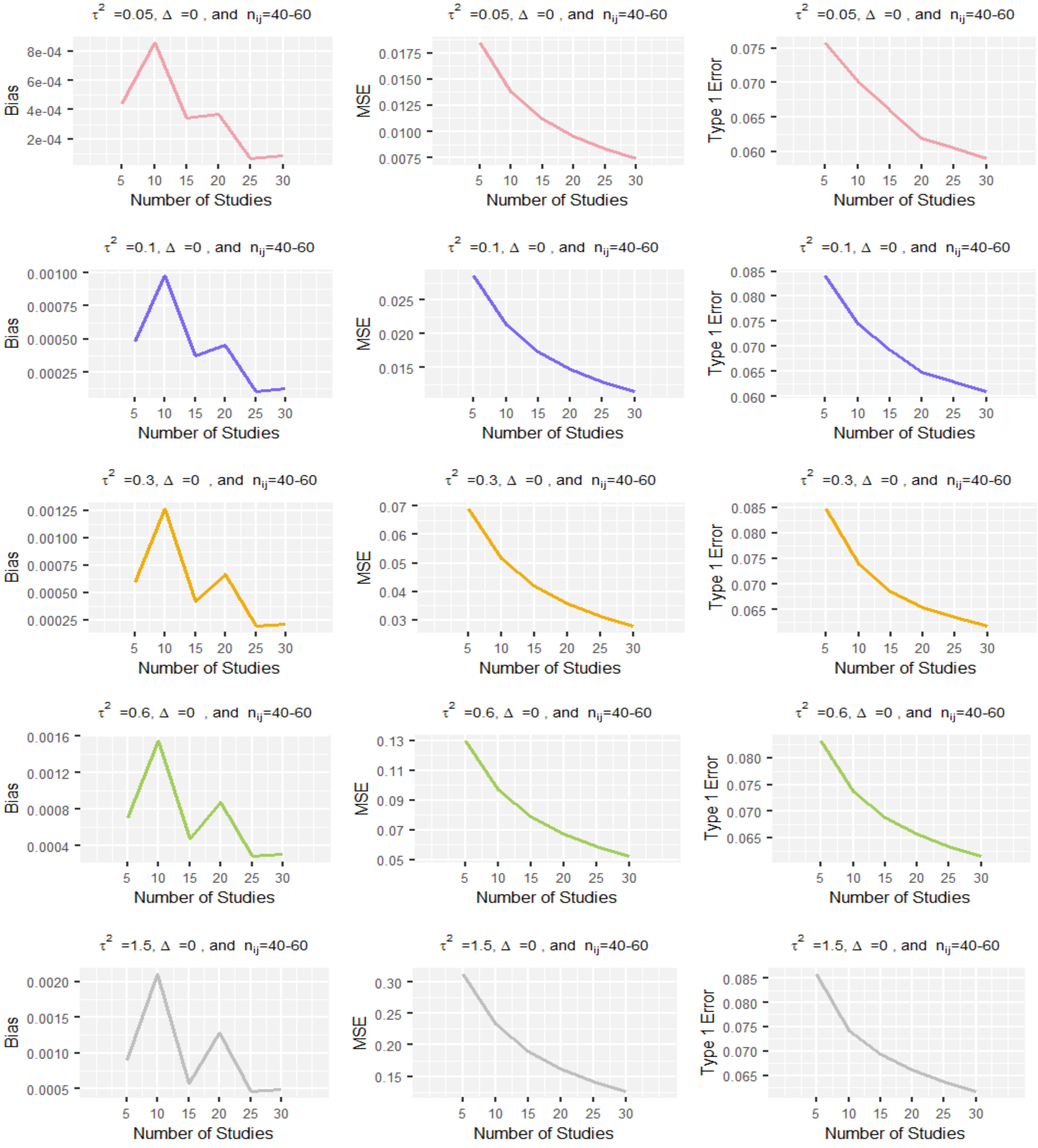


Figure B.64: *Properties of ML estimator under mixed model with not necessary equal large sample size and no effect*

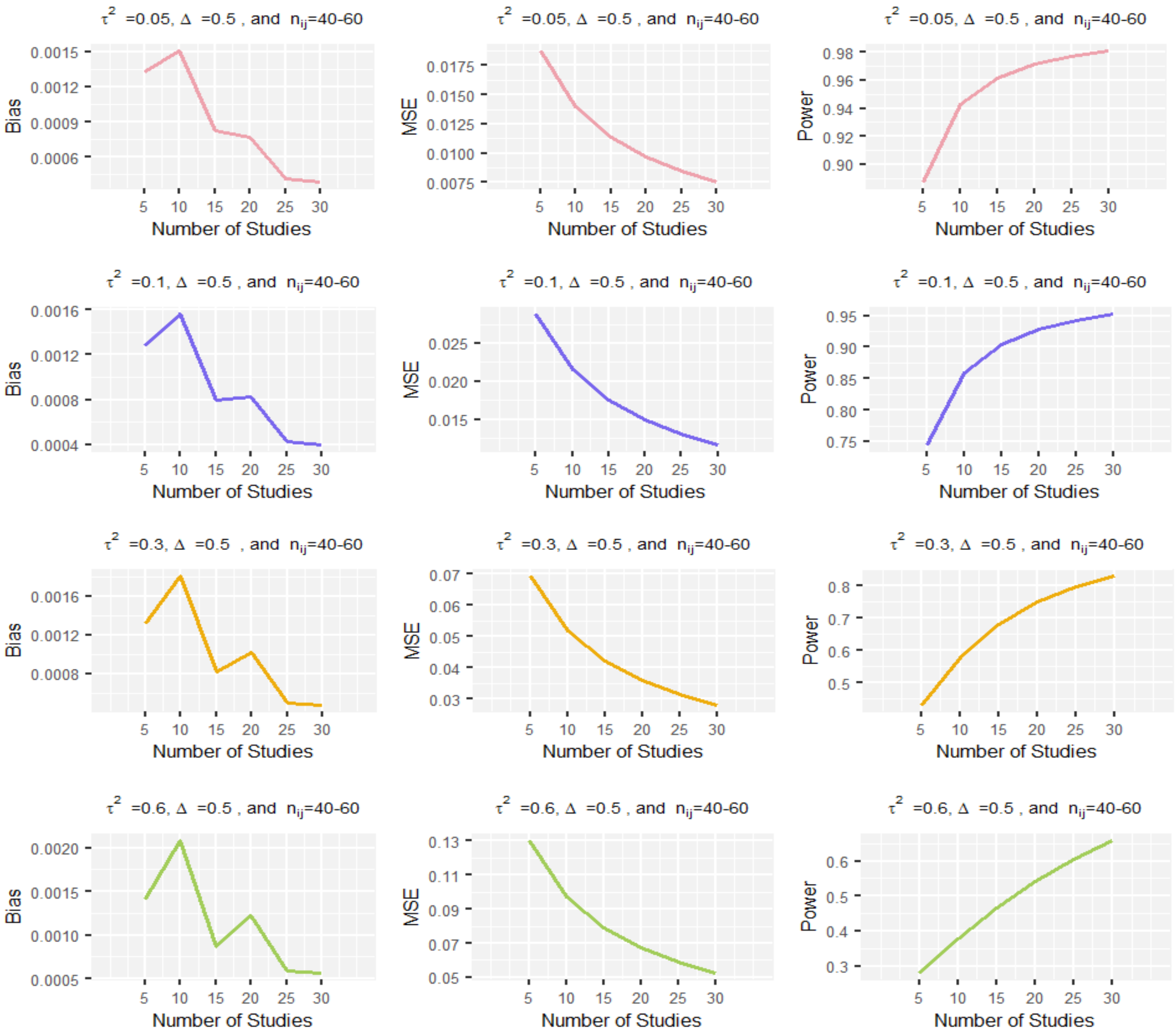


Figure B.65: *Properties of ML estimator under mixed model with not necessary equal large sample size and medium effect*

Appendix C

Sensitivity Analyses Approach in Chapter 3

We develop a sensitive analysis of the environmental effect ratio (EER) for determining the significance of effect size estimates. Various values for EER will be considered to determine for what values the effect size estimation is statistically significant. We apply this method to both meta-analyses and mixed model, see Section C.1 and C.2 respectively.

C.1 Meta-Analysis

Under different scenarios, $\hat{\tau}^2_{REML}$ and $\hat{\tau}^2_{ML}$ are estimated after using the optimal estimating of within-study variance, which is

$$s^2_{H.j} = J^2_j \tilde{n_j} + \frac{\hat{\Delta}^2_H}{2(n_{1j} + n_{2j})}$$

where $\hat{\Delta}^2_H$ is estimated by inverse variance method, see Section 2.5 for more details. In the figures below, τ indicates heterogeneity variance, Δ indicates overall effect size, n_{ij} indicates sample size for each group (n_{1j} for treatment 1 and n_{2j} for control) and m indicates small size. To making inferences about $\mu_1 - \mu_2$, the t-statistic by both t-distribution and Hartung-

Knapp t-distribution are used.

Under each parameter value combination, five steps would be done. First, we evaluate the performance of estimating between-study variance (REML and ML) in giving false significant (non-significant) results. We compare the result$\Big($ significance (non-significant) $\Big)$ that is given by known value $\tau^2 = 2EER^2$ with REML (ML) estimation. Second, we evaluate each method's performance in producing Type 1 errors. Thus, known value of $\tau^2 = 2EER^2$ is used to find when $\Delta = 0$ gives significance result. Third, the relationship between sample size, number of studies, heterogeneity, effect size, and statistically significant results is found. We want to show which term plays an important role in replication and which does not have a significant effect. Fourth, we evaluate the work of sensitive analysis by showing if it can help the researchers determine for what EER values the effect size estimate is statistically significant.

a) Meta-analysis with t-distribution

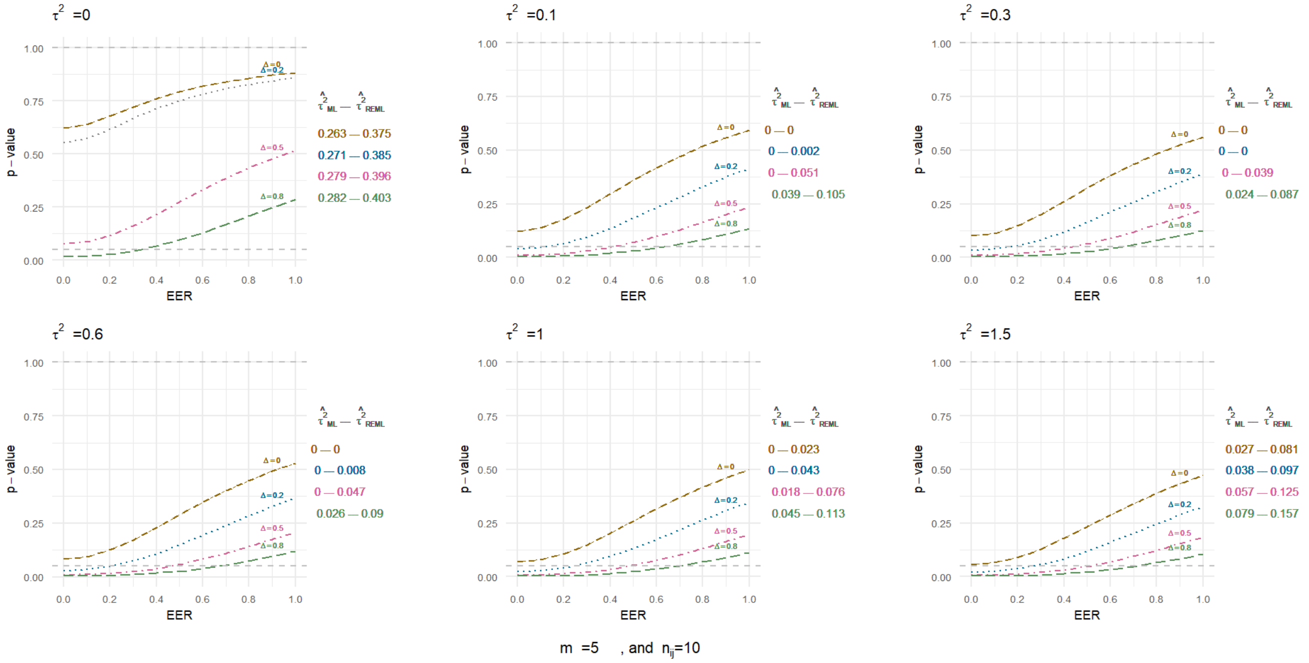


Figure C.1: *Meta-analysis by t-distribution: Sensitive analysis with various value of τ and Δ under small n_{ij}, where $n_{1j} = n_{2j}$ and small m. For further details see Section C.1*

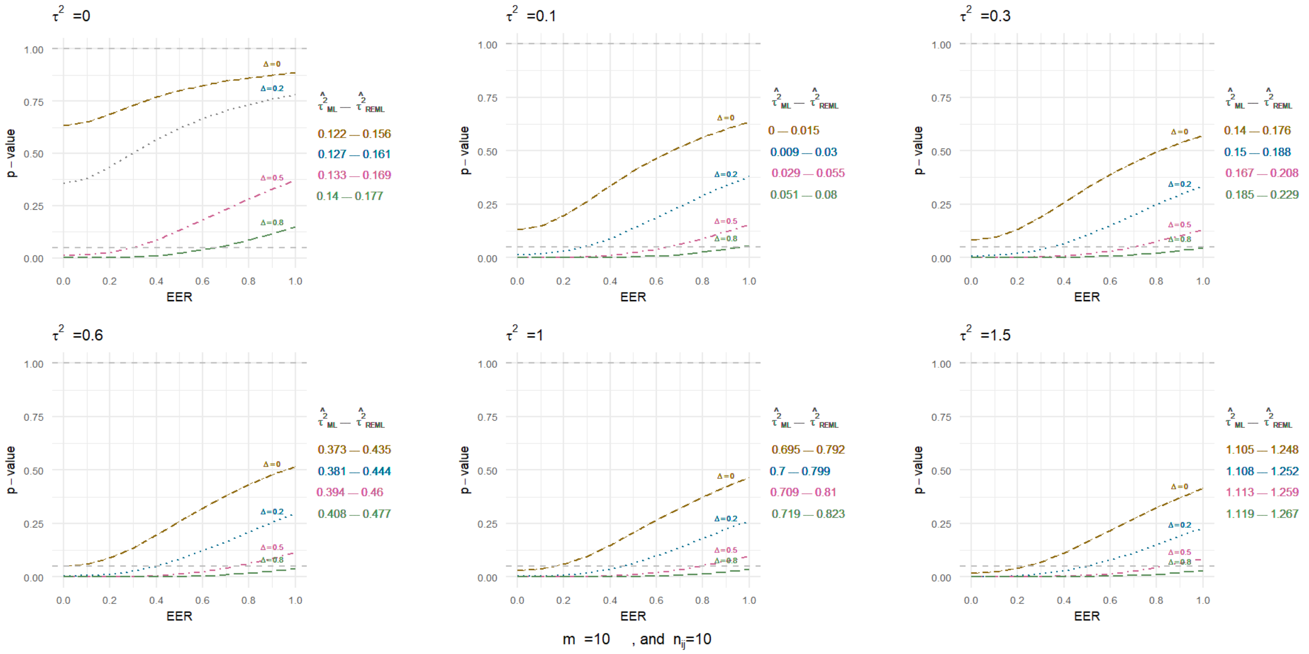


Figure C.2: *Meta-analysis (t-distribution): Sensitive analysis with various value of τ and Δ under small n_{ij}, where $n_{1j} = n_{2j}$ and medium m. For further details see Section C.1*

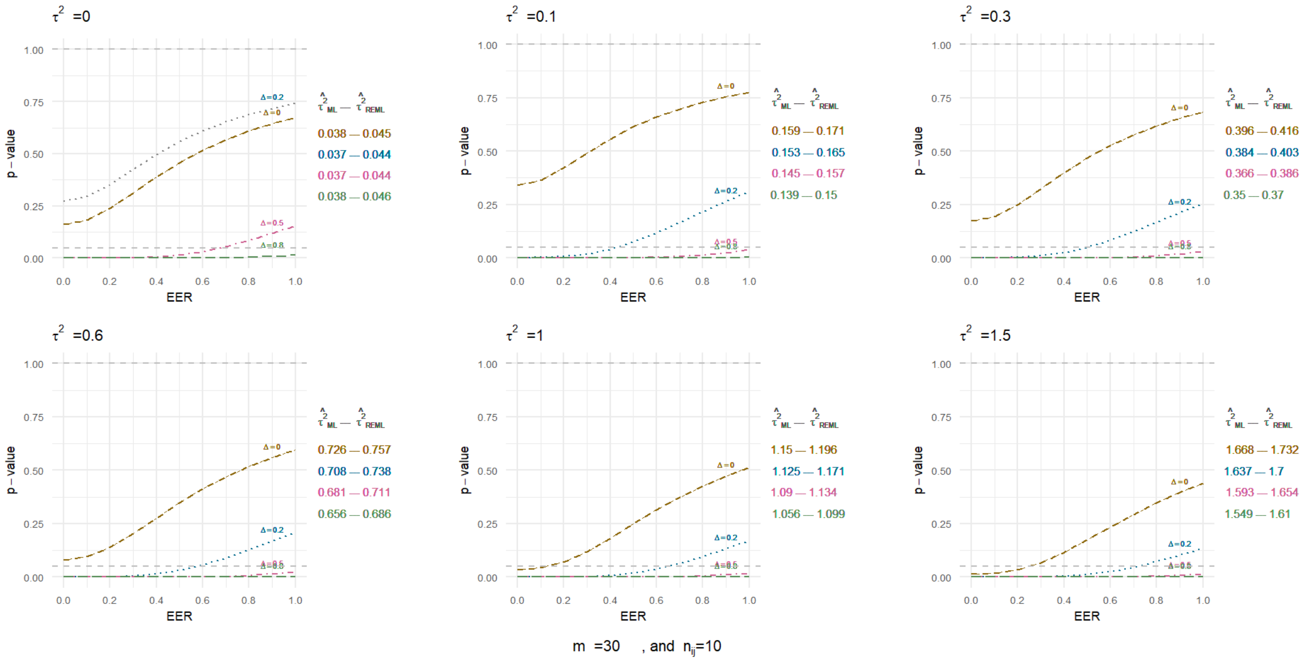


Figure C.3: *Meta-analysis by t-distribution: Sensitive analysis with various value of τ and Δ under small n_{ij}, where $n_{1j} = n_{2j}$ and large m. For further details see Section C.1*

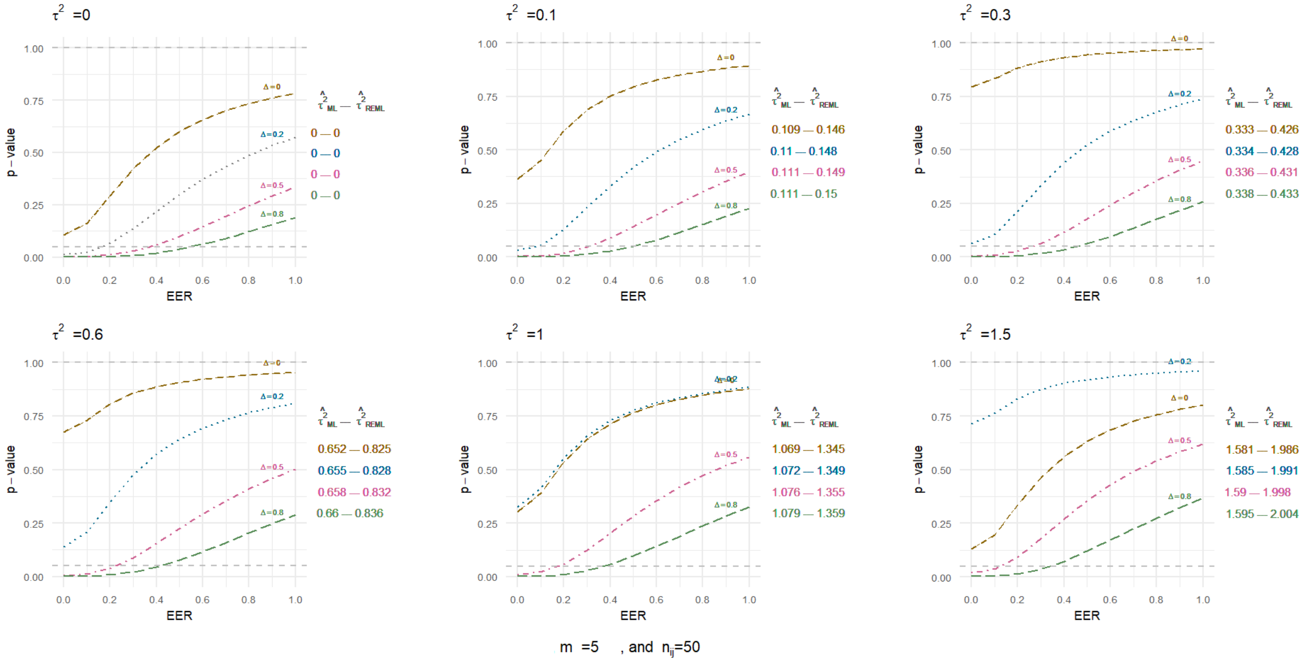


Figure C.4: *Meta-analysis by t-distribution: Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and small m. For further details see Section C.1*

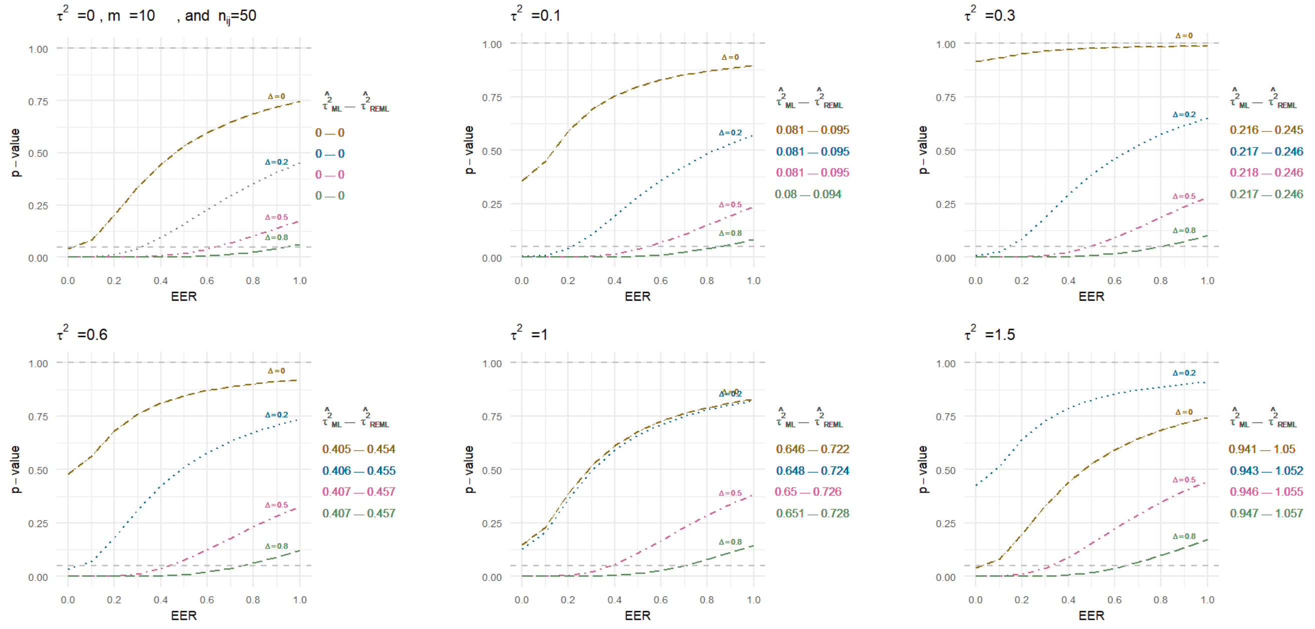


Figure C.5: *Meta-analysis by t-distribution: Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and medium m. For further details see Section C.1*

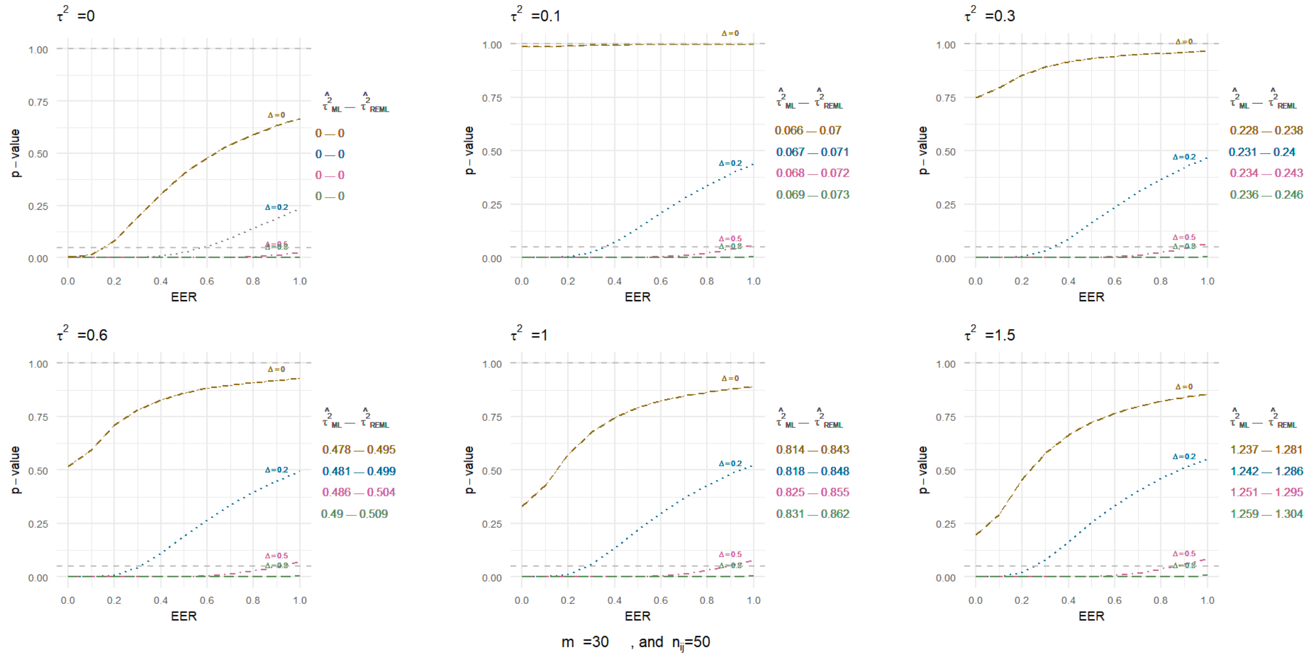


Figure C.6: *Meta-analysis (t-distribution): Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and large m. For further details see Section C.1*

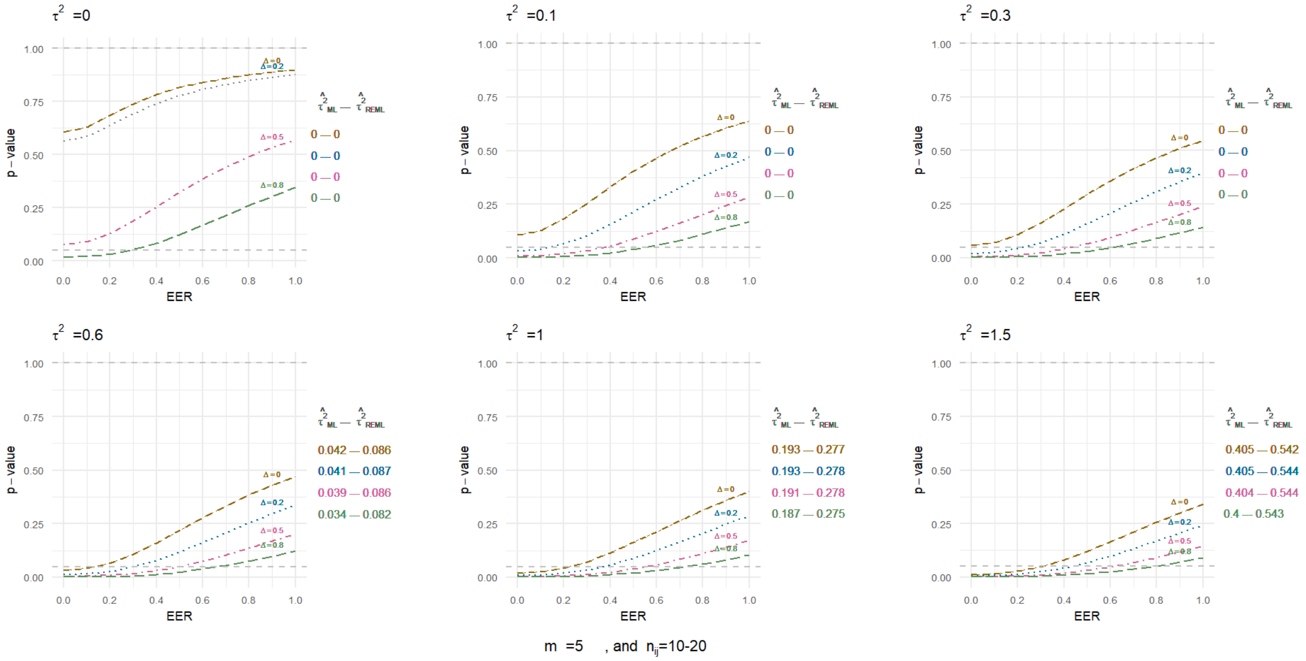


Figure C.7: *Meta-analysis by t-distribution: Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and small m. For further details see Section C.1*

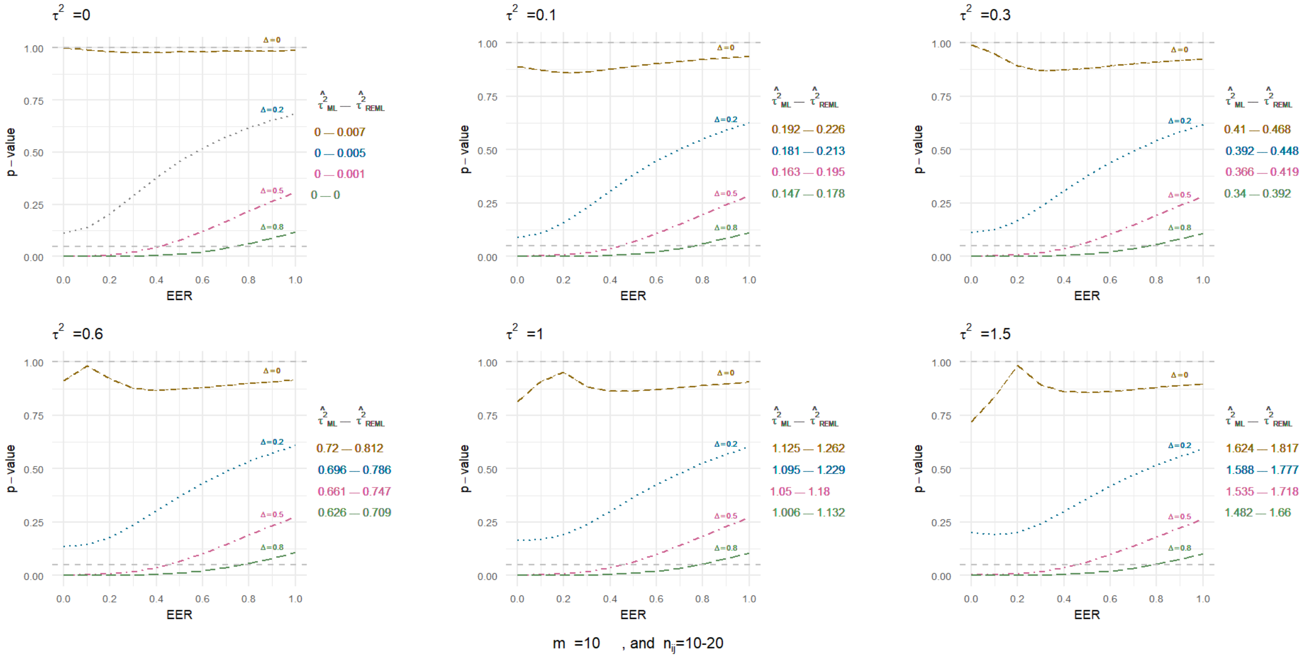


Figure C.8: *Meta-analysis by t-distribution: Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and medium m. For further details see Section C.1*

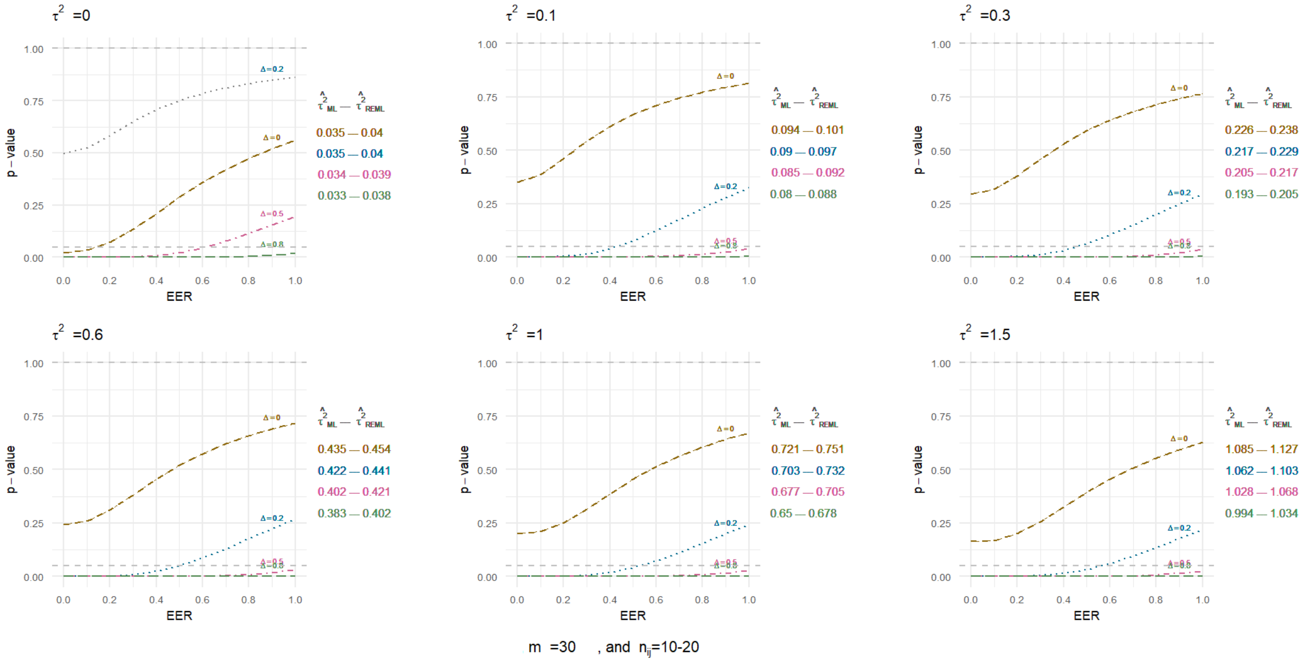


Figure C.9: *Meta-analysis by t-distribution: Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and large m. For further details see Section C.1*

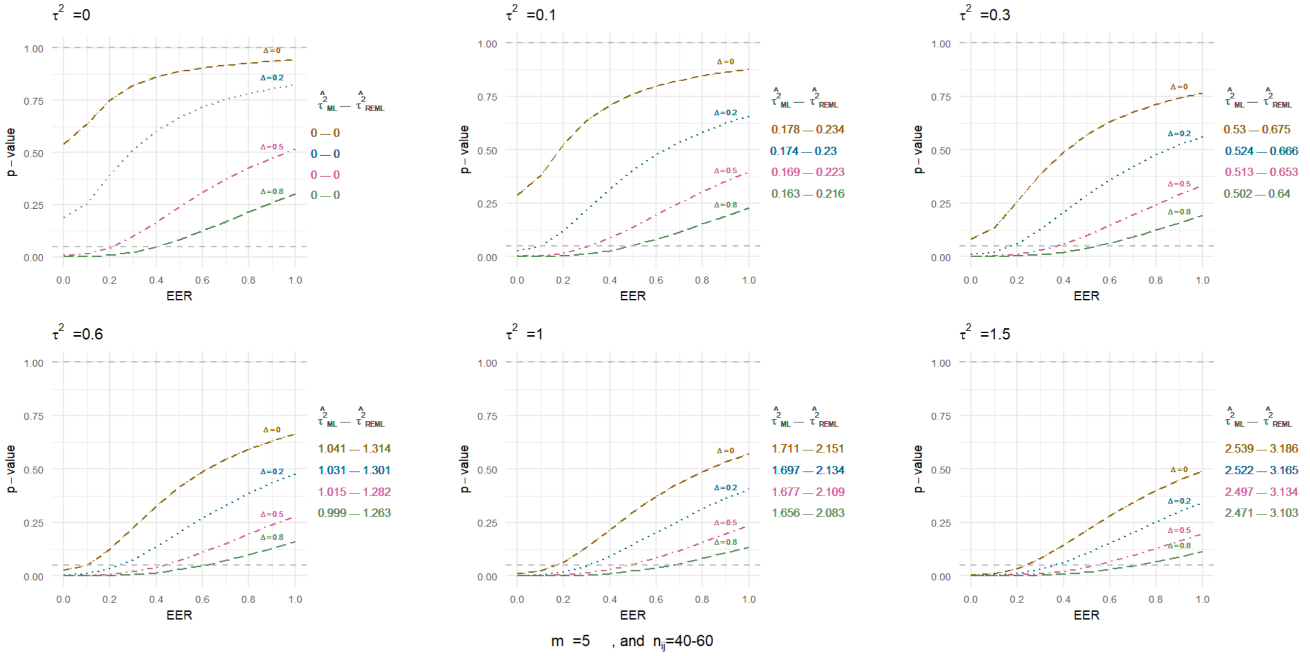


Figure C.10: *Meta-analysis (t-distribution): Sensitive analysis with various value of τ and Δ under large n_{ij}, where n_{1j} not necessary equal to n_{2j} and small m*

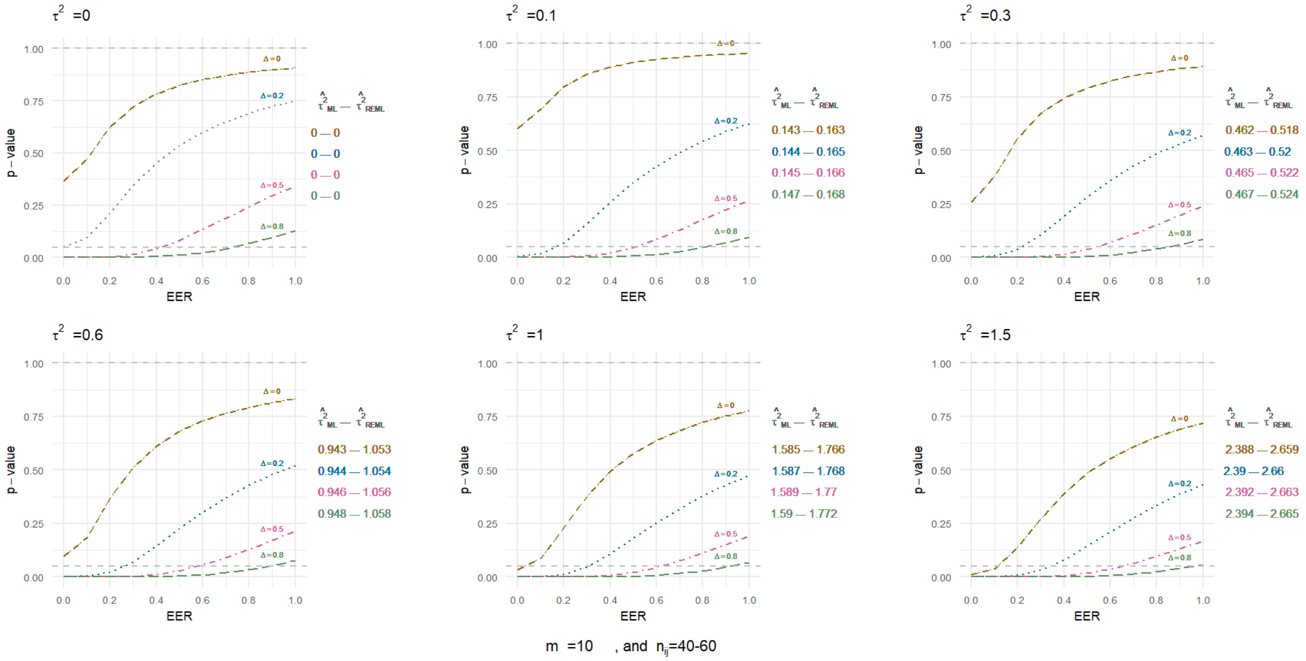


Figure C.11: *Meta-analysis by t-distribution: Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and medium m. For further details see Section* C.1

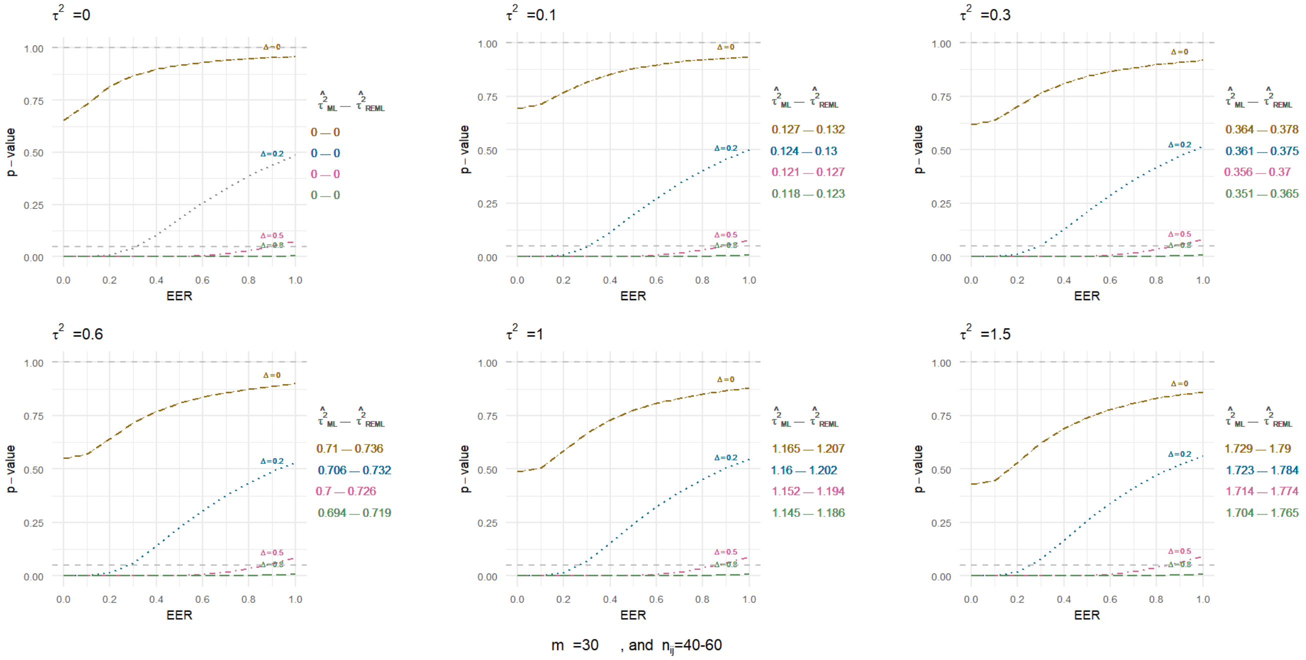


Figure C.12: *Meta-analysis by t-distribution: Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and large m. For further details see Section* C.1

b) Meta-analysis with Hartung-Knapp t-distribution vs t distribution

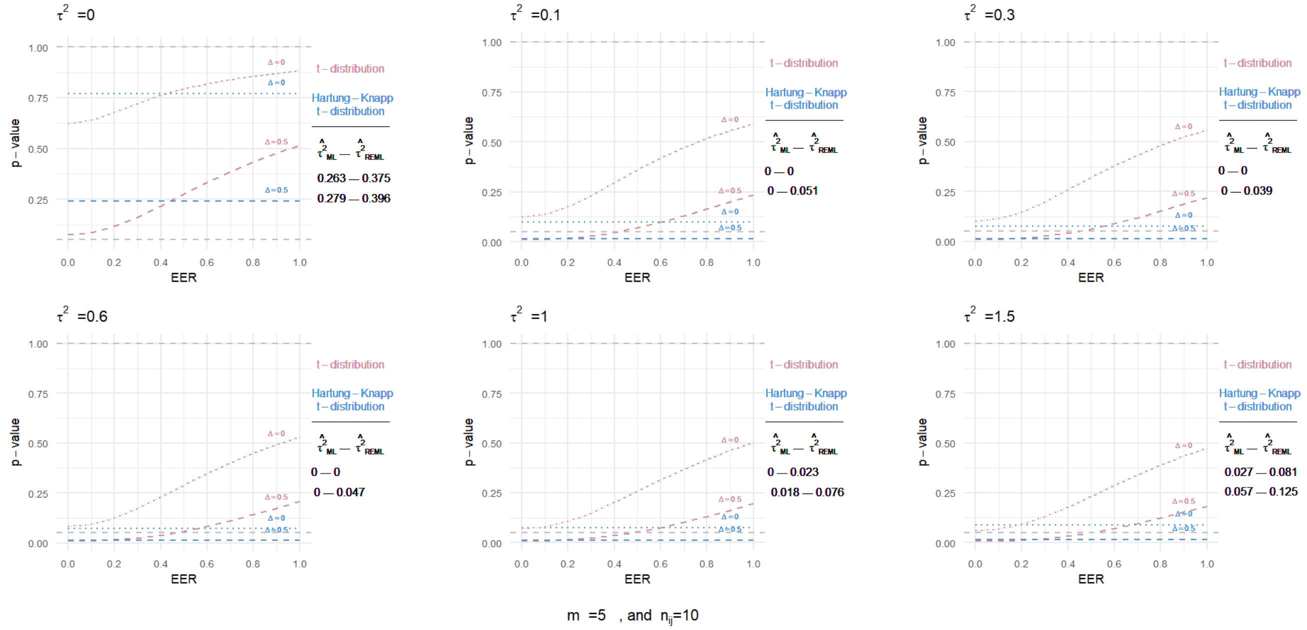


Figure C.13: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* $n_{1j} = n_{2j}$ *and small m. For further details see Section C.1*

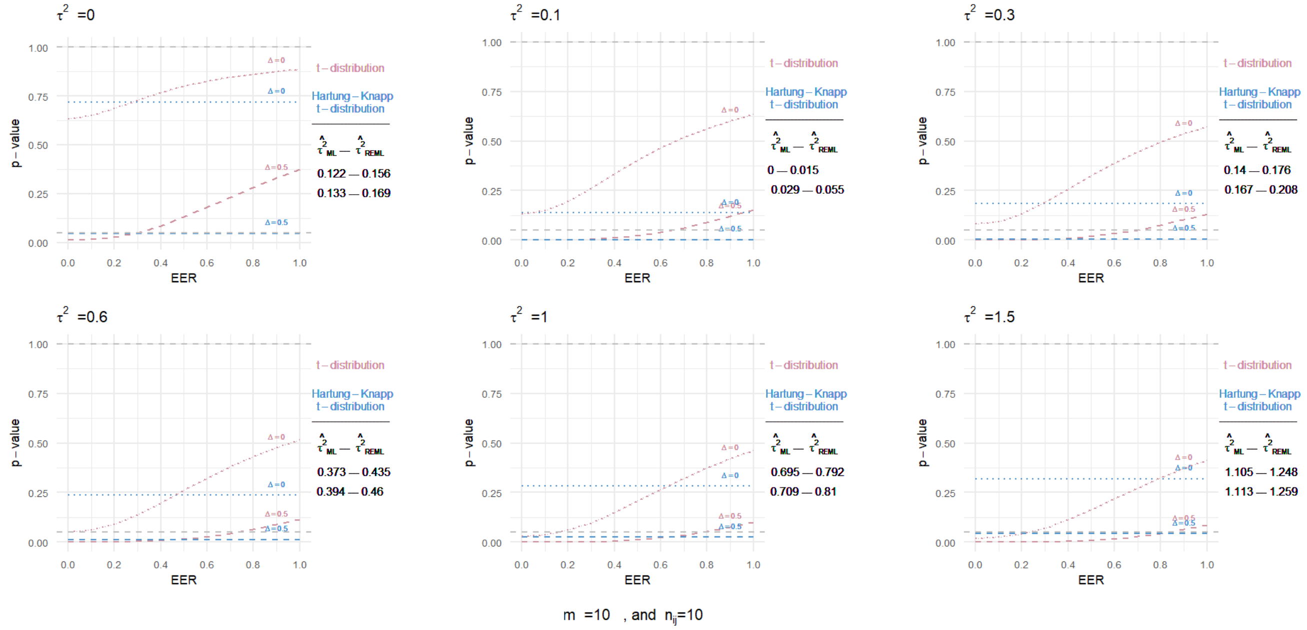


Figure C.14: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* $n_{1j} = n_{2j}$ *and medium m. For further details see Section C.1*

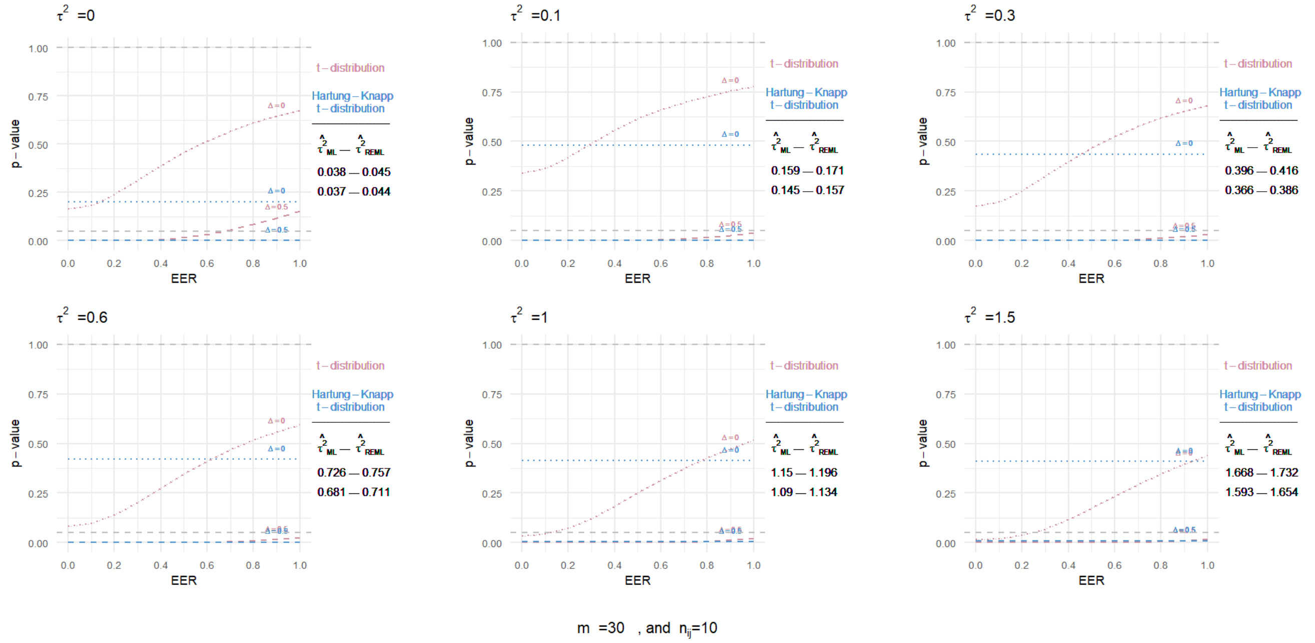


Figure C.15: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under small n_{ij}, where $n_{1j} = n_{2j}$ and large m. For further details see Section C.1*

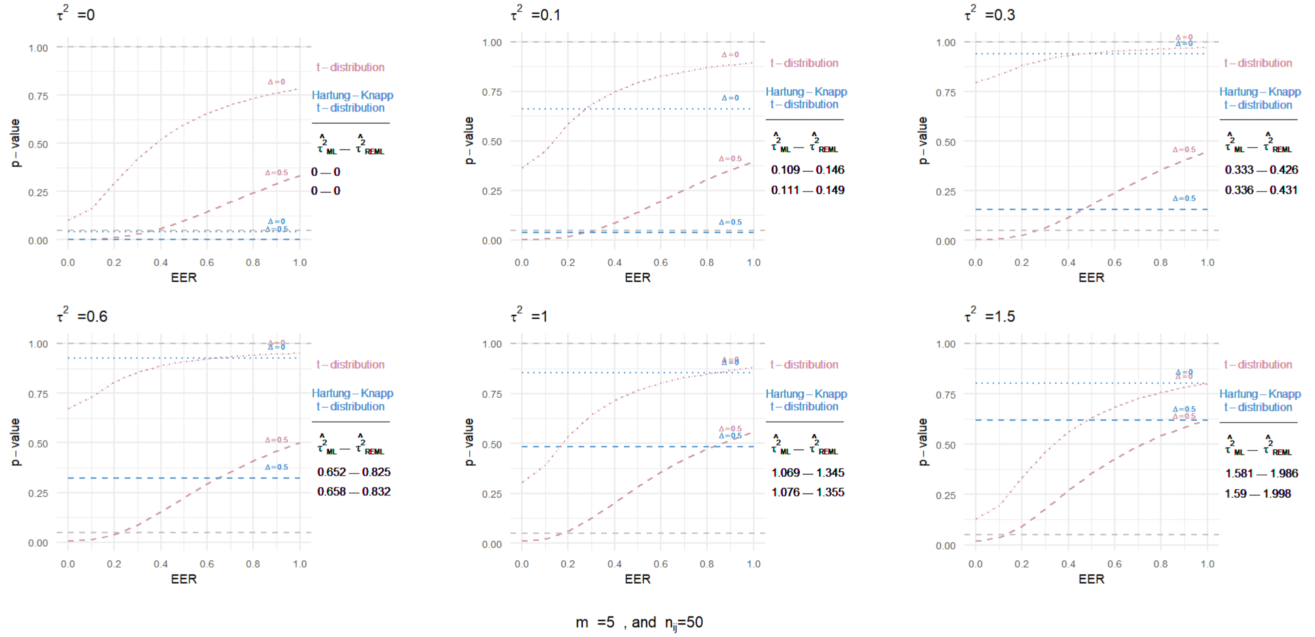


Figure C.16: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and small m. For further details see Section C.1*

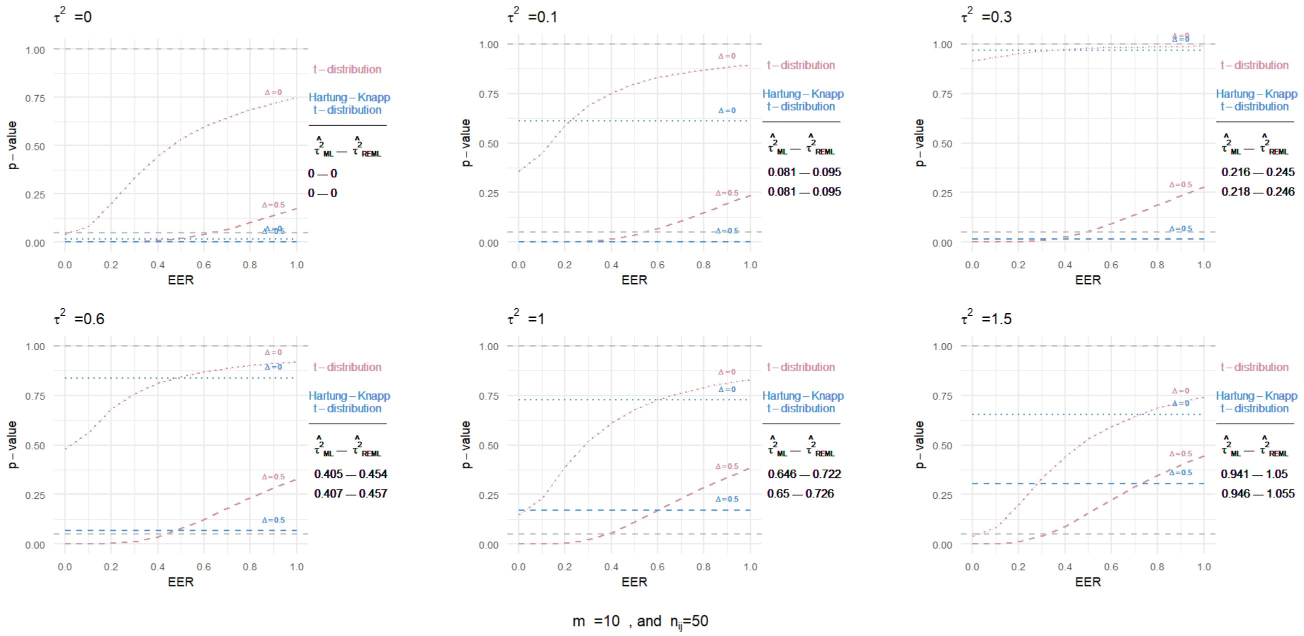


Figure C.17: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and medium m. For further details see Section C.1*

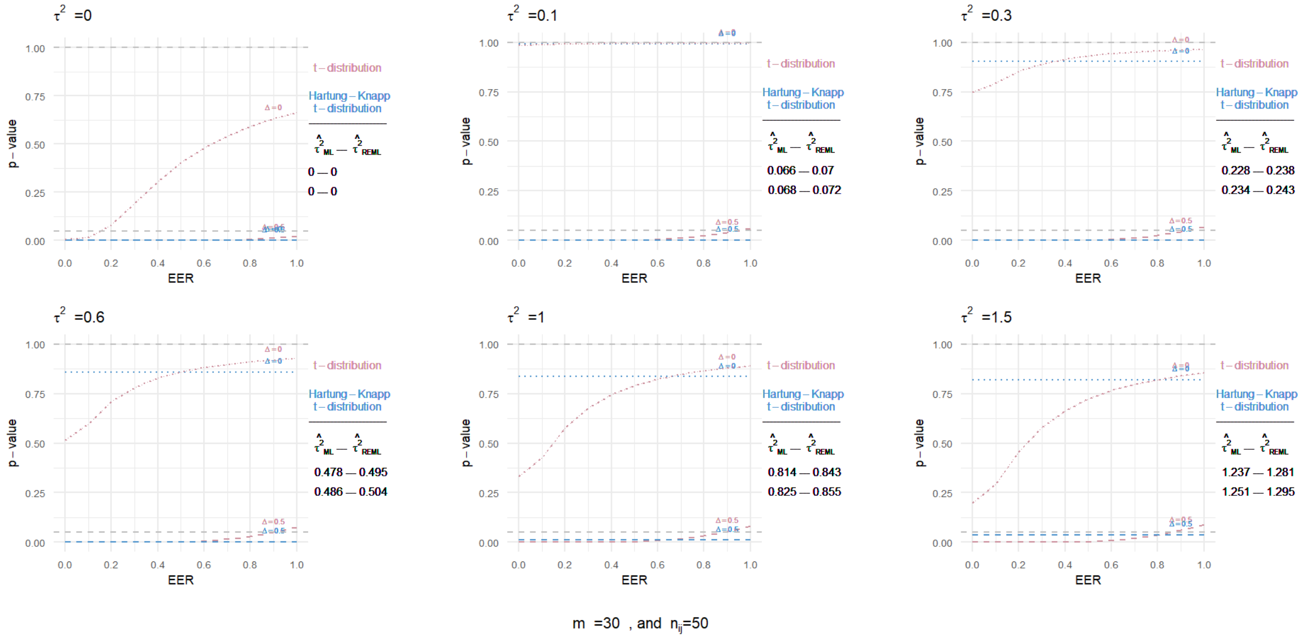


Figure C.18: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and large m. For further details see Section C.1*

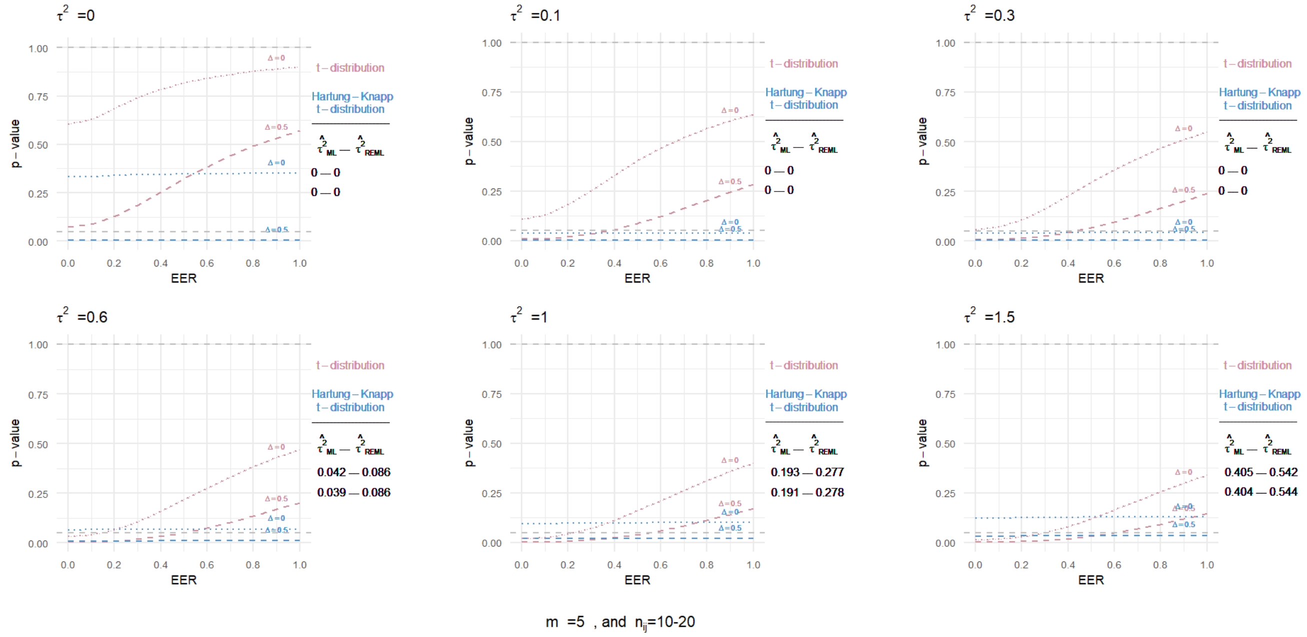


Figure C.19: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and small m. For further details see Section C.1*

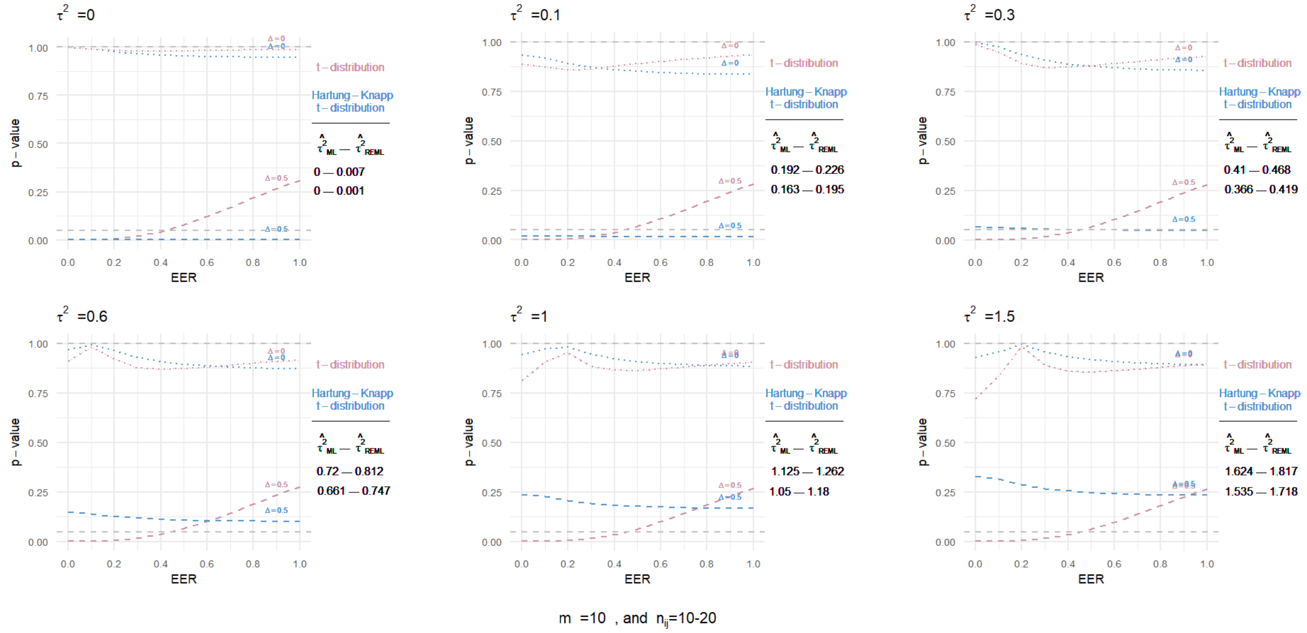


Figure C.20: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and medium m.For further details see Section C.1*

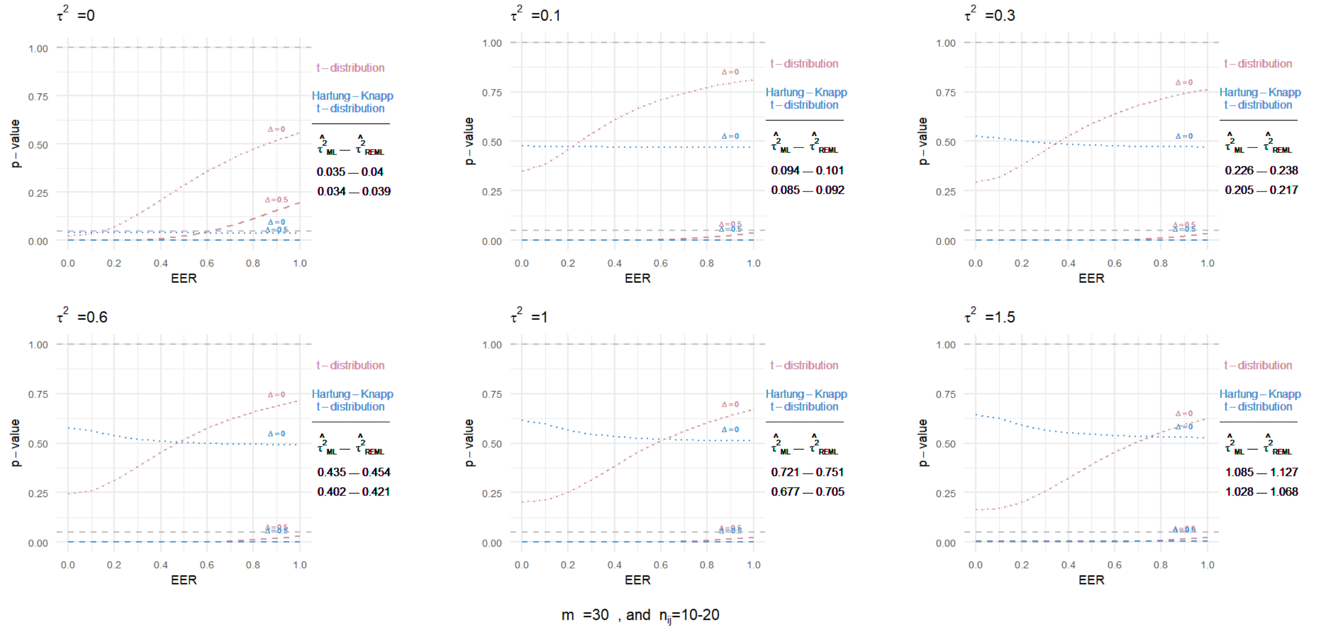


Figure C.21: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and large m. For further details see Section C.1*

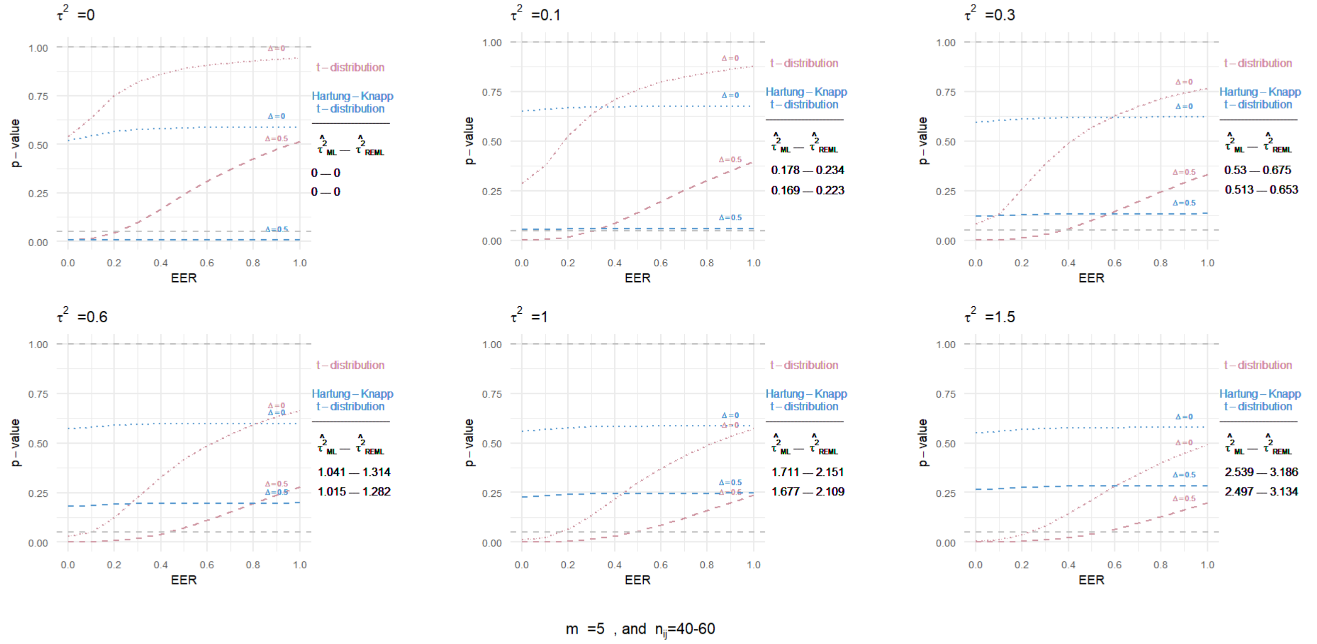


Figure C.22: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under large n_{ij}, where n_{1j} not necessary equal to n_{2j} and small m. For further details see Section C.1*

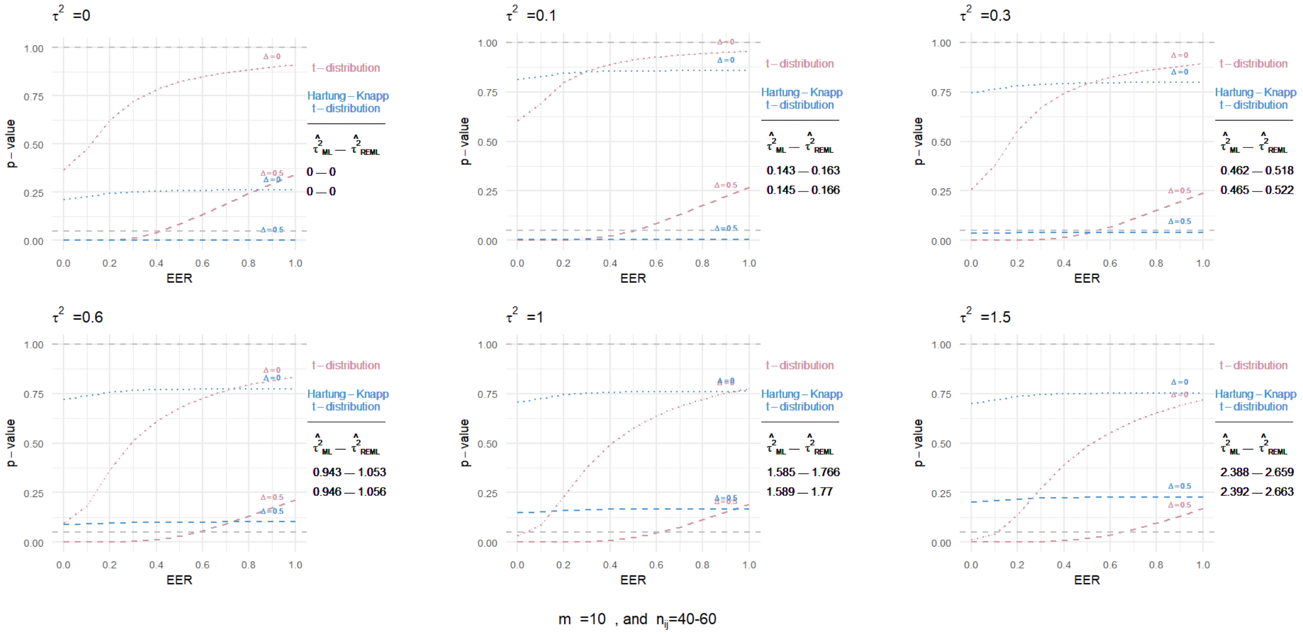


Figure C.23: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under large n_{ij}, where n_{1j} not necessary equal to n_{2j} and medium m.For further details see Section C.1*

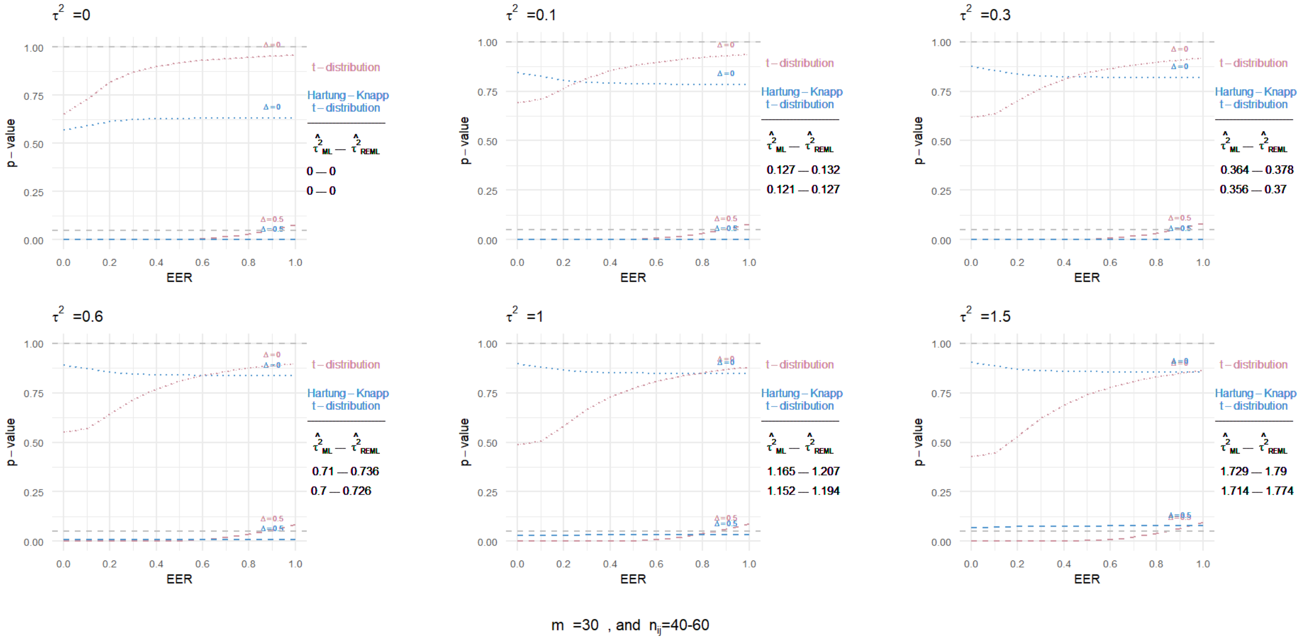


Figure C.24: *Meta-analysis (Hartung-Knapp t-distribution vs t distribution): Sensitive analysis with various value of τ and Δ under large n_{ij}, where n_{1j} not necessary equal to n_{2j} and large m. For further details see Section C.1*

C.2 Mixed Model

The measure of between-study variance under mixed model is described as the standard deviations of environment-by-treatment interaction divided by the experimental.

$$\hat{\tau}^2 = 2\,\frac{\hat{\sigma}_\zeta^2}{\hat{\sigma}_e^2}$$

The estimation of the quantity $\sigma_\zeta^2/\sigma_e^2$ can be found in Section 2.4. Simulations in this section is followed the same strategies as in Section C.1. However, here REML and ML are estimated by *lme*4 function. In the figures below, τ indicates heterogeneity variance, Δ indicates overall effect size, n_{ij} indicates sample size for each group (n_{1j} for treatment 1 and n_{2j} for control) and m indicates small size. To making inferences about $\mu_1 - \mu_2$, the t-statistic by both Satterthwaite approximation and N-m-1 degree of freedom are used.

a) Mixed model with Satterthwaite approximation

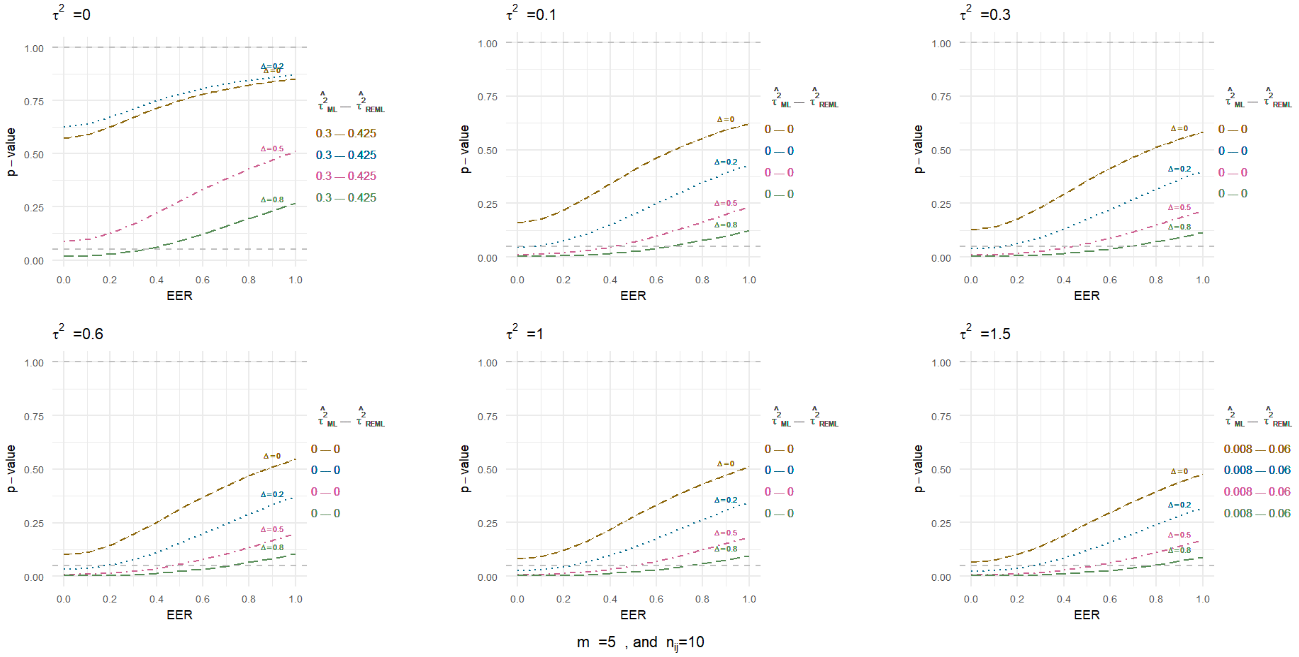


Figure C.25: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* $n_{1j} = n_{2j}$ *and small m*

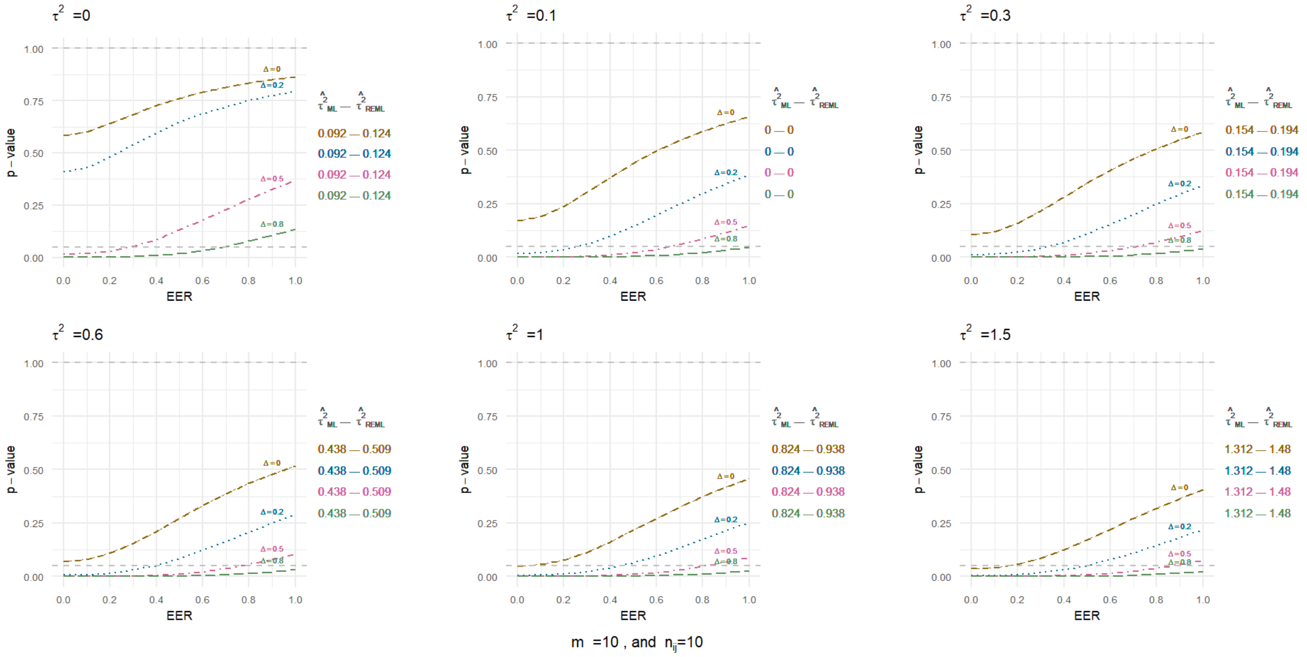


Figure C.26: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* $n_{1j} = n_{2j}$ *and medium* m

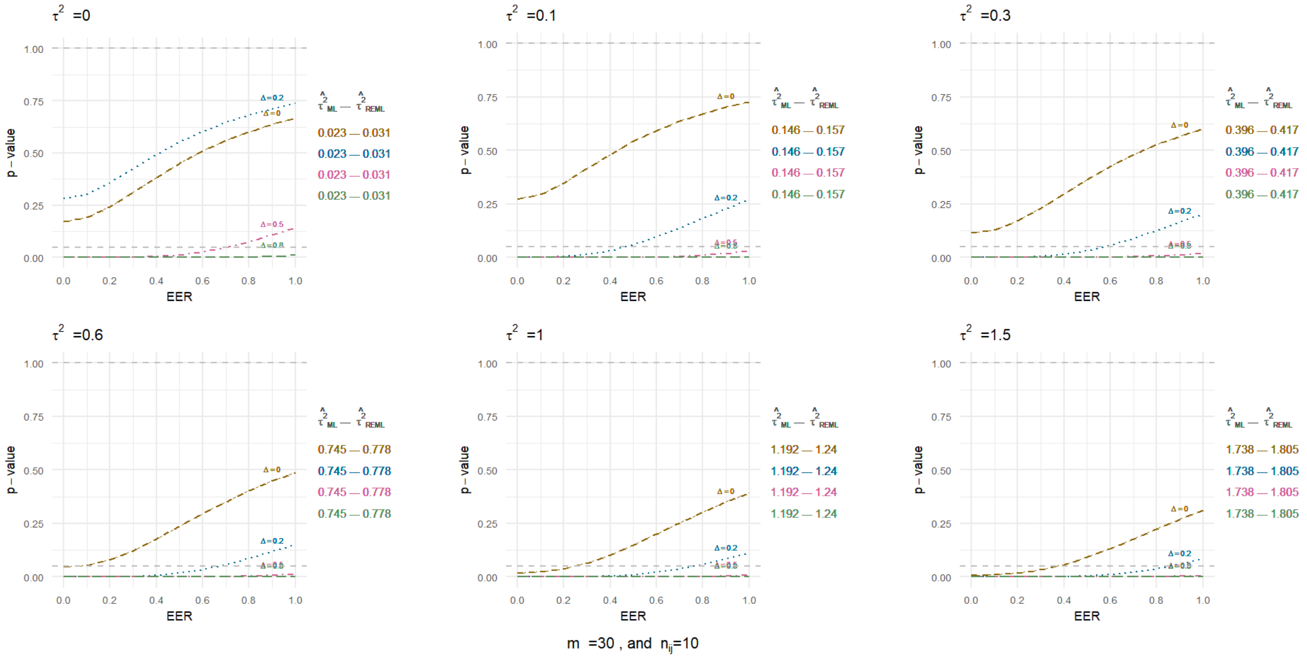


Figure C.27: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* $n_{1j} = n_{2j}$ *and large* m

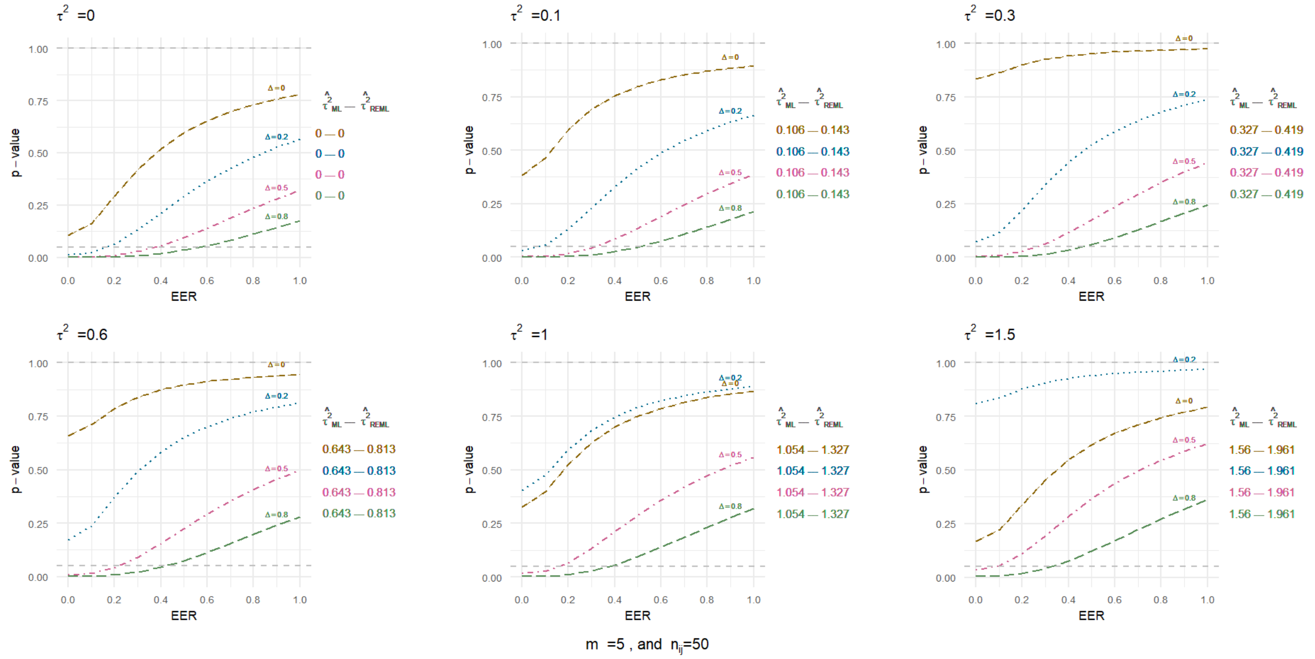


Figure C.28: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and small m*

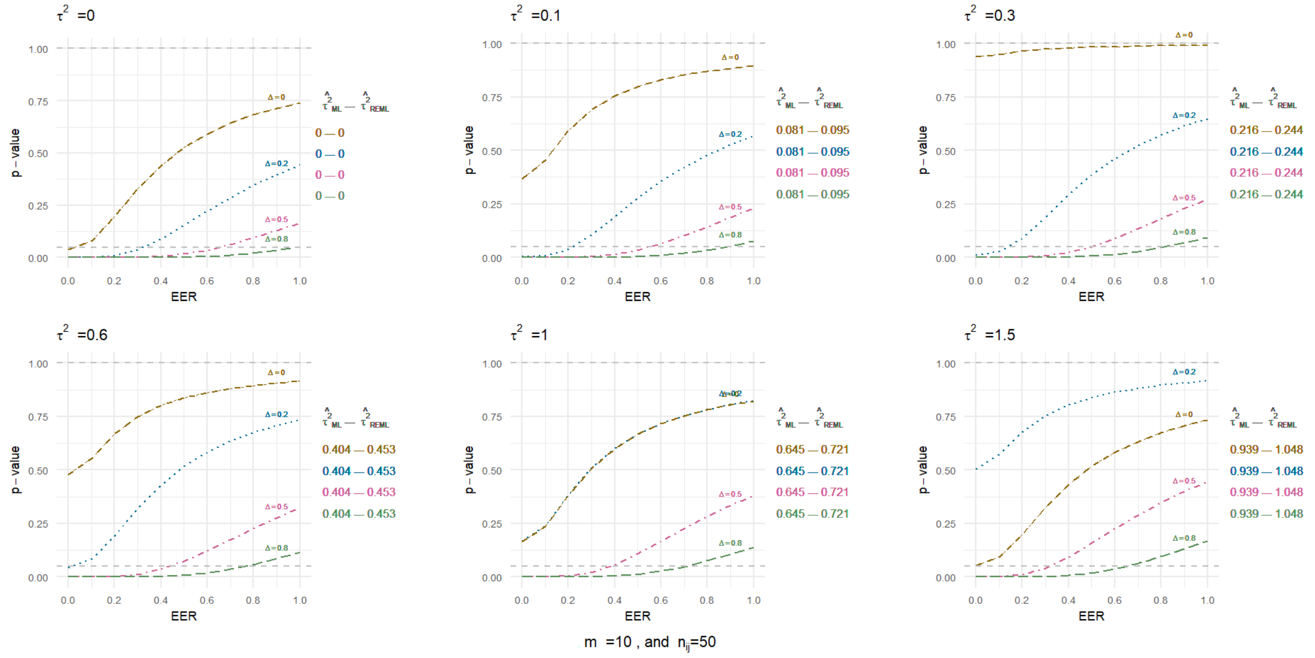


Figure C.29: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and medium m*

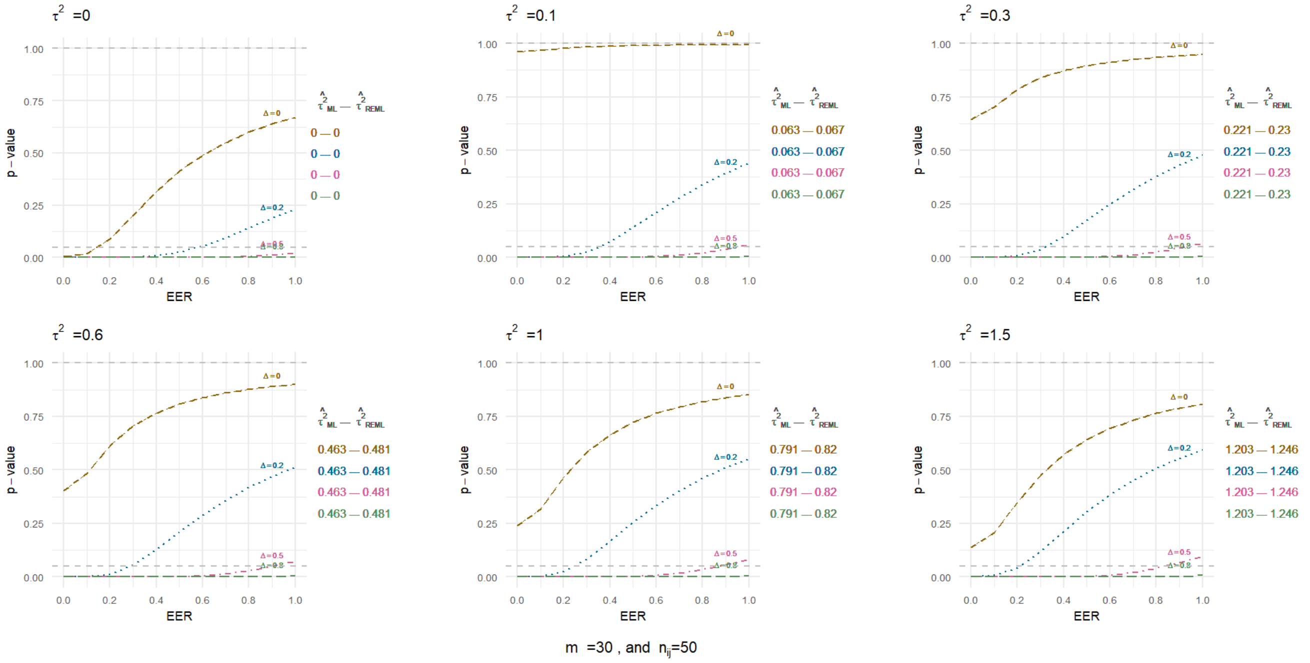


Figure C.30: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* $n_{1j} = n_{2j}$ *and large* m

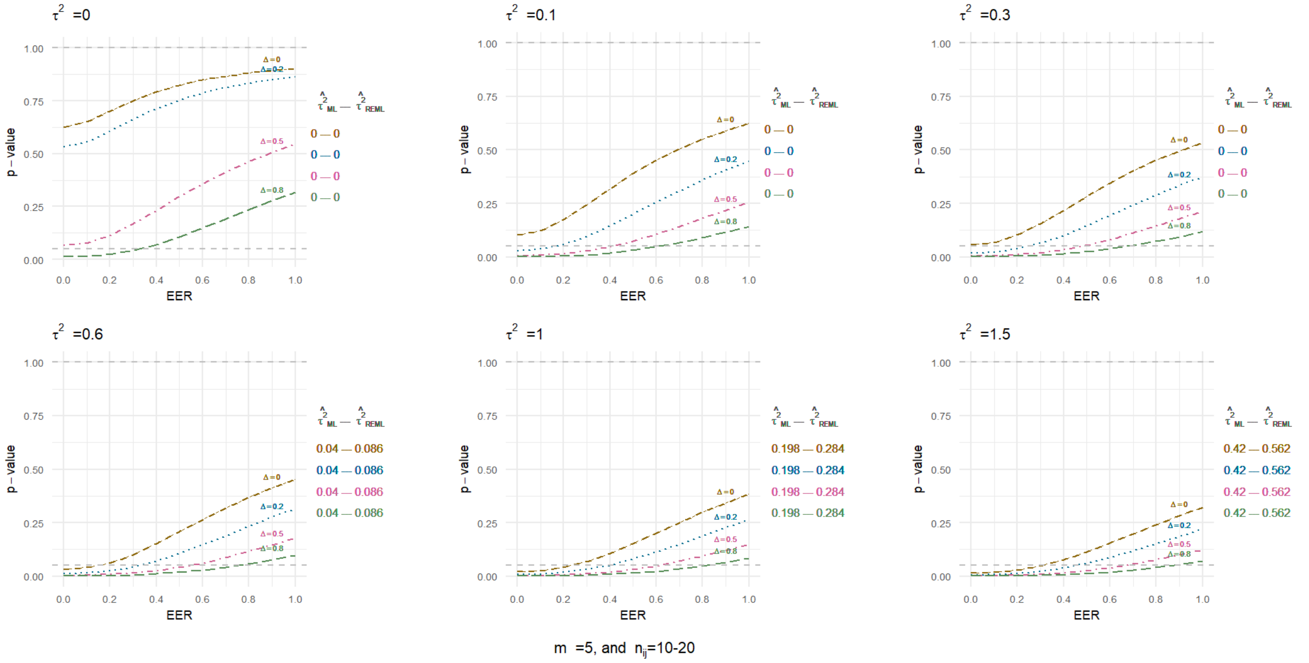


Figure C.31: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and small* m

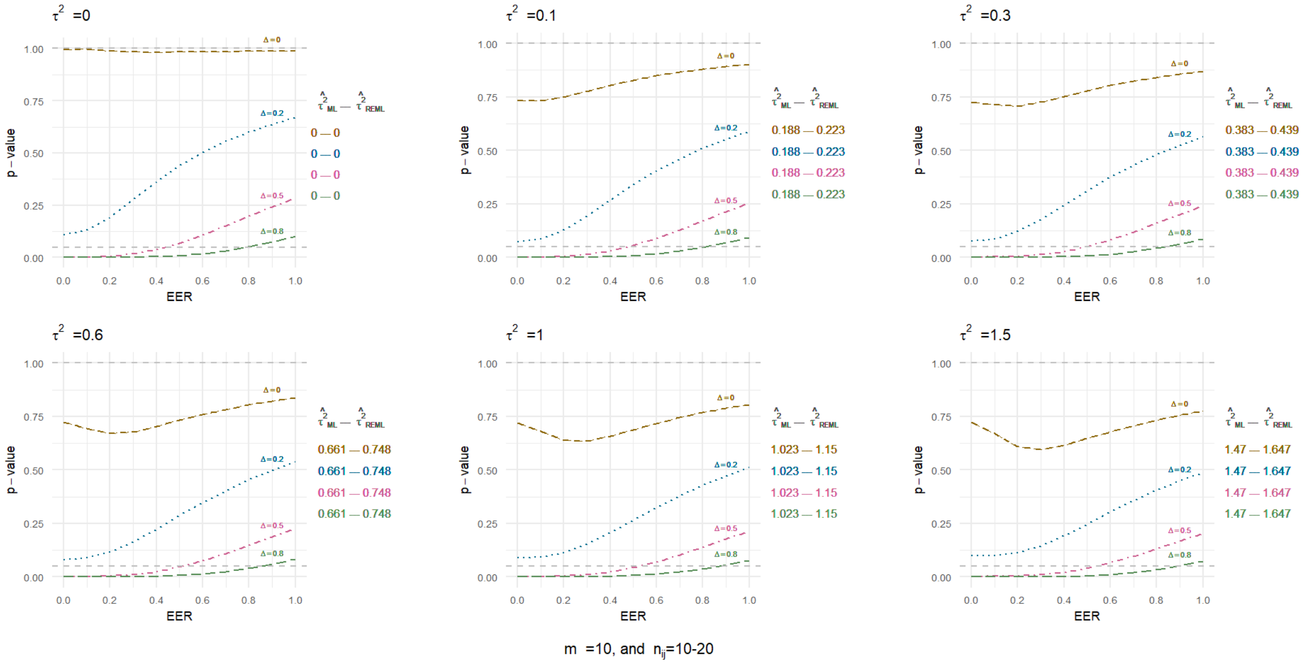


Figure C.32: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and medium m*

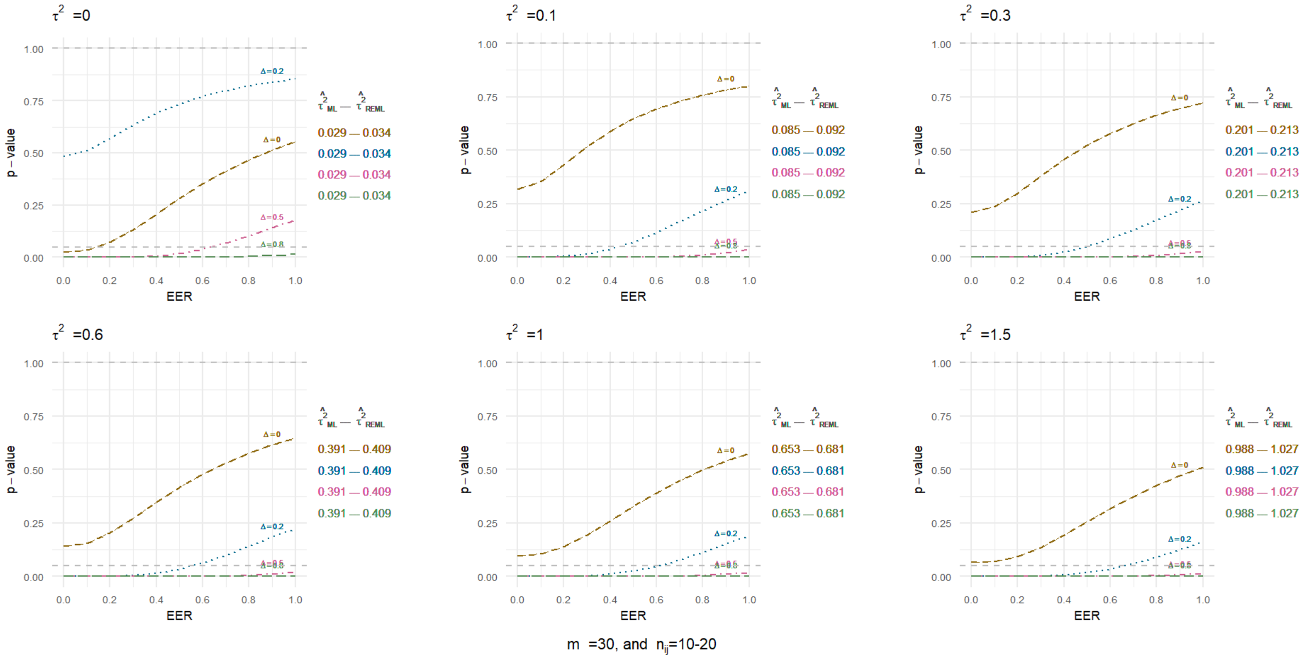


Figure C.33: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and large m*

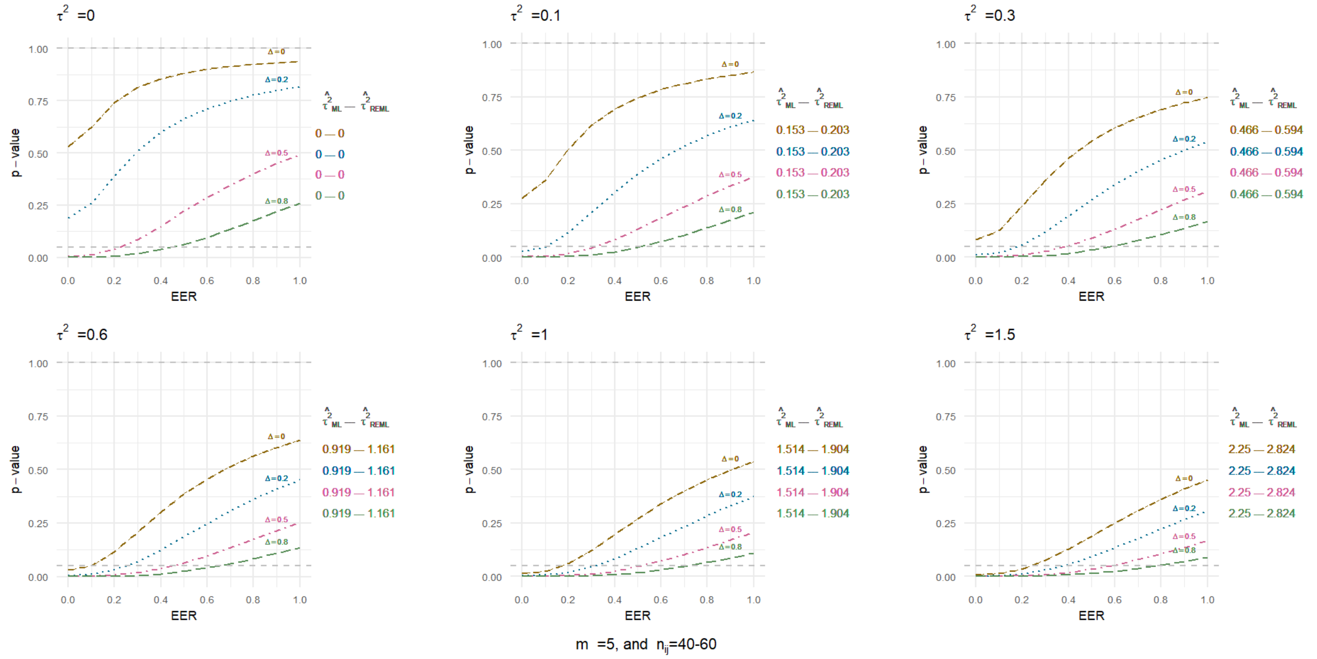


Figure C.34: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and small* m

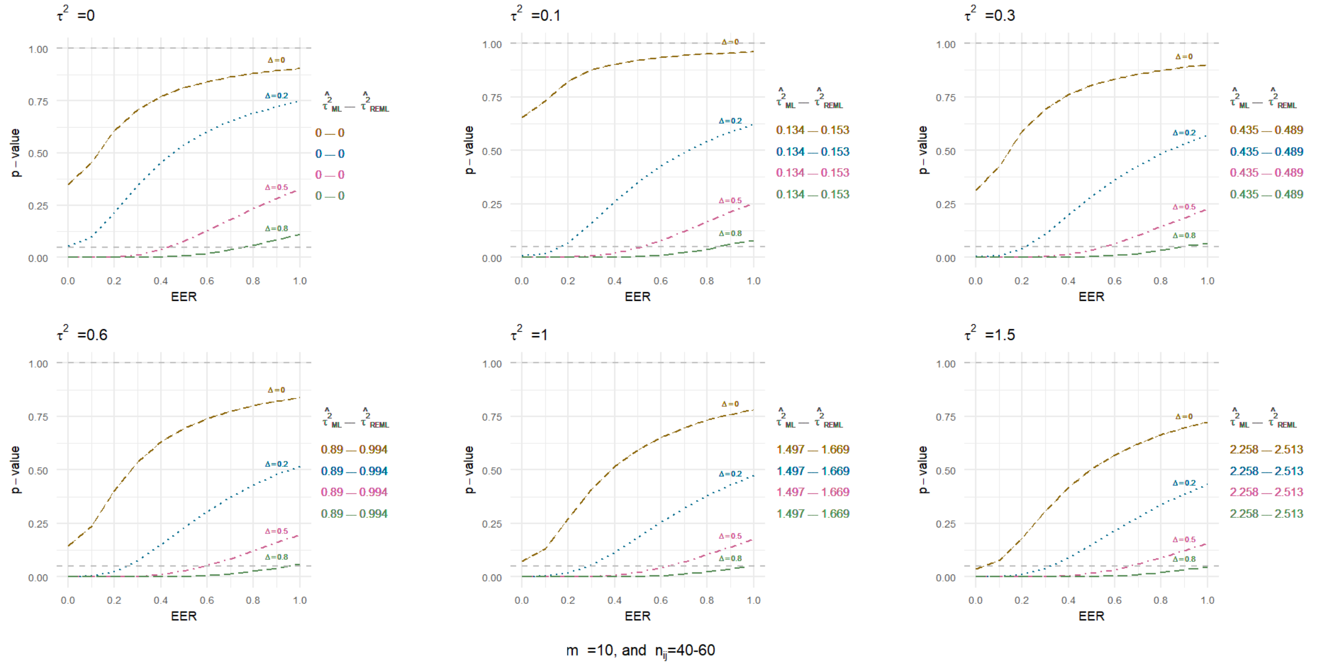


Figure C.35: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and medium* m

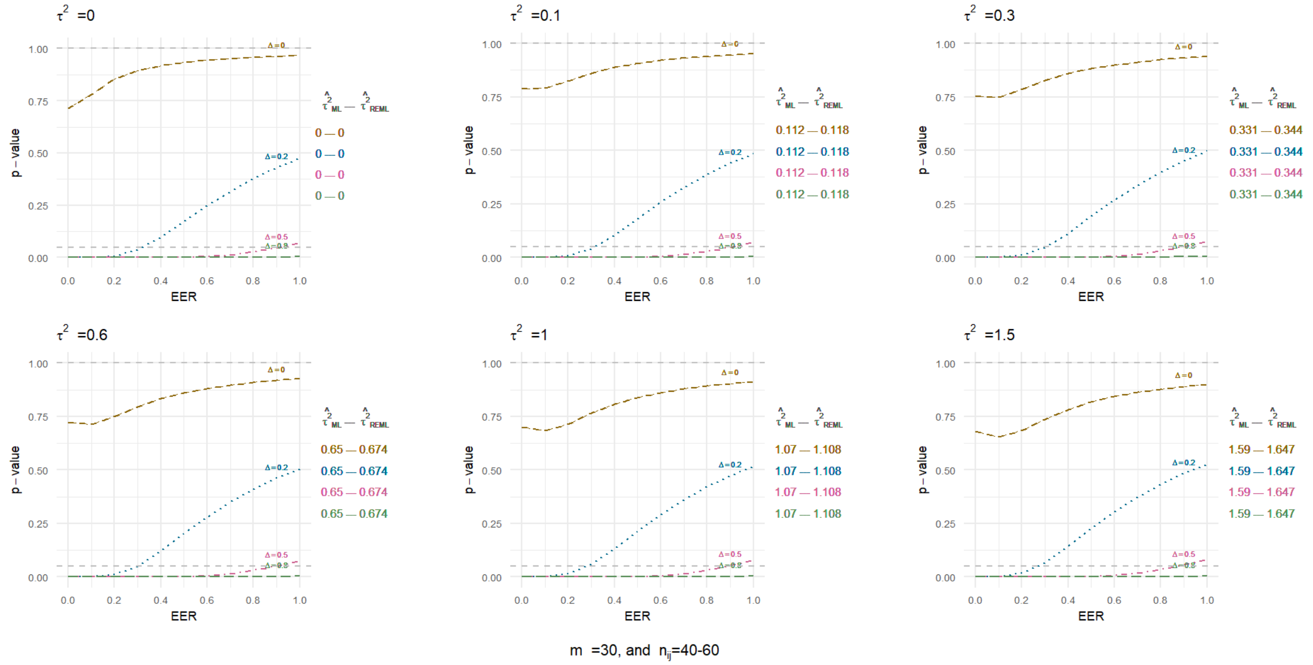


Figure C.36: *Mixed model (Satterthwaite approximation): Sensitive analysis with various value of τ and Δ under large n_{ij}, where n_{1j} not necessary equal to n_{2j} and large m*

b) Mixed model with N-m-1 Degree of Freedom

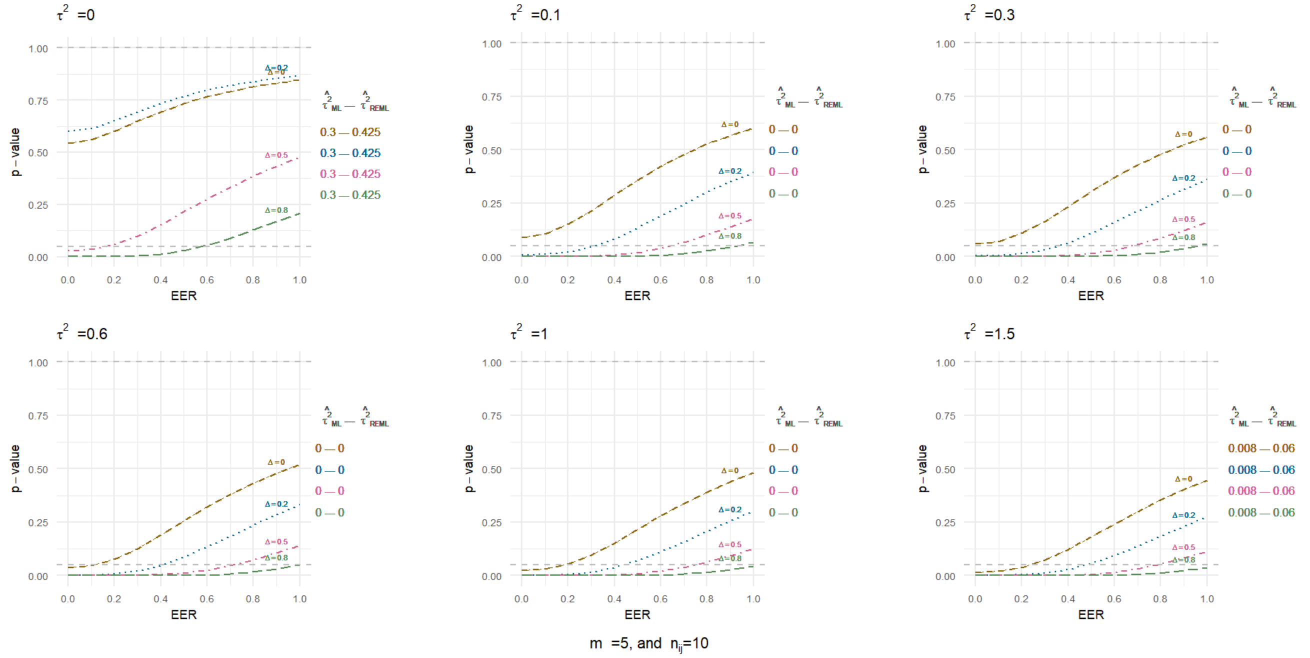


Figure C.37: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of τ and Δ under small n_{ij}, where $n_{1j} = n_{2j}$ and small m*

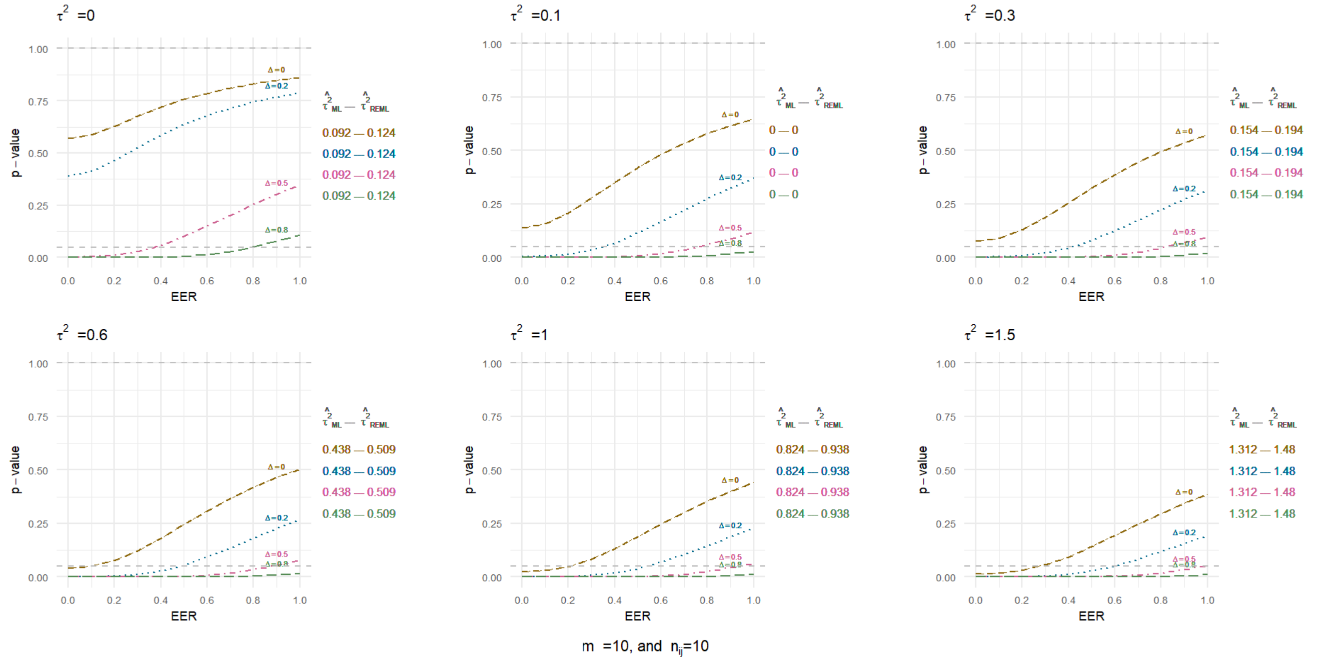


Figure C.38: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* $n_{1j} = n_{2j}$ *and medium m*

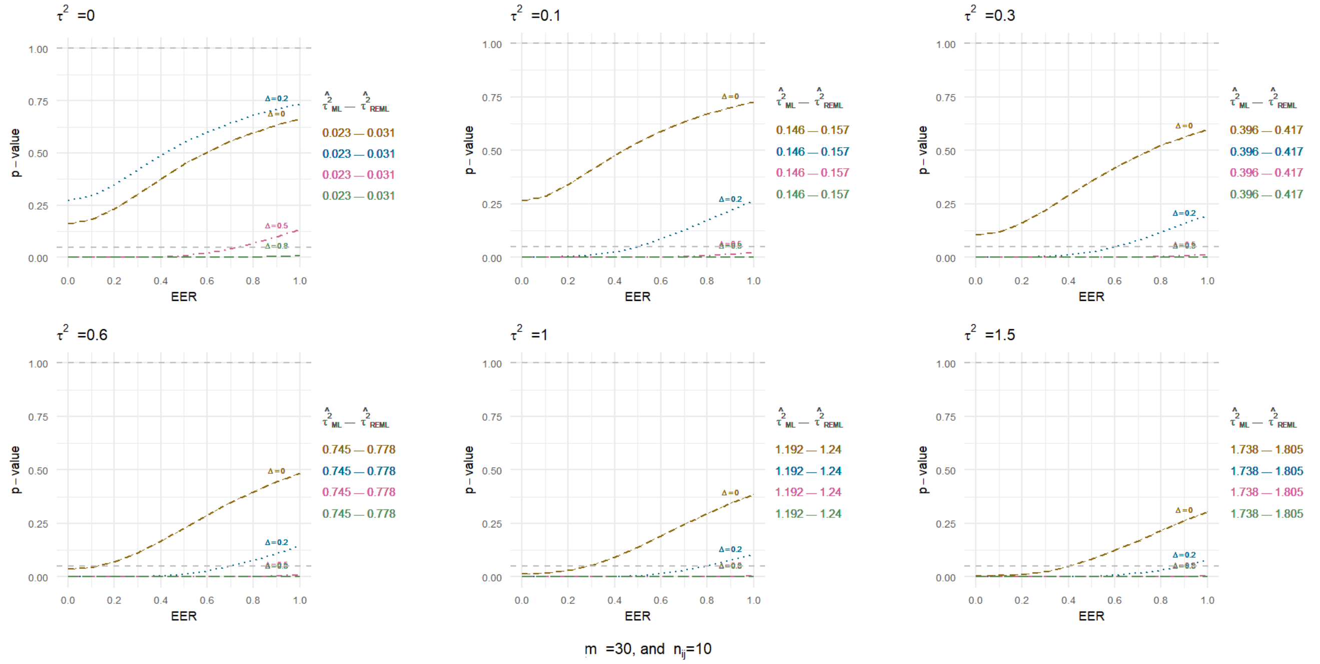


Figure C.39: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of* τ *and* Δ *under small* n_{ij}*, where* $n_{1j} = n_{2j}$ *and large m*

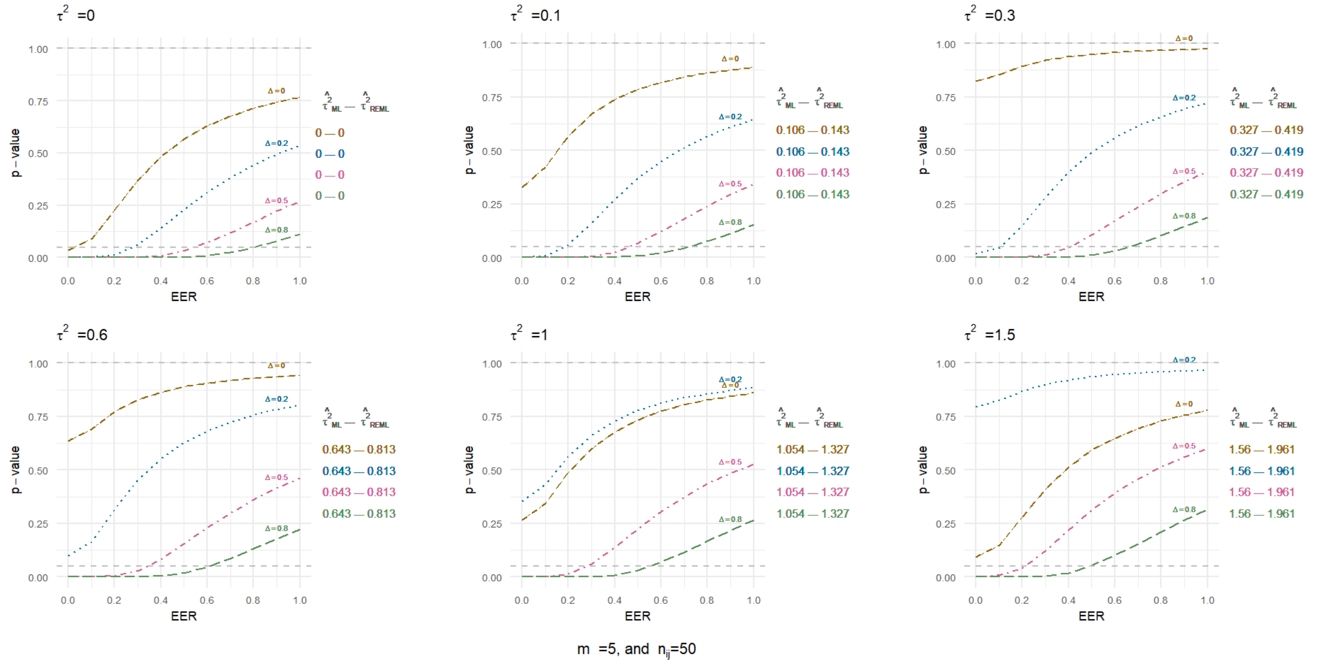


Figure C.40: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* $n_{1j} = n_{2j}$ *and small* m

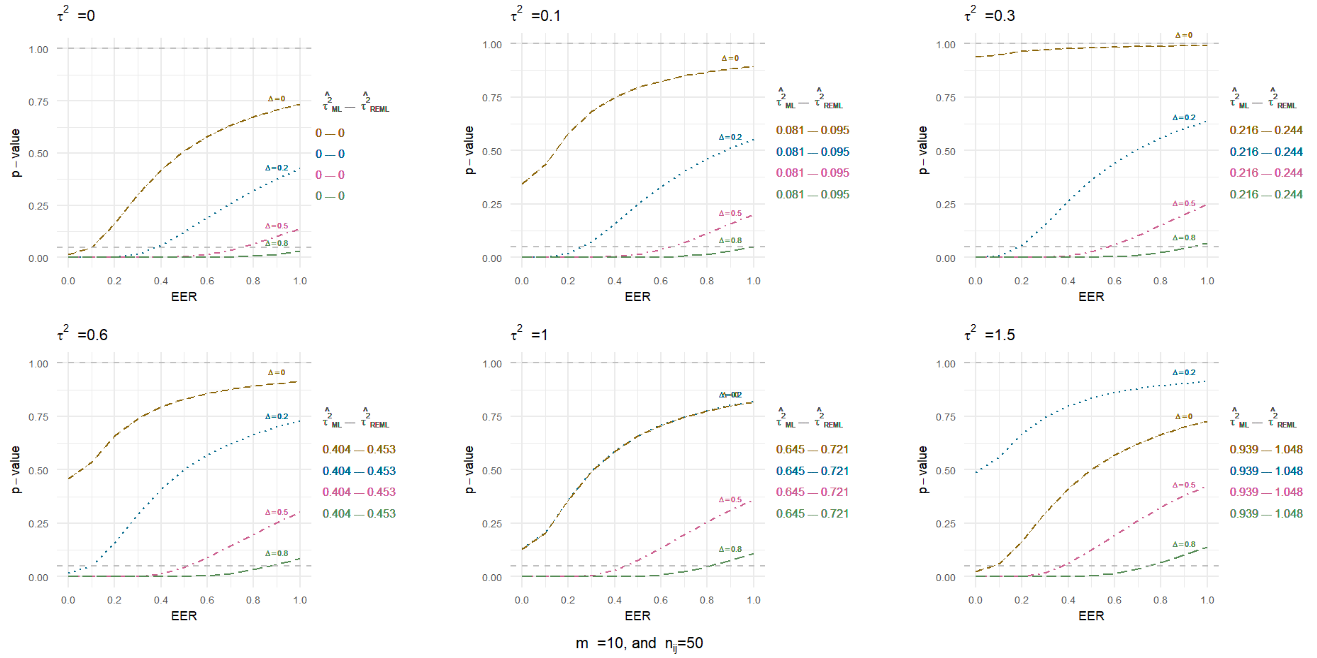


Figure C.41: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* $n_{1j} = n_{2j}$ *and medium* m

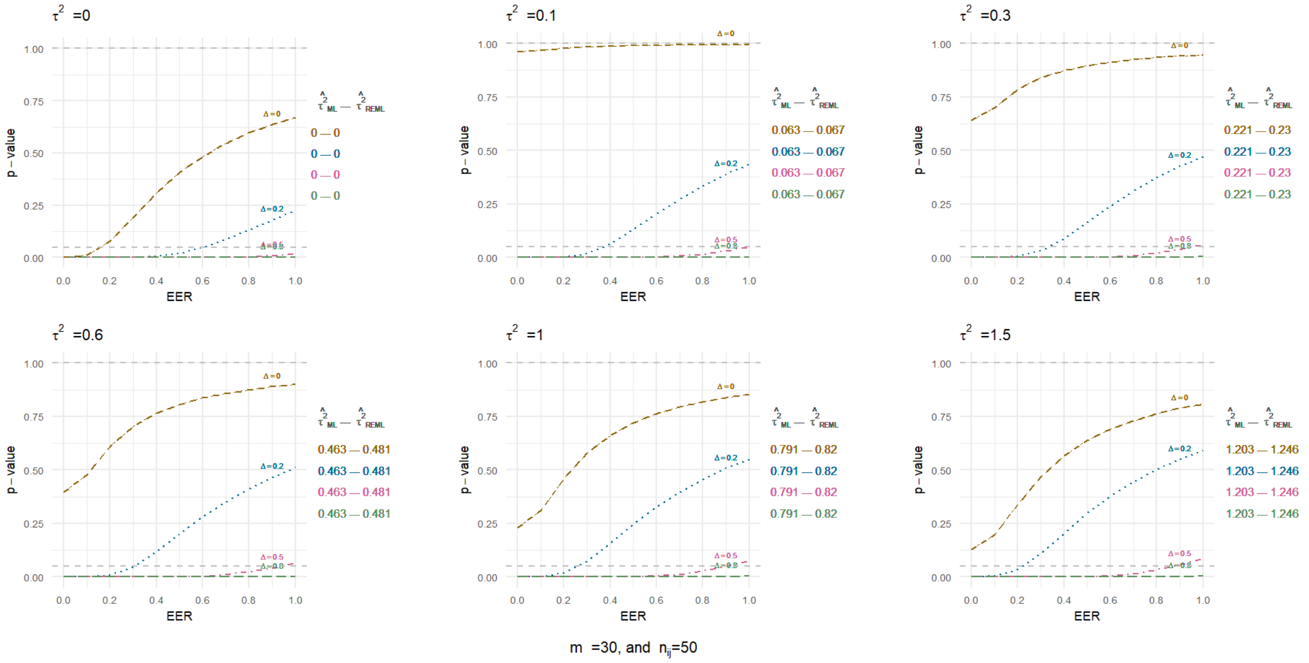


Figure C.42: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of τ and Δ under large n_{ij}, where $n_{1j} = n_{2j}$ and large m*

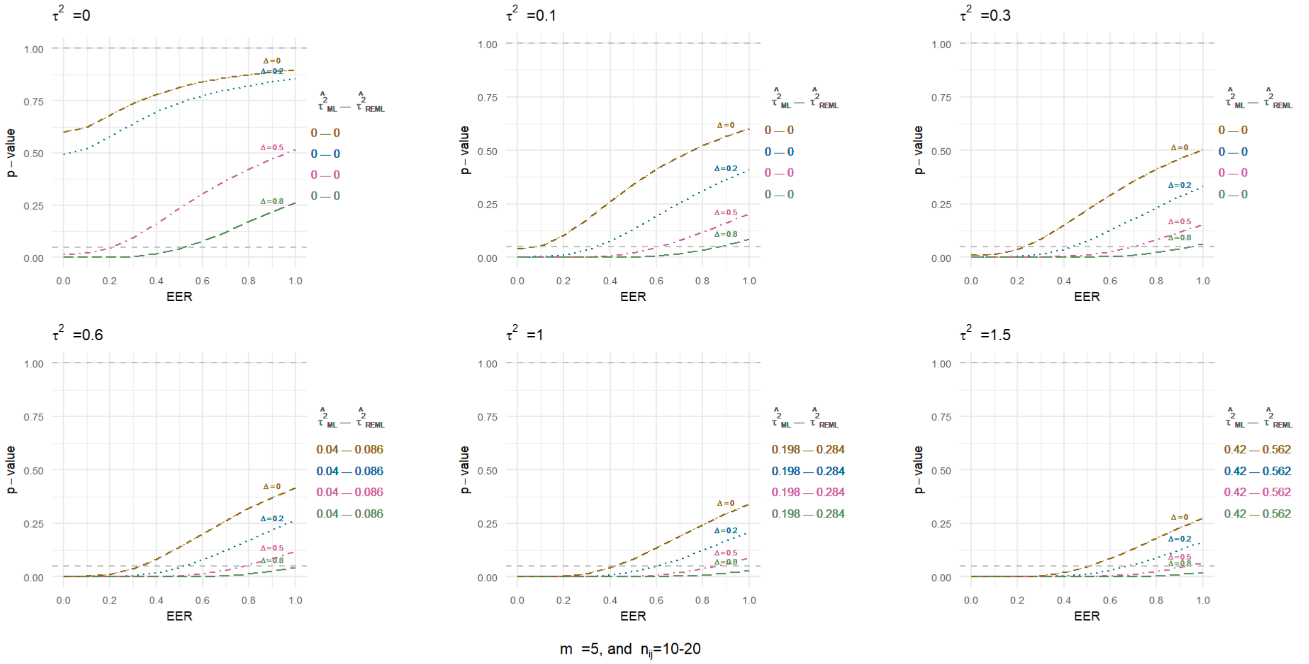


Figure C.43: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and small m*

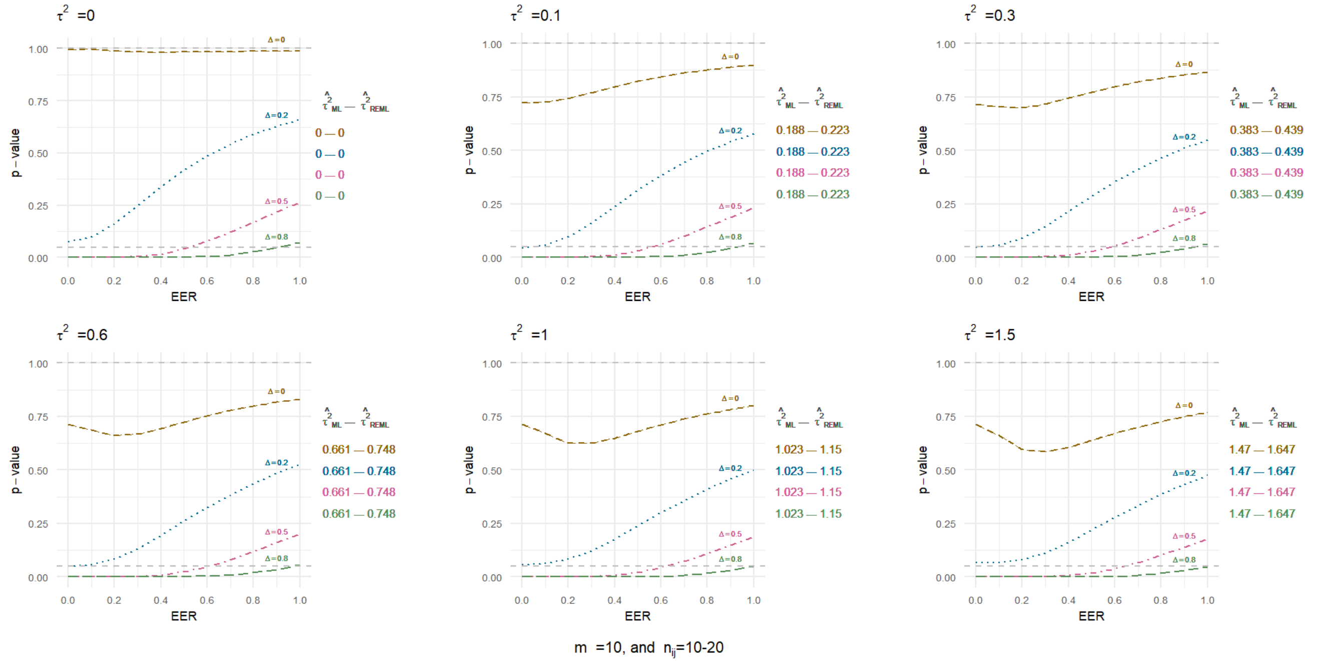


Figure C.44: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and medium m*

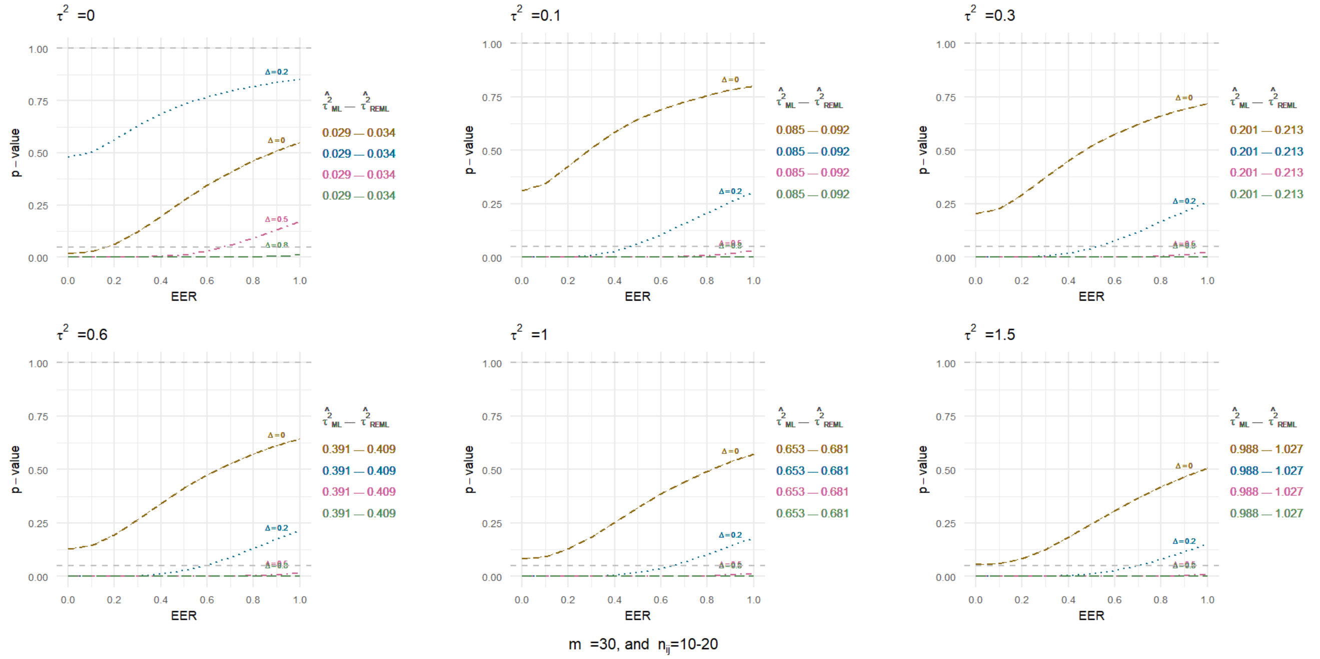


Figure C.45: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of τ and Δ under small n_{ij}, where n_{1j} not necessary equal to n_{2j} and large m*

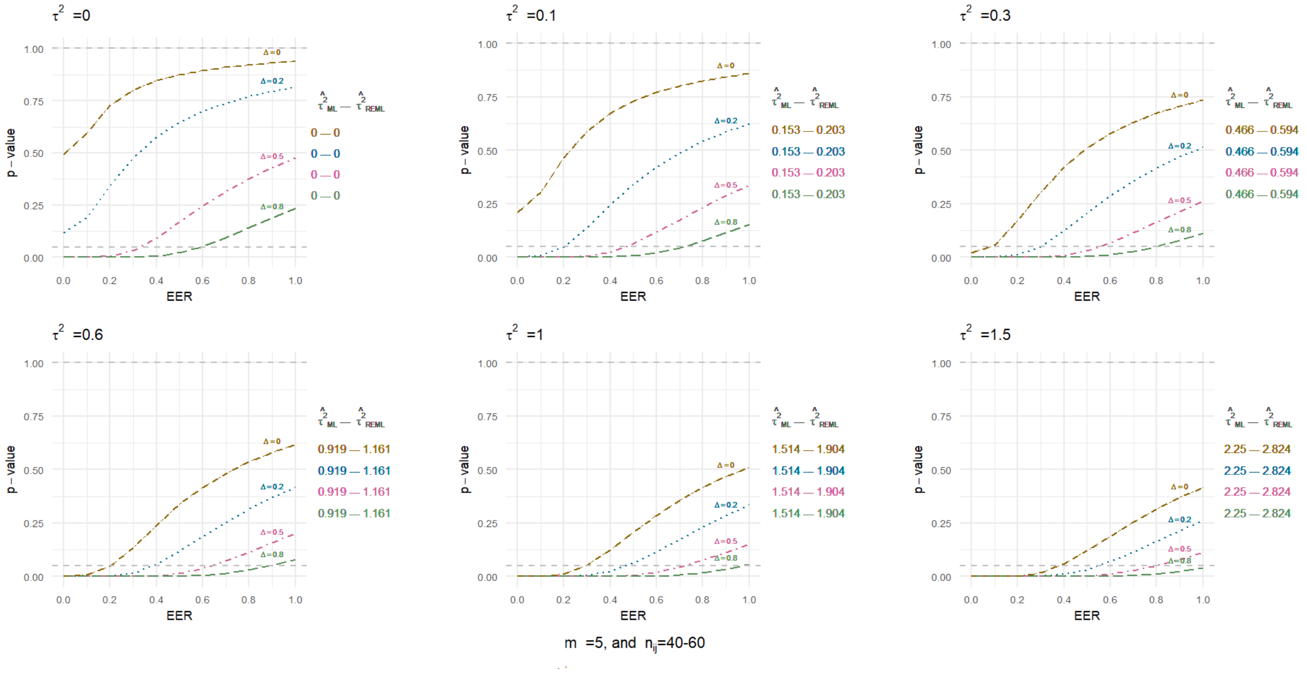


Figure C.46: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and small* m

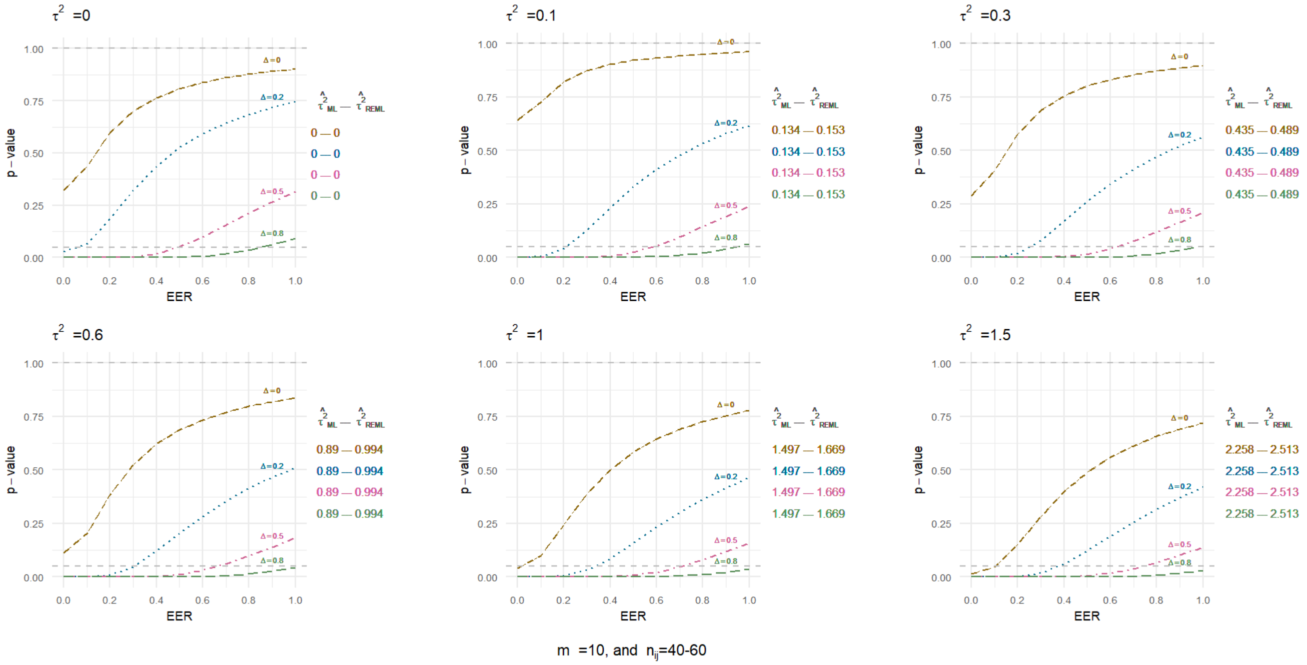


Figure C.47: *Mixed model (N-m-1 Degree of Freedo): Sensitive analysis with various value of* τ *and* Δ *under large* n_{ij}*, where* n_{1j} *not necessary equal to* n_{2j} *and medium* m

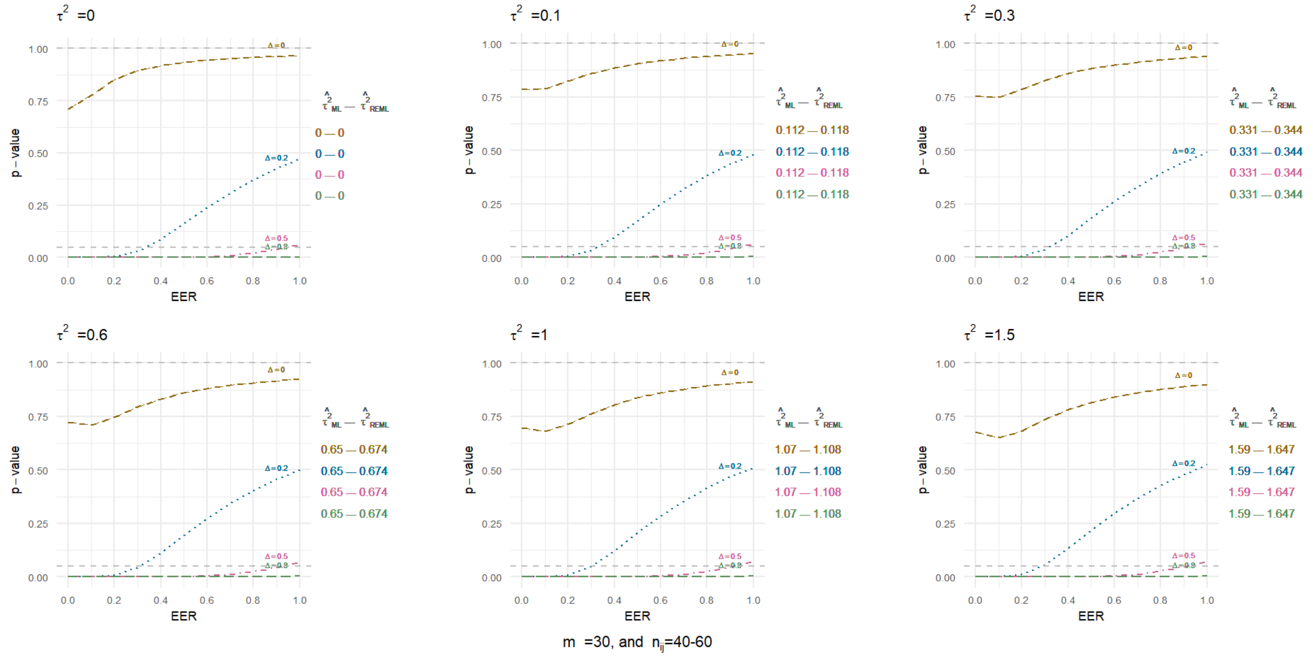


Figure C.48: *Mixed model (N-m-1 Degree of Freedom): Sensitive analysis with various value of τ and Δ under large n_{ij}, where n_{1j} not necessary equal to n_{2j} and large m*

Appendix D

Supplemental Materials For Chapter 4

D.1 Many Labs 4

D.1.1 Prior Publication

Klein et al. (2019) attempted to replicate Study 1 of Greenberg et al. (1994) in 21 labs. The 21 labs were randomly assigned to either expert protocol or in-house protocol under two conditions. There are 13 in-house Labs and 9 Expert Labs with total number 2281 participants.

Klein analyzed individual study under three exclusion sets. Therefore, there were three summary data. Klein provided these aggregate data in his supplemental materials. However, after we re-estimate the summary data, and the whole data, we find that For Exclusion Set 2, some participants who were not White Americans were included in the analysis, which led to a total number of 1,916 instead of 1,874, see Table D.4. Similarly, for Exclusion Set 3, some participants who did not strongly identify with the U.S. were included for a total of 1,723 instead of 1,693, see Table D.5. The total number of participants in in-house labs is 1,421, which is correct in both Exclusion Sets 2 and 3.

Table D.1: *Data Collection Sites: 21 Labs*

Locations	Author Advised Labs (AA) / In House Labs (IH)	Sample Sizes N
1. Ashland University(AU)	AA	56
2. Azusa Pacific University(APU)	IH	30
3. Brigham Young University(BYU)	AA	81
4. The College of New Jersey(TCNJ)	AA	136
5. University of Illinois(UI)	IH	87
6. Ithaca College(IC)	IH	177
7. University of Kansas(UK)	IH	75
8. University of Kansas(UK)	AA	43
9. Occidental College(OC)	AA	88
10. Pace University(PU)	IH	58
11. Pace University(PU)	AA	106
12. Pacific Lutheran University(PLU)	IH	246
13. University of California, Riverside(UCR)	AA	107
14. Southern Oregon University(SOU)	IH	29
15. University of Florida (UF)	IH	252
16. University of Pennsylvania(UP)	IH	86
17. University of Wisconsin(UW)	IH	68
18. University of Wisconsin (UW)	AA	79
19. Virginia Commonwealth University(VCU)	AA	103
20. Wesleyan University(WU)	IH	174
21. Worcester Polytechnic Institute(WPI)	IH	139

Table D.3: *Summary 1 of Exclusion 1 as reported*

Labs	AA/IH	N(TV)	N(MS)	Mean(TV)	Mean(MS)	SD (TV)	SD (MS)	Hedges' g	p-value
Ashland University	AA	28	28	1.88	1.14	1.75	1.93	-0.39	0.14
Azusa Pacific University	IH	15	15	0.22	0.09	1.01	1.54	-0.1	0.78
Brigham Young University	AA	41	40	1.77	2.65	2.06	2.06	0.42	0.06
The College of New Jersey	AA	60	76	1.14	1.57	1.96	1.69	0.24	0.18
University of Illinois	IH	44	43	-0.78	0.64	1.8	2.01	0.74	0
Ithaca College	IH	87	90	-0.41	-0.53	2.14	2.78	-0.05	0.76
University of Kansas	AA	18	25	1.07	0.89	2.31	2.51	-0.07	0.81
University of Kansas	IH	40	35	1.32	1.21	3.03	3.04	-0.03	0.88
Occidental College	AA	42	46	0.33	0.52	1.66	2.26	0.09	0.66
Pace University	AA	53	53	1.33	1.18	1.89	2.05	-0.08	0.68
Pace University	IH	34	24	2.2	0.11	5.93	5.33	-0.36	0.17
Pacific Lutheran University	IH	125	121	0.37	0.36	1.85	1.97	-0.01	0.96
University of California, Riverside	AA	52	55	0.79	0.73	1.68	1.72	-0.04	0.84
Southern Oregon University	IH	14	15	-1.02	-0.18	1.68	1.25	0.56	0.14
University of Florida	IH	107	145	0.99	0.76	1.47	1.23	-0.17	0.19
University of Pennsylvania	IH	45	41	-0.16	0.2	1.75	2.23	0.17	0.42
University of Wisconsin	AA	39	40	0.94	0.91	1.72	2.16	-0.02	0.94
University of Wisconsin	IH	36	32	-0.74	-0.93	2.05	1.77	-0.1	0.69
Virginia Commonwealth University	AA	42	61	1.33	1.3	1.15	1.99	-0.01	0.94
Wesleyan University	IH	97	77	-0.2	-0.09	1.81	1.78	0.06	0.67
Worcester Polytechnic Institute	IH	68	71	0.53	0.51	1.96	1.61	-0.01	0.95

Table D.4: *Summary 2 of Exclusion 2 as reported*

Labs	AA/IH	N(TV)	N(MS)	Mean(TV)	Mean(MS)	SD (TV)	SD (MS)	Hedges' g	p-value
Ashland University	AA	26	23	1.83	1.39	1.85	2.02	-0.22	0.45
Azusa Pacific University	IH	15	15	0.22	0.09	1.01	1.54	-0.1	0.78
Brigham Young University	AA	34	38	1.83	2.61	2.17	2.02	0.37	0.12
The College of New Jersey	AA	42	59	1.27	1.68	2.21	1.74	0.21	0.34
University of Illinois	IH	44	43	-0.78	0.64	1.8	2.01	0.74	0
Ithaca College	IH	87	90	-0.41	-0.53	2.14	2.78	-0.05	0.76
University of Kansas	IH	40	35	1.32	1.21	3.03	3.04	-0.03	0.88
University of Kansas	AA	11	18	1.76	0.8	2.41	2.49	-0.38	0.32
Occidental College	AA	18	28	0.57	0.79	1.74	2.76	0.09	0.76
Pace University	IH	34	24	2.2	0.11	5.93	5.33	-0.36	0.17
Pace University	AA	34	33	1.69	1.53	2.01	2.19	-0.08	0.77
Pacific Lutheran University	IH	125	121	0.37	0.36	1.85	1.97	-0.01	0.96
University of California, Riverside	AA	13	13	0.44	2.33	0.89	1.25	1.69	0.05
Southern Oregon University	IH	14	15	-1.02	-0.18	1.68	1.25	0.56	0.14
University of Florida	IH	107	145	0.99	0.76	1.47	1.23	-0.17	0.19
University of Pennsylvania	IH	45	41	-0.16	0.2	1.75	2.23	0.17	0.42
University of Wisconsin	IH	36	32	-0.74	-0.93	2.05	1.77	-0.1	0.69
University of Wisconsin	AA	31	34	0.88	0.86	1.68	2.27	-0.01	0.97
Virginia Commonwealth University	AA	12	28	1.22	1.49	1.13	2	0.15	0.59
Wesleyan University	IH	97	77	-0.2	-0.09	1.81	1.78	0.06	0.67
Worcester Polytechnic Institute	IH	68	71	0.53	0.51	1.96	1.61	-0.01	0.95

Total number is 1916 because some participants who were not white Americans were included in the analysis. Total number should be 1,874. The red numbers indicate incorrect numbers.

Table D.5: *Summary 3 of Exclusion 3 as reported*

Labs	AA/IH	N(TV)	N(MS)	Mean(TV)	Mean(MS)	SD (TV)	SD (MS)	Hedges' g	p-value
Ashland University	AA	18	14	2.43	1.76	1.85	2.29	-0.32	0.39
Azusa Pacific University	IH	15	15	0.22	0.09	1.01	1.54	-0.1	0.78
Brigham Young University	AA	28	31	1.89	2.69	2.23	1.99	0.37	0.15
The College of New Jersey	AA	23	32	1.98	1.74	2.31	1.66	-0.12	0.69
University of Illinois	IH	44	43	-0.78	0.64	1.8	2.01	0.74	0
Ithaca College	IH	87	90	-0.41	-0.53	2.14	2.78	-0.05	0.76
University of Kansas	AA	6	11	2.89	0.67	2.79	2.22	-0.87	0.13
University of Kansas	IH	40	35	1.32	1.21	3.03	3.04	-0.03	0.88
Occidental College	AA	6	14	0.5	2.1	1.38	3.42	0.51	0.21
Pace University	AA	17	19	2	2.21	2.21	2.16	0.09	0.79
Pace University	IH	34	24	2.2	0.11	5.93	5.33	-0.36	0.17
Pacific Lutheran University	IH	125	121	0.37	0.36	1.85	1.97	-0.01	0.96
University of California, Riverside	AA	10	8	1	2.33	0.88	1.53	1.05	0.28
Southern Oregon University	IH	14	15	-1.02	-0.18	1.68	1.25	0.56	0.14
University of Florida	IH	107	145	0.99	0.76	1.47	1.23	-0.17	0.19
University of Pennsylvania	IH	45	41	-0.16	0.2	1.75	2.23	0.17	0.42
University of Wisconsin	AA	19	23	1.18	1.44	1.65	2.28	0.13	0.67
University of Wisconsin	IH	36	32	-0.74	-0.93	2.05	1.77	-0.1	0.69
Virginia Commonwealth University	AA	7	16	1.62	2.19	1.22	1.88	0.32	0.41
Wesleyan University	IH	97	77	-0.2	-0.09	1.81	1.78	0.06	0.67
Worcester Polytechnic Institute	IH	68	71	0.53	0.51	1.96	1.61	-0.01	0.95

Total number is 1723 because some participants who did not strongly identify with the U.S. were included in the analysis. Total number should be 1,693. The red numbers indicate incorrect numbers.

D.1.2 Publication

Klein et al. (2022) published their results after following their preregistration plan by excluding four labs that has less than 60 participants, and excluding some data in In House labs that collected before the analysis plan was preregistered (545 participants were excluded). Therefore, only 17 labs were remained with a total 1,578 participants. Note, the Pace University in In House site is included because they provided more than 60 participants in total.

Table D.6: *Date Collection Sites: 17 Labs*

Locations	Author Advised Labs (AA) / In House Labs (IH)	Sample Sizes N
1. Brigham Young University(BYU)	AA	81
2. The College of New Jersey(TCNJ)	AA	135
3. University of Illinois(UI)	IH	89
4. Ithaca College(IC)	IH	81
5. University of Kansas(UK)	IH	75
6. Occidental College(OC)	AA	88
7. Pace University(PU)	IH	58
8. Pace University(PU)	AA	106
9. Pacific Lutheran University(PLU)	IH	60
10. University of California, Riverside(UCR)	AA	107
11. University of Florida (UF)	IH	98
12. University of Pennsylvania(UP)	IH	86
13. University of Wisconsin(UW)	IH	68
14. University of Wisconsin (UW)	AA	79
15. Virginia Commonwealth University(VCU)	AA	103
16. Wesleyan University(WU)	IH	95
17. Worcester Polytechnic Institute(WPI)	IH	141

Klein analyzed individual study under three exclusion sets. Therefore, there were three summary data. Klein provided these aggregate data in their supplemental materials. Here, we provide part of his summary data, the p-value.

Table D.7: *17 Labs: p-value for each study under 3 differed exclusions*

		$p - value$		
Labs	AA/IH	Exclusion Set 1	Exclusion Set 2	Exclusion Set 3
Brigham Young University	AA	0.06	0.12	0.15
Ithaca College	IH	0.9	0.9	0.9
Occidental College	AA	0.66	0.76	0.21
Pace University	AA	0.68	0.77	0.79
Pace University	IH	0.17	0.17	0.17
Pacific Lutheran University	IH	0.33	0.33	0.33
The College of New Jersey	AA	0.22	0.34	0.69
University of California, Riverside	AA	0.84	0.05	0.28
University of Florida	IH	0.83	0.83	0.83
University of Illinois	IH	0	0	0
University of Kansas	IH	0.88	0.88	0.88
University of Pennsylvania	IH	0.42	0.42	0.42
University of Wisconsin	AA	0.94	0.97	0.67
University of Wisconsin	IH	0.69	0.69	0.69
Virginia Commonwealth University	AA	0.94	0.59	0.41
Wesleyan University	IH	0.32	0.32	0.32
Worcester Polytechnic Institute	IH	0.92	0.92	0.92

D.2 Follow-up Studies

Hoogeveen et al. (2023) conducted a Bayesian hierarchical modeling and bayesian model-averaged Meta-analysis instead of meta-analytic approach. These method had applied to different exclusions criteria. The data and the R code are at (`https://github.com/SuzanneHoogeveen/ml4-reanalysis`).

Table D.8: *Re-analyze ML4 by bayesian hierarchical modeling*

	Exclusions	Exclusion Criteria	Labs	$\hat{\Delta}$	95% CI
First Re-analysis	Set 1	Apply on AA	17 Labs	0.02	$[-0.12,\ 0.16]$
	Set 2	Apply on AA	17 Labs	0.04	$[-0.11,\ 0.19]$
	Set 3	Apply on AA	17 Labs	0.05	$[-0.11,\ 0.21]$
Second Re-analysis	Set 1	Apply on AA	Expert Labs (7)	0.08	$[-0.12,\ 0.28]$
	Set 2	Apply on AA	Expert Labs (7)	0.14	$[-0.11,\ 0.39]$
	Set 3	Apply on AA	Expert Labs (7)	0.18	$[-0.12,\ 0.49]$
Third Re-analysis	Set 1	Apply on both	21 Labs	0.01	$[-0.11,\ 0.12]$
	Set 2	Apply on both	16 Labs	−0.04	$[-0.20,\ 0.12]$
	Set 3	Apply on both	9 Labs	0.05	$[-0.22,\ 0.32]$

Effect size estimates similar to Cohen's d

Table D.9: *Re-analyze ML4 by bayesian model-averaged Meta-analysis*

	Exclusions	Exclusion Criteria	Labs	$\hat{\Delta}$	95% CI
First Re-analysis	Set 1	Apply on AA	17 Labs	0.06	$[-0.06,\ 0.18]$
	Set 2	Apply on AA	17 Labs	0.09	$[-0.05,\ 0.22]$
	Set 3	Apply on AA	17 Labs	0.08	$[-0.06,\ 0.23]$
Second Re-analysis	Set 1	Apply on AA	Expert Labs (7)	0.08	$[-0.10,\ 0.25]$
	Set 2	Apply on AA	Expert Labs (7)	0.16	$[-0.08,\ 0.41]$
	Set 3	Apply on AA	Expert Labs (7)	0.18	$[-0.10,\ 0.47]$
Third Re-analysis	Set 1	Apply on both	21 Labs	0.03	$[-0.07,\ 0.13]$
	Set 2	Apply on both	16 Labs	−0.03	$[-0.17,\ 0.13]$
	Set 3	Apply on both	9 Labs	0.07	$[-0.20,\ 0.34]$

The overall effect size was estimated across studies (Hedges' g), without constraining the direction