

MULTIPLE TESTING, MULTILEVEL SOLUTIONS: REVISITING THE MULTIPLICITY
PROBLEM WITH FREQUENTIST AND BAYESIAN HIERARCHICAL MODELING

UDI ALTER

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

APRIL 2026

© UDI ALTER, 2026

ABSTRACT

The longstanding multiplicity problem—the inflation of false positives that arises from conducting multiple significance tests—remains a source of confusion and disagreement in psychology and many other sciences. Traditional solutions, such as familywise error rate (FWE; e.g., Bonferroni) and false discovery rate (FDR; e.g., Benjamini-Hochberg) corrections, impose control externally by adjusting p values and confidence intervals. These post-hoc adjustments can reduce false positives but often at the expense of statistical power, making it harder to detect true effects when many hypotheses are tested simultaneously. Hierarchical modeling incorporates multiplicity as part of the model itself and allows information to be shared appropriately across tests. As a result, the inferences drawn about the effect sizes, uncertainty intervals, and statistical significance are often more justifiable. The first simulation study evaluated frequentist multilevel modeling (MLM) as an alternative to post-hoc adjustment procedures. Results provided initial evidence that partial pooling and shrinkage can help address the multiplicity problem, particularly when cluster sizes and effect magnitudes are large. Although MLM is not a universal solution, it offers a substitute for mechanical corrections, but only under certain conditions. The second simulation study extended the investigation into Bayesian multilevel modeling (BMLM). Findings were mixed: when the joint null hypothesis was true (i.e., all effects were null), most BMLM prior variants maintained very low false-positive rates—well below the nominal Type I error rate and other correction methods—while retaining high power. Yet in partial-null settings, where true and null effects coexisted, strong shrinkage and narrow credible intervals raised familywise error above that of MLM, FWE, and FDR. Together, the two studies demonstrate that hierarchical models can, under the right conditions, embed multiplicity control within the model itself rather than impose it after the fact. However, their performance depends critically on the structural and design conditions, effect configuration, and modeling assumptions. This work adds empirical evidence to ongoing theoretical discussions of how

multilevel and Bayesian approaches can promote more transparent, coherent, and defensible inference.

DEDICATION

*To Adi and Milo,
Mom and Dad,
and Saftas and Sabas*

ACKNOWLEDGMENTS

There is a saying, perhaps best captured by Richard McElreath, that once you learn about multilevel modeling, you start seeing it everywhere. The variability in the graduate school experience has, in many ways, a multilevel structure. Both the journey and outcome vary considerably across students, but much of that variability lives at the supervisor level. The intraclass correlation, I suspect, is high. I was already fortunate that Rob Cribbie's cluster had a remarkably high intercept, but little did I know how steep that slope was going to be.

Rob is, in the truest sense of the word, a mentor. He is also a brilliant teacher, a colleague, a coach, and a friend. And somehow, all of these roles at once, without any of them diminished by the others. What has struck me most throughout this journey is not just his generosity or his kindness, though both are extraordinary on their own, but his unwavering commitment to my best interest. Rob always had my back, always. Not sometimes, not when it was convenient, always; he advocated for me at every opportunity. If the opening paragraph of these acknowledgments reads as if it came from somebody who loves what they do, Rob Cribbie is the reason why.

As is often the case with multilevel models, observations are not always nested within a single cluster. In a cross-classified model, a lower-level unit can simultaneously be nested within multiple, overlapping contexts. I was fortunate enough to be that kind of unit:

Dave Flora was, in many ways, my introduction to the world that this dissertation inhabits. It was his multilevel modeling course (and the privilege of later TAing for it) that first drew me into this material seriously. Dave is an extraordinary teacher and a genuinely rigorous thinker, and I always found myself looking forward to hearing his opinion, especially when it

was a strong one. The discussions we had, both conceptual and technical, were among the best pleasures of this degree.

Georges Monette is, to put it plainly, a gem. A mind of rare depth and precision, paired with a humility and warmth that make him one of the most disarming people I have ever had the pleasure of learning from. Despite being, by any measure, the most sought-after presence at the Statistical Consulting Service, Georges never made me feel like an imposition or impostor. He gave his time generously, his explanations patiently, and his perspectives—always philosophically rich and genuinely open—changed the way I think about this work. I am deeply thankful.

Monique Herbert chaired my examining committee, and I am so glad she did. There is something about having someone so sincerely invested in your success at the head of that table that makes even the most stressful day feel manageable. But Monique has been much more than a committee chair. She is a mentor and a friend, someone I could go to for honest advice about teaching, career, and the bigger questions that come with building a life in academia. I am truly grateful for all of it.

I would also like to thank Jonathan Fawcett for serving as my external examiner. His questions and comments during the defence were incredibly insightful and pushed my thinking in ways I did not anticipate. I am so appreciative that he agreed to take on this role, and I look forward to what I hope will be many more conversations.

Phil Chalmers is, and I say this with full awareness of what it means, a once-in-a-generation mind. What makes Phil remarkable is not just the creativity and depth of his thinking, but the fact that none of it comes with any pretense. He is as generous and approachable as he is brilliant, and he was indispensable to this dissertation in ways that go well beyond

acknowledgment. The simulations that form the backbone of this work would simply not exist without his guidance (or his software!), his patience, and his willingness to wade into the technical weeds with me at every stage. I am flattered to call him a colleague, and even more so, a friend.

Jodi Martin has been a constant and grounding presence throughout this degree. As a teaching and career mentor, she helped me think about who I want to be, not just as a researcher but as an educator and a professional. As a friend, she offered something even more rare still: a safe space to think out loud, to vent, to ask the questions I was not sure I was allowed to ask. Her door was always open. I am so lucky to have had her in my corner.

Michael Friendly occupies a category all his own. If Rob is my academic father, then Michael is, without question, my academic uncle—present, generous, and influential in that particular way that the best mentors are, not by holding an official supervisory role, but simply by being who they are. His support and encouragement throughout this journey have meant more than he probably knows.

I also want to thank the Cribbie Lab. Nataly Beribisky deserves special mention. In many ways, she paved the path I walked; a few steps ahead of me, always generous enough to look back and help me find my footing. Nataly opened the door to this lab for me, and her kindness and warmth throughout my time here have meant more than she knows. She is a close and trusted friend, and I owe her a great deal. I could not have asked for a better person to have walked alongside through this.

Naomi is my cohort, my friend, and my lab buddy. And to Victoria (my academic younger sister), Linda, and Evelyn, thank you for the friendship, the conversations, and the kind of everyday support that makes the harder days much easier.

To Miranda, Adina, Sarah, and Sam: kindred spirits and close friends I could always talk to freely, confide in, and laugh with. You made this journey lighter and much more enjoyable.

A special thank you to Avi Inbar, my closest friend and “other significant other,” for being the voice of reason and helping me through difficult times, pleasant times, and everything in between.

To Milo, my son: you arrived in the middle of all of this, somewhere between the literature review and writing the simulation code for Study 1, and you made everything simultaneously harder and infinitely more consequential. I still do not fully understand how I survived nine months of sleepless nights and Rob’s 8:30 am classes, but by some miracle, I did, and so did you, and here we are. You are my greatest joy.

To Adi, my wife: there are not enough words, and the ones that exist do not feel enough. You have been my anchor throughout this entire journey; patient past reason, unconditionally supportive beyond measure, and sane in moments when I was anything but. I will spend a lifetime being grateful for that, and for you. And as if all of that is not enough, as I write this, you are preparing to leave everything familiar behind to move to a new place, far away, where neither of us has ever been, simply so I can do the work I love. That is not a small thing. That is everything. You are everything. This dissertation, and everything it means to me, belongs to both of us.

TABLE OF CONTENTS

Abstract	ii
Dedication	iv
Acknowledgments	v
Table of Contents	ix
List of Tables	xiii
List of Figures	xiv
 Chapter One: General Introduction	 1
 Chapter Two: Shrinking Concerns Over Multiple Testing: Addressing the Multiplicity Problem from a Multilevel Modeling Perspective (Study 1)	 4
The Multiplicity Problem	5
The Never-Settled Multiplicity Adjustment Debate: To Adjust or Not to Adjust, That is the Question	6
Not to Adjust: Philosophical Considerations	7
Not to Adjust: Practical Considerations and Issues with Multiplicity Control Adjustments	9
To Adjust: Popular Multiplicity Control Approaches and Procedures	11
Familywise Error (FWE) Rate	11
False Discovery Rate (FDR)	12
Controlling Without Adjusting: Introducing a Multilevel Modeling (MLM) Approach to Multiplicity Control	13
Estimating Cluster-Specific Effects Through Partial Pooling	19
Best Linear Unbiased Prediction (BLUP) and Empirical BLUP (EBLUP)	19
Precision of the Cluster-Specific Predictions	21
Constructing Uncertainty Intervals for the Cluster-Specific Predictions	22
Determining Statistical Significance for Cluster-Specific Effects	23
Hypothesis Testing	23
Shrinkage and Its Implications for Significance Testing	24
Prior Research on Multilevel Modeling and the Multiplicity Problem	25

Current Research	27
Method.....	28
Monte Carlo Simulation Study	28
Objective.....	28
Design.....	29
Procedure	30
Evaluating the Performance of Testing Approaches	32
Results	33
Type I Error Rates.....	33
Complete-Null Conditions.....	33
Partial-Null Conditions	34
Effect Size.....	34
Within-cluster Sample Size	35
Number of Clusters.....	35
Within-Cluster Variability.....	36
Power Rates	36
Power-Only Conditions	36
Partial Null Conditions	37
Effect Size.....	38
Within-cluster Sample Size	38
Number of Clusters.....	38
Within-Cluster Variability.....	38
Discussion.....	39
Summary of Findings	39
Implications for Practice and Recommendations	41
Limitations and Future Directions	43
Conclusion	45

Chapter Three: A Bayesian Solution to a Frequentist Problem? Addressing the Multiplicity Problem with Bayesian Hierarchical Modeling (Study 2).....	46
How Bayesians View and Approach the Multiplicity Problem.....	48
A Bayesian Framework for Multiplicity.....	50
Hierarchical Models: Model-Based Control.....	51
Bayesian Multiplicity Control (in Variable Selection)	53
Shrinkage Priors.....	55
Spike-and-Slab Priors.....	55
Horseshoe Priors.....	56
Bayesian Multilevel Modeling (BMLM).....	57
A Note on Statistical Significance	63
Current Research	63
Method.....	64
Monte Carlo Simulation Study	64
Objective.....	64
Design.....	64
Procedure	65
Evaluating the Performance of Testing Approaches	68
Results	68
Type I Error Rates.....	68
Complete-Null Conditions.....	68
Partial-Null Conditions.....	69
Proportion of Null Clusters.....	70
Effect Size.....	70
Within-cluster Sample Size	70
Number of Clusters.....	72
Prior Specifications.....	72
Power Rates	74
Discussion.....	75
Summary of Findings	75

Beyond Error Rates: Effect Sizes, Precision, and Just a Touch of Context.....	77
Limitations and Future Directions	82
Conclusion	84
 Chapter Four: General Discussion	 86
 References.....	 89
Tables	98
Figures.....	100
Appendices.....	108
Appendix A.....	108
Appendix B.....	110

LIST OF TABLES

Table 1.1 Summary of Manipulated Factors and Simulation Conditions in Study 1

Table 2.1 Summary of Manipulated Factors and Simulation Conditions in Study 2

LIST OF FIGURES

- Figure 1.1 Familywise Type I Error Rates in Complete-Null Conditions in Study 1
- Figure 1.2 Familywise Type I Error Rates in Partial-Null Conditions in Study 1
- Figure 1.3 Power Rates in Study 1
- Figure 2.1 Simulation Prior Specifications in Study 2
- Figure 2.2 Familywise Type I Error Rates in Complete-Null Conditions in Study 2
- Figure 2.3 Familywise Type I Error Rates of MLM and BMLM with Different Prior
Specifications in Study 2
- Figure 2.4 Illustrative Example of Cluster Effects Using MLM and Different Priors
- Figure 2.5 Power Rates in Small Effect Sizes ($b_{1j} = 1$) Conditions in Study 2
- Figure A1 Familywise Type I Error Rates Across Within-Cluster Variability and Effect Size
Levels in Half-Null Conditions (Proportion of Null Clusters is .5) in Study 1
- Figure A2 Power Rates Across Within-Cluster Variability and Effect Size Levels in Half-
Null Conditions (Proportion of Null Clusters is .5) in Study 1
- Figure B1 Per-Comparison Type I Error Rates of MLM and BMLM with Different Prior
Specifications in Study 2

CHAPTER ONE

GENERAL INTRODUCTION

Researchers across nearly every scientific discipline have long grappled with the inflation of false positives that occurs when conducting multiple statistical tests. This issue—known as the multiple comparisons, or multiplicity, problem—is particularly relevant to frequentist inference, the dominant approach in psychological research. Approaches to control the overall Type I error rate, such as the familywise error rate (FWE; e.g., Bonferroni procedure) and false discovery rate (FDR; e.g., Benjamini-Hochberg procedure) adjustments, have been available for decades, and yet confusion and inconsistency in their application (and even in their perceived necessity) are still actively debated.

Although the risk of increased false positives due to multiple testing is widely acknowledged, there is still no agreement on whether it is a feature that demands proactive control or it is simply an unavoidable byproduct of statistical inference. Experts have taken firm stances at opposite ends of this spectrum. For instance, Rothman (1990) argued that “no adjustments are needed for multiple comparisons,” while others have insisted that any setting involving multiple tests must, without exception, apply formal adjustments. In practice, most researchers seem to fall somewhere in between; they recognize that multiplicity control is sometimes, but perhaps not always, necessary.

Even with this recognition, multiplicity adjustments have neither become routine nor consistent in many disciplines, including psychology (Martinez Gutierrez et al., 2025). Poole (1991) illustrated this tension with a humorous anecdote, recalling a time when he and a group of 15 of his seasoned statistician colleagues were asked about a familiar epidemiological problem and how they addressed multiple comparisons. He writes:

We responded by fidgeting slightly in our seats. When our disappointed leader asked if any of us had made any multiple-comparisons adjustments in the past five years or so, not a single hand went up. Someone finally broke the tension by joking about the distinction between theory and practice. (p. 241)

This anecdote captures the current state of affairs in many fields: researchers generally support the idea of multiplicity control in theory but rarely implement it in practice, often due to confusion and disagreement about when and how such adjustments should be applied.

Such ongoing confusion and disagreement have also been shown empirically. Martinez Gutierrez et al. (2025) reviewed recent articles from leading psychology journals and found that roughly 82% of studies did not apply any form of multiplicity control. Even among those that did, adjustments were inconsistently implemented—often limited to a subset of analyses—and many lacked a clear rationale. This pattern has consequences that extend far beyond individual studies. Undermanaged error rates can threaten the reliability of reported findings and contribute directly to the broader reproducibility crisis in psychological research (Ioannidis, 2005; Open Science Collaboration, 2015).

Given its potential implications for the credibility, replicability, and interpretability of psychological science, the multiplicity problem is far more than a technical nuisance or a box researchers must check as part of their analyses—it reflects deeper questions about how evidence is analyzed, interpreted, and reported. The studies that follow approach this issue from a modeling perspective, framing multiplicity not as a problem of post-hoc adjustment, but as a question of modeling. Specifically, the studies in this dissertation explore the extent to which multilevel frameworks—both frequentist and Bayesian—can provide a principled and integrated approach to managing error inflation while maintaining interpretability and statistical power.

This dissertation consists of two complementary simulation studies that together examine the multiplicity problem from both frequentist and Bayesian multilevel modeling (MLM) perspectives. Although each study is designed to stand on its own as a complete, publishable manuscript, the two form a coherent and cumulative line of inquiry. Study 1 evaluates whether MLM, within the frequentist framework, can serve as a model-based alternative to traditional multiplicity control procedures. Specifically, it compares the performance of MLM against conventional approaches such as FWE and FDR adjustments, focusing on their ability to balance familywise Type I error control with statistical power across multiple research conditions. Study 1 thus establishes empirically the foundation for viewing multiplicity as a modeling challenge rather than a post-hoc correction problem.

Building on that foundation, Study 2 extends the investigation into the Bayesian realm. Study 2 examines whether Bayesian multilevel modeling (BMLM) can offer an even more coherent and flexible framework for managing multiplicity by incorporating prior information and quantifying uncertainty. The study systematically varies prior specifications to assess how different assumptions—ranging from completely uninformative or skeptical to confident and precise—affect familywise Type I error and statistical power rates. Through this design, Study 2 explores how priors can serve as a form of regularization, influencing the degree of estimate shrinkage and, consequently, the overall false-positive rate. Together, these two studies provide a coherent examination of how multilevel modeling principles can be applied to address one of the oldest and most persistent dilemmas in quantitative methodology.

CHAPTER TWO

SHRINKING CONCERNS OVER MULTIPLE TESTING: ADDRESSING THE MULTIPLICITY PROBLEM FROM A MULTILEVEL MODELING PERSPECTIVE (STUDY 1)

Researchers from nearly every scientific discipline have long grappled with the issue of inflated Type I error (false positive) rates when conducting multiple significance tests. This issue, commonly known as the *multiple comparisons* or *multiplicity* problem, is of particular concern in frequentist statistics, the traditional and most common mode of statistical inference in psychology. To combat this problem, recommendations for how to control the Type I error rate across a set of statistical tests, such as familywise error rate control and false discovery rate control, have been in place for decades.

When Wilkinson and the Task Force on Statistical Inference (1999) wrote about statistical issues in psychological research, multiplicity control received more page space than any other issue. Yet, the multiplicity problem still presents a complicated and confusing landscape, lacking clear guidelines and consensus regarding when and how researchers should implement control (e.g., what constitutes a “family” of hypotheses? under which scenarios is multiplicity control required?). Remarkably, even experts in the field remain divided on whether and how to address Type I errors in the context of multiple analyses (see Cribbie, 2017). Therefore, it is no surprise that multiplicity control is still inconsistently applied, even two decades following the Wilkinson report. In a systematic review of top psychology journals in 2021, Martinez Gutierrez et al. (2025) found that 82% of articles did not implement any form of multiplicity control, and those that did showed little consistency. Even among articles that imposed control, adjustments were

not made for all analyses; only 8% of tests across all research articles were adjusted, in some manner, for multiplicity.

The Multiplicity Problem

Often, modern data analysis and psychological research consist of numerous, simultaneous hypothesis testing. In fact, Martinez Gutierrez et al. (2025) found that the median number of hypotheses tested per article in psychology is 76. Hypothesis testing includes *any* statistical significance test with a null hypothesis, an alternative hypothesis, and a threshold (e.g., $\alpha = 0.05$) that, if surpassed, the evidence against the null hypothesis is determined sufficient to reject the null hypothesis. Of course, this rejection of the null does not mean that the alternative hypothesis is necessarily true; instead, rejecting the null hypothesis implies that the observed effect is large enough relative to the variability in the data that, if the null hypothesis is indeed true (i.e., a Type I error), we would only expect to reject the null every $1/\alpha$ times the same test was performed simply due to random sampling.

Multiple comparisons or multiple tests refer to evaluating two or more significance tests of the same set or family, such as comparing multiple means in a pairwise manner, testing the correlations among a set of variables, or testing regression coefficient estimates from a model (e.g., multiple regression, structural equation model). However, a set can also include a mix of different types of analyses or hypothesis tests. When more than a single significance test is performed simultaneously, the risk that *at least one* Type I error will occur will necessarily be greater than α . For instance, suppose three independent *t*-tests are conducted in a study, each with a Type I error probability of $\alpha = .05$, then the probability that one or more of the null hypotheses will be incorrectly rejected will approach $1-(1-.05)^3 = 1-.86 = .14$ (assuming the tests are independent of one another). Now, instead of maintaining our agreed-upon threshold of an

“acceptable” 5% chance of a false-positive, we drift into 14%. And this probability grows rapidly with every additional test added to the set, quickly reaching nearly 100% as the number of tests increases, ensuring that at least one false-positive will take place once enough nulls are true. Hence, the multiplicity problem. But must researchers care about or solve this problem? And if so, how?

The consequences of this unchecked inflation of Type I error extend beyond individual studies. When many of the reported “statistically significant” effects in the literature are in fact false positives, the likelihood that independent replication attempts will confirm these findings is greatly reduced. This dynamic—exacerbated by the large number of tests often conducted within a single study, selective reporting, flexibility in designs, definitions, outcomes, and analytical decisions, and the inconsistent use of multiplicity adjustments—has been identified as a key contributor to the reproducibility crisis in psychology and other sciences (Ioannidis, 2005; Open Science Collaboration, 2015).

The Never-Settled Multiplicity Adjustment Debate: To Adjust or Not to Adjust, That is the Question

Games (1971) wrote “[t]he area of multiple comparisons is one of the more confusing areas of statistics, and is one that receives a widely differing set of recommendations from many applied statistics texts in the behavioral sciences” (p. 531). More than half a century has passed since Games’s article, but the same philosophical and statistical discussions remain. The debate on when and how to adjust for multiple significance tests dates back to 1935, when Fisher suggested Bonferroni-type corrections (Frane, 2020). Over the years, the literature has been dominated by techniques and recommendations aimed at adjusting for multiple tests, specifically to control familywise error rates (Cribbie, 2017; Hancock & Klockars, 1996). However, by the

end of the last century, opposing arguments emerged, challenging the necessity of such adjustments. Notable anti-adjustment arguments by Rothman (1990), Saville (1990, 2003), and Perneger (1998), along with a growing dissatisfaction with the null hypothesis significance testing framework and a more recent call to shift focus toward effect size and precision (e.g., Cummings, 2014), have reformed and reignited the old debate (Cribbie, 2017).

Not to Adjust: Philosophical Considerations

Pressured by the academic jungle to produce statistically significant results at every turn, the lone researcher would find it more convenient and self-serving to ignore the multiplicity problem, thus avoiding adjustment complexities and increasing their chances of rejecting the null hypotheses by not compromising statistical power. However, neither of these reasons—a greater chance of favourable results and avoiding complexities—is valid or justifiable on its own for abandoning multiplicity control. The multiplicity problem is, at heart, philosophical. As such, if researchers decide not to adjust for multiple tests, they should do so for the right reasons. Here are some of the more prominent ones:

1. “There is no spoon”: The goal of adjusting for multiple testing is to put a cap on the probability of at least one Type I error. The probability of at least one Type I error depends on the number of hypotheses tested that are truly null in the population. But, what if there is no such thing as a Type I error? When would one expect a perfectly null (more specifically, nil) relationship in nature? Try to think of an association of interest in psychological research that equals an absolute 0, a complete and utter absence of association. With any realistic relationship in psychology, one should expect at least some effect or difference, however minor (Rothman, 1990). A Type I error is, therefore, a little like a unicorn. But, even if a

truly null relationship existed, finding another is unlikely, and both being tested simultaneously in the same set or family is downright nearly impossible. Because almost no relationships or effects are ever completely null, researchers need not worry about the multiplicity problem (Gelman et al., 2012).

2. Define “set”: How should we define a set or family of hypotheses? Is it all hypotheses within a study, an article, by the same researcher, or within the same journal issue, publisher, or discipline? The answers to these questions are subjective and can lead to inconsistent or arbitrary decisions. Imagine a researcher who successfully rejects the null hypothesis of Relationship A. In a parallel universe, another version of the same researcher also tests Relationships B and C using the same data. If that researcher were to follow the familywise approach, they would adjust the nominal significance threshold at the test level for each of the three hypothesis tests. Although the p -value for Relationship A remains the same, controlling for multiplicity means it might no longer meet the adjusted significance threshold. Thus, the same result leads to different conclusions. “Why should a research finding be punished because of the company it keeps?” (Hancock & Klockars, 1996, p. 272). Research findings should not be contingent on the number of associations explored (Cribbie, 2017). To maintain consistency across studies, the natural unit of analysis should be a single test, hypothesis, or relationship (Rothman, 1990; Saville, 1990; Wilson, 1962).
3. Let p be the sidekick: In most research contexts, the emphasis should be on obtaining precise parameter estimates rather than on achieving statistical significance (Saville, 1990) or merely determining whether an effect exists. It is

more informative, productive, and useful to evaluate its direction, magnitude, and precision (Cumming, 2014; Cumming & Calin-Jageman, 2016). Indeed, results from null hypothesis significance tests can offer evidence against the null, but a genuine, insightful understanding of the association in question comes from effect sizes, confidence intervals, replications, and meta-analyses (Cribbie, 2017; Cumming, 2014; Gelman et al., 2012). As Tukey (1977) beautifully put it, “black or white is dangerous, misleading, generally unsatisfactory” (p. 101) in the context of hypothesis testing. Thus, keep the p -value the Robin to the effect size’s Batman; by making hypothesis testing the supplementary rather than the primary tool for drawing conclusions, Type I errors (and consequently multiplicity) become less critical.

Not to Adjust: Practical Considerations and Issues with Multiplicity Control Adjustments

Although multiplicity control procedures like the ones discussed shortly effectively reduce the probability of false positives, nothing comes without a cost. Curbing the Type I error rate with these methods necessarily involves a trade-off with other crucial aspects of statistical inference. The most obvious trade-off is the loss of statistical power that accompanies these adjustments, particularly with methods like the Bonferroni correction (Aickin & Gensler, 1996). When stricter thresholds for significance are imposed, we risk overlooking true and meaningful effects, especially in studies with small effects, limited sample size, or high variability. By decreasing the probability of a Type I error, we inadvertently increase the probability of a Type II error, where real discoveries are potentially missed because the bar for significance is set higher. Indeed, there are situations where a Type I error is more costly than a Type II error, making multiplicity control adjustments a prudent choice. However, this is not always the case. The

trade-off is ultimately at the discretion of the researcher, and the decision regarding which error is more dangerous must be carefully considered alongside how to handle the Type I error rate, if at all.

Another critical issue is the complexity of implementation in advanced statistical analyses. Although applying the Bonferroni or Benjamini-Hochberg procedures to a series of t tests or correlations is straightforward and can be easily executed in most statistical software, the same cannot be said for more complex methods like structural equation modeling (SEM). These sophisticated statistical frameworks, which are increasingly prevalent in psychological research, introduce additional layers of complexity that might complicate both theory and application of multiplicity control procedures and increase the risk of misinformed decisions. If researchers, including those well-versed in the multiplicity problem, struggle to reach a consensus about whether, when, and how to invoke control even with simpler analyses, what will happen with more sophisticated techniques?

For instance, in SEM, controlling for multiple tests might be a challenging task due to the interdependencies between paths and the overall model structure (Cribbie, 2007; Smith & Cribbie, 2013). These complexities make crucial decisions (e.g., what constitutes a family or set of hypotheses?) even more burdensome. But, even if clear and universally accepted decisions were available, the practical implementation of multiple testing corrections within advanced modeling techniques is not commonly supported by standard statistical software packages. Researchers are often left to devise their own solutions or rely on more advanced statistical tools that require a deeper understanding of underlying algorithms or mechanics. Custom scripts or specialized software capabilities may be helpful, but these are not always readily available or user-friendly.

To Adjust: Popular Multiplicity Control Approaches and Procedures

Whether these arguments justify abandoning multiplicity control altogether is open to interpretation. Still, there is no debate that the chance of at least one Type I error increases with more hypothesis tests. This argument alone is valid and sufficiently convincing for many to invoke some form of control, if not in practice, at least in theory (e.g., by considering it or expressing intention; see Martinez Gutierrez et al., 2025). And, given the vast literature on multiplicity control, most researchers find a familywise approach more suitable, both theoretically (Hancock & Klockars, 1996) and practically (Martinez Gutierrez et al., 2025).

Familywise Error (FWE) Rate

One solution to the exponential rise in the risk of at least one Type I error is to focus on the overall Type I error rate, where “overall” refers to the set, or *family*, of tests over which we are concerned about error inflation (Cribbie, 2017). If we set our threshold of an acceptable chance of a false-positive (i.e., nominal Type I error rate; e.g., $\alpha = 0.05$) as the overall, or familywise (FWE) rate, and split it across the number of tests in the family, we can effectively curb the rising risk of at least one Type I error, thereby compensating for multiple, simultaneous tests. The most frequently used FWE control in psychological research is the Bonferroni method (Martinez Gutierrez et al., 2025).

The Bonferroni method comes from a probabilistic proof by Bonferroni in 1937, but was used in the context of multiplicity control procedures by Dunn in 1961. The Bonferroni method divides the FWE Type I error rate equally across the number of tests (T) such that the significance level for any given test in the family (α_t) is $\alpha_t = \alpha/T$. The Bonferroni method is a

simultaneous multiple-testing procedure, also known as a one-step procedure, because it tests all hypotheses at once using the same significance level, $\alpha_t = \alpha/T$ (Cribbie, 2017).

False Discovery Rate (FDR)

Formally, the false discovery rate (FDR) is defined as the expected ratio of incorrectly rejected hypotheses to the total number of rejected hypotheses. Of course, one can never know how many hypotheses are correctly and incorrectly rejected. The FWE approach discussed above splits the overall Type I error rate across an entire set of hypotheses regardless of their results. FDR, on the other hand, acknowledges that a Type I error cannot occur with a statistically nonsignificant test. That is, we cannot incorrectly reject the null hypothesis if the null hypothesis is not rejected in the first place. FDR is, thus, a slightly different approach to dealing with the multiplicity problem. FDR controls the overall Type I error rate, but only across the statistically significant tests instead of the entire set of tests (Benjamini & Hochberg, 1995; Keselman et al., 1999). Because this approach skips over the non-rejected hypothesis tests, FDR-based procedures provide more liberal adjustment, and thus greater statistical power (Cribbie, 2003; Cribbie, 2017). Among the FDR-based procedures, the Benjamini-Hochberg (1995) is the most used in psychological research (Martinez Gutierrez et al., 2025).

Unlike the Bonferroni correction, which adjusts p values in a single step and maintains a uniform significance threshold across all tests, the Benjamini-Hochberg procedure employs a stepwise, specifically step-up, approach. In this method, the p values from multiple tests are first ranked in ascending order from smallest to largest. Each p value is then compared to a critical value which increases according to its rank (i). The critical value is calculated using $\alpha i/T$, and α is the desired FDR. Once the largest p value that satisfies the condition $p_i \leq \alpha i/T$ is identified, all

hypotheses corresponding to p values up to this threshold are declared significant, whereas all the hypotheses beyond this step are declared insignificant.

Controlling Without Adjusting: Introducing a Multilevel Modeling (MLM) Approach to Multiplicity Control

Popular multiplicity adjustment procedures like FWE and FDR have been used for decades to deal with multiple testing (i.e., control Type I error inflation). These procedures are simple, after-the-fact corrections applied to the null hypothesis test statistics, p values, or confidence intervals. The use of random-effects techniques has been recently suggested to handle a large number of comparisons by way of *modeling* differences in social epidemiology research (Bell et al., 2019; Evans et al., 2018; Jones et al., 2016; Merlo, 2019). This approach grew out of the need to address the limitations of conventional fixed-effects models, particularly when dealing with complex, high-dimensional interactions and testing numerous related hypotheses simultaneously. These limitations include challenges with interpretability, model parsimony, scalability, small subgroup sample sizes, redundant testing of interactions (i.e., overlapping terms), and, of course, Type I error inflation (Evans et al., 2018; Jones et al., 2016). Such issues frequently arise in the context of quantitative intersectionality analysis.

Intersectionality emphasizes that social factors such as race, gender, and class are not isolated but intersect in ways that create unique identities that cannot be explained simply by summing up the individuals' social categories; individuals typically belong to multiple subgroups simultaneously, and the interaction between these social characteristics can magnify their effects. For instance, the compounded disadvantages faced by individuals who are both Black and have low income are greater than the mere sum of each single group's impact (Bell et al., 2019). Yet, these effects are rarely represented with interaction terms because these terms would saturate a

model. For example, a model incorporating the effects of SES, sex, age, and ethnicity/race along with all possible interactions can quickly become unwieldy and difficult to interpret, especially when dealing with covariates with multiple categories, which is often the case with such variables (in a model with SES, sex, age and ethnicity, there would be 10 omnibus two-way interactions alone). Although dealing with two-way interactions is usually manageable and interpretable, adding three-way or higher-order interactions increases complexity to the point where models become difficult to untangle and comprehend. Furthermore, “many-way” interactions can be impractical and suffer from a lack of statistical power, often referred to as the “curse of dimensionality,” which results from the limited number of observations within each potential combination of variables (Bell et al., 2019; Green et al., 2017).

Researchers can address these challenges by using multilevel modeling, where social identities are treated as a random effect. In this framework, membership in a particular social group—or a combination of multiple groups (e.g., a middle-aged European male of moderate SES)—is considered a subset, or a sample from a population of infinite covariate combinations. This approach models individuals (Level 1 units) within their intersectional positions or social identities (Level 2 units). Although many psychological applications emphasize fixed effects (i.e., estimates of mean effects across groups), multilevel models also allow researchers to extract cluster-specific (i.e., social identity or subgroup) estimated effects when modeling between-group variation is theoretically or statistically warranted. In intersectional research, investigators may be interested not only in overall mean effects but also in whether specific identity groups differ meaningfully in baseline levels of an outcome or in the strength of relationships between predictors and outcomes. This approach enables the partitioning of variance across social strata, providing a more nuanced examination of both between- and within-group differences. As Bell et

al. (2019) highlight, this method helps address two key questions. First, using the intraclass correlation (ICC), researchers can assess the extent to which total variation in the outcome variable is attributable to intersectionality. Second, it allows for understanding how different social groups differ initially on the outcome (i.e., cluster intercepts) and in the relationships of interest (i.e., cluster slopes).

The same approach can be applied to additional scenarios that are common in psychological research. One such scenario involves evaluating a treatment across multiple sites, where MLM can effectively model numerous simultaneous differences and tests (as demonstrated in Gelman et al., 2012). For instance, consider testing a new therapy technique across a network of psychotherapy clinics, distributed across various provinces and territories. Given the potential variations in clinic characteristics—such as client and therapist demographics, language differences, or geographic contexts—it could be essential to move beyond assessing only the overall treatment effect. Hence, we are particularly interested in the difference in the effectiveness of the new and current gold standard treatments at each clinic on clients' mental health. To evaluate the difference in effectiveness between the two treatments, the new therapeutic approach is compared to a widely used method, such as cognitive behavioural therapy (CBT). Patients arriving at each clinic are randomized into one of two groups: those receiving the new therapy and those undergoing the familiar treatment, which serves as the control. Given the nested structure of the data (i.e., patients within clinics) and our primary interest in estimating the treatment effects specific to each clinic rather than an overall, mean effect, researchers have traditionally resorted to conducting separate mean-difference comparisons for each site (e.g., Petry et al., 2012, Sandell et al., 2015, Wagstaff et al., 2019).

The multi-site trial illustration was selected to align with the working example discussed in Gelman et al. (2012). However, the same modeling approach can be applied in contexts where the joint hypothesis is more plausibly the primary criterion for decision-making. For instance, in an educational setting, a policymaker might wish to implement a new program teaching students how to ethically use AI tools across a school district only if the program demonstrates a positive effect in at least 90% of participating schools. Similarly, in clinical research, psychedelic-based PTSD therapy practitioners might seek regulatory approval for general use only if the new treatment is shown to be effective in all predefined patient subgroups (e.g., combinations of age, sex, and comorbidity status). In both cases, the substantive research question hinges on rejecting the joint null hypothesis across multiple tests, making it a natural setting for examining the advantages of MLM over traditional multiplicity-control methods.

This approach assumes that patient outcomes within a clinic may be more similar to one another than to those from other clinics (e.g., due to contextual factors such as demographics and therapists' styles). This approach can be represented as a simple two-mean comparison (e.g., using a t test) or as a regression model:

$$y_{ij} = b_{0j} + b_{1j}X_{1j} + e_{ij} \quad (1.1)$$

where y_{ij} is the mental health measurement score for client i in clinic j , b_{0j} is the mean mental health score for the control group in clinic j , b_{1j} represents the effect of treatment in clinic j , X_1 is an indicator for treatment status (i.e., 0 for control, 1 for new therapy), and e_{ij} is the difference between client i 's mental health score and their treatment group's mean in clinic j , which is distributed as $e_{ij} \sim N(0, \sigma^2)$.

To evaluate whether the new treatment effect is statistically significant, we can test the null hypothesis of no effect, $H_{0j}: b_{1j} = 0$, against an alternative hypothesis, $H_{1j}: b_{1j} \neq 0$ for each of the J clinics. Assuming a nominal Type I error rate of $\alpha = .05$, each of these tests has a 5% chance of incorrectly rejecting the null hypothesis when, in fact, there is no difference between the new and current gold standard treatment. However, conducting J separate significance tests raises concerns about inflating the overall risk of making a false conclusion. The same issue arises if we wish to examine whether the treatment effects significantly differ between clinics or from the overall treatment effect.

Instead, researchers can extend the model introduced in Equation 1.1 into a random-slopes MLM framework, treating clients as Level 1 units and clinics as Level 2:

$$\text{Level 1: } y_{ij} = b_{0j} + b_{1j}X_{1j} + e_{ij} \quad (1.2)$$

$$\text{Level 2a: } b_{0j} = \gamma_{00} + \delta_{0j}$$

$$\text{Level 2b: } b_{1j} = \gamma_{10} + \delta_{1j}$$

where y_{ij} , b_{0j} , b_{1j} , X_{1j} , and e_{ij} are defined as in Equation 1.1. Here, γ_{00} represents the grand mean, which is the mean mental health score of the control group across all clinics, and δ_{0j} is the deviation of clinic j 's control group mean from the grand mean, distributed as $\delta_{0j} \sim N(0, \tau_{00})$. Next, γ_{10} is the overall treatment effect, representing the mean difference in mental health score between the treatment and control groups across all clinics, and δ_{1j} is the deviation of clinic j 's treatment effect from the overall effect, distributed as $N(0, \tau_{11})$. Further, the random intercept and slope are allowed to covary, $\text{Cov}(\delta_{0j}, \delta_{1j}) = \tau_{01}$, such that

$$\begin{pmatrix} \delta_{0j} \\ \delta_{1j} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, T \right), \quad (1.3)$$

where T is the variance-covariance matrix of the random effects,

$$T = \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix}. \quad (1.4)$$

We can then merge the Level 1 and Level 2 equations into a single reduced-form equation such that:

$$y_{ij} = \gamma_{00} + \delta_{0j} + (\gamma_{10} + \delta_{1j})X_{1j} + e_{ij}, \quad (1.5)$$

which can be simplified to:

$$y_{ij} = \mu_{00} + \delta_{0j} + \mu_{10}X_{1j} + \delta_{1j}X_{1j} + e_{ij} = \quad (1.6)$$

$$y_{ij} = (\mu_{00} + \mu_{10}X_{1j}) + (\delta_{0j} + \delta_{1j}X_{1j} + e_{ij}),$$

where the terms between the first pair of brackets are the fixed effects and, in the second, the random effects.

Researchers typically estimate and interpret the overall, fixed effects—i.e., μ_{00} and μ_{10} —and the random effect variance parameters, τ_{00} , τ_{11} , and σ^2 . To evaluate the effectiveness of the new treatment in each clinic, however, we would need to estimate the cluster-specific slopes (b_{1j}) and test whether each is statistically different from zero.

Although the cluster-specific slope (b_{1j}) in Equation 1.2 is denoted the same as the treatment effect parameter for clinic j in Equation 1.1, the estimation approach and results differ under MLM. If we had used the traditional approach, as in Equation 1.1, where J separate regression models are estimated via ordinary least squares (OLS), each clinic's treatment effect would be estimated independently, while ignoring the other b_1 estimates from the other $J-1$ clinics; in other words, each regression model is estimated in isolation from its sibling models. In

this case, b_{1j} estimates are unaware and unaffected by other b_{1j} estimates and thus no information is shared or “pooled” across clinics. The opposite extreme to this “no pooling” approach would be complete pooling, where information from all clinics is combined into a single model, ignoring any distinction between clinics. Rather than estimating J separate models with, say, n observations each, this approach treats all data as one set of $J \times n$ observations. MLM strikes a balance between these extremes, allowing for partial pooling of information across clinics.

Estimating Cluster-Specific Effects Through Partial Pooling

In MLM, random cluster effects (e.g., clinic-specific effects) are not directly estimated; instead, they are predicted because they are treated as latent variables rather than fixed statistical parameters (Snijders & Bosker, 2012). In linear mixed models, cluster-specific effects are typically obtained using Best Linear Unbiased Prediction (BLUP); when the variance components are estimated from the data rather than treated as known or assumed, these predictions are more precisely referred to as empirical BLUPs (EBLUPs). Because variance components are rarely known in practice, most reported cluster-specific effects are EBLUPs, obtained by substituting estimated variance components into the BLUP formula.

Best Linear Unbiased Prediction (BLUP) and Empirical BLUP (EBLUP)

In traditional Bayesian analysis, a prior distribution is specified based on prior beliefs about the parameters. By combining this prior with observed data, we derive the posterior distribution—a probability distribution that reflects the updated inference given the model, data, and prior. With EBLUP, we do not specify a prior in the Bayesian sense. Instead, we assume a random-effects distribution, and estimate variance components directly from the data (by relying

on maximum likelihood methods). Although this derivation resembles Bayesian updating under conjugate normal assumptions, it is obtained within a frequentist framework.

For example, we might assume that the treatment effect across clinics follows a normal distribution and plug the estimates of their multivariate means and variance components. Thus, each of the b_{1j} EBLUP predictions in Equation 1.2 is defined as the conditional expected value given the observations in clinic j and the estimated model parameters, assuming that the b_{1j} s are normally distributed (Snijders & Berkhof, 2008; Snijders & Bosker, 2012). Had we ignored all other clinics and focused only on clinic j , b_{1j} would be estimated by the clinic's mean difference in treatments ($\hat{b}_{1j} = \bar{y}_{j_treat} - \bar{y}_{j_control}$), which is also the OLS estimate (no pooling). Alternatively, if we relied solely on the fixed effects, our best estimate would be $\hat{\gamma}_{10}$ (complete pooling). Another option is to integrate the information from clinic j with the fixed effect. The result is *partial pooling* or *shrinkage*, a compromise between these two extremes, whereby cluster j 's intercepts and slope are EBLUP-predicted by combining both OLS-estimated vectors $\hat{b}_j = \begin{pmatrix} \hat{b}_{0j} \\ \hat{b}_{1j} \end{pmatrix}$ and population estimate $\hat{\gamma} = \begin{pmatrix} \hat{\gamma}_{00} \\ \hat{\gamma}_{11} \end{pmatrix}$ with weights proportional to their respective estimated precision matrices, $\frac{1}{\sigma^2} X_j' X_j$ and T^{-1} :

$$\begin{pmatrix} \tilde{b}_{0j} \\ \tilde{b}_{1j} \end{pmatrix}^{\text{EBLUP}} = \lambda_j \hat{b}_j + (I - \lambda_j) \hat{\gamma} \quad (1.7)$$

Where I is the 2x2 identity matrix, and λ_j is the 2x2 weight matrix, which is also considered the reliability of clinic j 's OLS estimates, is defined as:

$$\lambda_j = \left(\frac{1}{\sigma^2} X_j' X_j + T^{-1} \right)^{-1} \frac{1}{\sigma^2} X_j' X_j, \quad (1.8)$$

such that, again, T is the variance-covariance matrix of the random effects.

The expression in Equation 1.7 represents the BLUP (or EBLUP when variance components are estimated), which under Gaussian assumptions coincides with the conditional mean of the random effects given the data. The slope component of the expression in Equation 1.7, $\tilde{b}_{1j}$, can also be interpreted as the adjusted or shrunken clinic-specific coefficient. The degree of shrinkage, or influence, acting on the estimate partly depends on the sample size in clinic j . In larger clinics, the EBLUP prediction approaches the OLS estimate.

When clinic-specific effects are treated as fixed but unknown parameters, the within-clinic OLS estimator is unbiased. In contrast, the EBLUP shrinks estimates toward the population mean and is therefore considered biased for the realized clinic-specific parameter. However, under the hierarchical model in which clinic effects are assumed to be random draws from a population distribution, the EBLUP coincides with the conditional expectation of the random effect given the data. Importantly, under the hierarchical model, the EBLUP minimizes mean squared prediction error among linear unbiased predictors. As a result, when predicting random clinic effects, it typically achieves lower mean squared error than the within-clinic OLS estimator, despite introducing shrinkage toward the population mean (Gelman et al., 2012; Snijders & Bosker, 2012).

Precision of the Cluster-Specific Predictions

The precision of each cluster-specific estimate, whether derived via EBLUP, can be quantified by its standard error (SE), reflecting the degree of uncertainty around the estimated effect for the specific cluster. Snijders and Bosker (2012) describe the comparative SE, which isolates the uncertainty of the prediction for a given cluster j without incorporating variability in the fixed effect. The SE comparative calculation comes from the mixed model theory (and

equivalently Bayesian updating), which includes the clusters' coefficients as the conjugate prior, $b_j \sim N(\gamma, T)$, and OLS estimates as the likelihood, $\hat{b}_j \sim N(b_j, \sigma^2(X_j'X_j)^{-1})$. Combining both Gaussian distributions' inverse variances gives the EBLUP's precision. The square root of the (2,2) element of this matrix gives the comparative standard error for the slope in cluster j , expressed as¹:

$$SE_{comparative} = SE(\tilde{b}_{1j}) = \sqrt{\left[\left(\frac{1}{\sigma^2} X_j'X_j + T^{-1} \right)^{-1} \right]_{22}}. \quad (1.9)$$

The comparative SE captures how much the estimates for clinic j would vary if we were to repeatedly draw similar samples. As the cluster sample size n_j increases, the comparative SE decreases, reflecting greater reliance on the specific data for each cluster.

Constructing Uncertainty Intervals for the Cluster-Specific Predictions

To evaluate the likely range of values for each cluster-specific effect, we can construct an approximate 95% Wald-type confidence interval (CI),² for example, around the EBLUP estimate. Using the comparative SE, the CI can be calculated as:

$$b_{1j} \pm 1.96 \times SE_{comparative} \quad (1.10)$$

This interval reflects uncertainty about the predicted random effect for cluster j under the hierarchical model. It quantifies how much the prediction would vary across repeated samples from the same population model.

¹ Because these are two independent estimates of the same parameters, we cannot add variances, but must invert, add precisions instead, and inverse back to variance.

² Strictly speaking, for predicted random effects this interval is a Wald-type interval based on the conditional variance of the predictor rather than a confidence interval for a fixed parameter. We retain the term “confidence interval” for consistency with usage throughout the dissertation.

For smaller clusters, where estimates shrink closer to the overall mean due to limited data, the CI tends to be wider, reflecting greater uncertainty around the estimate. In contrast, for larger clusters with more data, estimates rely more on clinic-specific information and less on the overall mean, leading to narrower CIs and greater precision.

Determining Statistical Significance for Cluster-Specific Effects

Determining statistical significance for cluster-specific effects in MLM requires careful consideration of the model's hierarchical structure and the inherent variability within and between clusters. While the precision of cluster-specific predictions can be evaluated through SEs or CIs, significance testing involves assessing whether the effect of interest is distinguishable from zero (or another specified null value).

Hypothesis Testing

In a frequentist context, the significance of a cluster-specific effect, b_{1j} , can be tested by constructing a null hypothesis such as $H_0: b_{1j} = 0$. This involves calculating a z statistic:

$$z = \frac{\tilde{b}_{1j}}{SE_{comparative}} \quad (1.11)$$

This z score is then compared against critical values from the standard normal distribution (e.g., ± 1.96 for a two-tailed test at $\alpha = .05$) to determine significance. Alternatively, a p value can be calculated, reflecting the probability of observing a z score as extreme as or more extreme than the observed value under the null hypothesis. If $p \leq \alpha$, the null hypothesis is rejected, suggesting that the cluster-specific effect is statistically significant.

Shrinkage and Its Implications for Significance Testing

The partial pooling inherent in EBLUP predictions introduces shrinkage, systematically pulling cluster-specific estimates toward the population mean. This effect is more pronounced for clusters with greater deviations from the population mean, smaller sample sizes, or higher within-cluster variability, as their estimates rely more heavily on population-level information. Consequently, if the true population-level effect is null (i.e., $\gamma_{10} = 0$), the estimates will shrink toward zero, closely aligning with the population mean. Pulling all cluster-specific predictions closer to 0 would affect the number of statistically significant effects. Specifically, having less extreme estimates is expected to reduce the probability of identifying statistical significance, and by extension, the probability of a Type I error. In the model we discuss here, and in the simulation that follows, the population mean treatment effect is represented by γ_{10} , and because this design assigns an equal proportion of treated individuals in every cluster ($p_j = .5$), γ_{10} corresponds to the pure within-cluster treatment effect. In designs where treatment proportions vary across clusters, γ_{10} would reflect a blend of the within- and between-cluster treatment effects (what Hoffman and Walters, 2022, call “smushed effects”), and shrinkage would then pull toward this blended population mean rather than necessarily toward zero. In more realistic situations where treatment proportions and/or cluster sizes vary, the within- and between-cluster effects can be “teased apart” by cluster-mean (often referred to as group-mean) centring the predictor and including a variable consisting of the cluster means as an additional covariate at Level 2 in the model (Hoffman & Walters, 2022).

Alternatively, if the true population-level effect is non-zero (e.g., $\gamma_{10} = 5$), shrinkage pulls cluster-specific estimates, again, toward the fixed effect; however, the centre of gravity will no longer be zero but rather around that non-zero value representing the population mean (e.g.,

$\hat{\gamma}_{10} = 4.91$). In our model, this fixed effect is γ_{10} , which equals the within-cluster treatment effect in our balanced design because the proportion of treated individuals is constant across clusters. In this case, partial pooling is expected to increase the probability of identifying statistically significant results, thereby enhancing statistical power. However, because the population-level effect represents a weighted average of all clusters' effects, shrinkage can introduce bias in cluster-specific estimates in cases where some clusters have truly null effects while others have non-null effects. In such cases, shrinkage toward the weighted average may distort the individual estimates.

It is important to note that partial pooling affects both the numerator of the test statistic (the cluster-specific estimate) and its denominator (the comparative SE). Partial pooling leads to lower comparative SE, on average, relative to the unpooled SE because the estimates integrate information from other clusters. Lower SEs can partially offset the shrunk cluster-specific estimate when the mean population effect is null, but they can also strengthen the ability to correctly identify statistically significant results when the mean population effect is not null.

Prior Research on Multilevel Modeling and the Multiplicity Problem

The potential of MLM to address the multiplicity problem has been previously proposed in a small but growing body of research. These studies demonstrate MLM's promise, but its underlying mechanisms, statistical properties, and applicability to psychological research contexts remain largely unknown.

Gelman et al. (2012) proposed Bayesian MLM as a solution to the multiplicity problem, emphasizing its ability to handle multiple comparisons through partial pooling. Their theoretical framework highlighted how borrowing strength across clusters benefits smaller or highly

variable groups by shrinking estimates toward the population mean. However, Gelman et al. did not empirically evaluate their claims through rigorous simulation studies, leaving critical questions unanswered. Most notably, they conflated the Bayesian aspects of their approach with the hierarchical structure of MLM itself, leaving it unclear whether reduced error rates stemmed from Bayesian priors, the partial pooling mechanism naturally occurring in cluster-specific estimates within MLM, or their combination. Untangling this relationship is crucial to understanding the true source of MLM's advantages and applying them in practical settings.

Bell et al. (2019) conducted one of the most relevant studies to the present research, empirically evaluating how non-Bayesian MLM addresses multiplicity. Their simulation study assessed Type I error rates in random-effects models under various conditions, demonstrating that MLM can effectively control error inflation in social epidemiological contexts. However, several limitations diminish its applicability to psychological research. First, because Bell et al. did not explicitly explain the mechanisms underlying MLM's ability to address multiplicity, it is unclear exactly how cluster-specific estimates were obtained or how statistical significance was determined, which are central to understanding how partial pooling impacts error control. Second, their reliance on MCMC methods for estimation, while appropriate in both Bayesian and frequentist settings, diverges from the typical frequentist MLM practices commonly used in psychological research. This methodological choice, although understandable (for successful estimation), limits the study's relevance for psychology, where maximum likelihood-based methods are almost exclusively used. Third, their simulation scope was relatively narrow, examining only a few conditions affecting Type I error rates without systematically investigating factors like cluster size, number of clusters, or within-cluster variability. Finally, Bell et al. did not evaluate statistical power, a critical component in assessing any solution to the multiplicity

problem. Traditional multiplicity control methods, like FWE and FDR controlling procedures, often sacrifice power to control error rates. Thus, a comprehensive evaluation of MLM requires assessing its performance across both dimensions, which is lacking in most studies.

Other studies, including Evans et al. (2018), Jones et al. (2016), and Merlo et al. (2018) have applied MLM in contexts where the multiplicity problem arises but did not directly evaluate its statistical properties for addressing this issue. Evans et al. and Jones et al. used MLM to model complex, high-dimensional interactions in intersectionality research, demonstrating its capacity to account for variability at multiple levels. Similarly, Merlo et al.'s multilevel analysis of individual heterogeneity and discrimination (MAIHDA) emphasized MLM's ability to reduce bias in subgroup analyses. However, these studies primarily focused on applications and did not empirically assess Type I error control, statistical power, or comparisons to traditional multiplicity control methods like FWE and FDR. Consequently, while they showcase MLM's flexibility and utility, they do not evaluate its potential as a solution to the multiplicity problem.

Current Research

Collectively, these studies highlight MLM's potential to address the challenges posed by multiple testing but underscore the need for a more rigorous evaluation. Key issues include:

1. **Mechanistic Understanding:** Prior research has not disentangled the effects of partial pooling from Bayesian influences, leaving the true source of error rate control unexamined at large.
2. **Comprehensive Simulation Studies:** Existing simulations have had a limited scope, often neglecting critical factors like cluster size, number of clusters, and variability, which are highly relevant to psychological research scenarios.

3. **Power Evaluation:** Few, if any, studies have assessed power alongside Type I error rates, leaving the trade-offs between error control and statistical power underexplored.
4. **Relevance to Psychology:** Studies like Bell et al. (2019) focus on social epidemiology or intersectionality contexts, which, while important, differ from most psychological research settings. Applications in psychology often involve testing interventions across sites (e.g., clinics, schools) or participant subgroups such as those arising from intersectionality, where realistic simulations tailored to these contexts are crucial.

The present study addresses these issues by systematically evaluating how MLM handles multiplicity in more realistic psychological research scenarios. By conducting a Monte Carlo simulation with a wide range of conditions, it assesses both Type I error rates and power, offering a comprehensive view of MLM's performance compared to traditional methods like FWE and FDR. Additionally, this study disentangles the effects of partial pooling from Bayesian methods, advancing the theoretical and practical understanding of MLM as a viable solution to the multiplicity problem. These contributions hold meaningful implications for improving methodological rigour and statistical practices in psychological research.

Method

Monte Carlo Simulation Study

Objective

The objective of this Monte Carlo simulation study is to evaluate the extent to which MLM addresses the multiplicity problem. Specifically, the study aims to provide empirical

evidence on MLM's performance with respect to Type I error rates and statistical power, comparing it to traditional adjustment techniques such as FWE and FDR controls. The simulation is designed to reflect common conditions encountered by psychology researchers, ensuring its relevance and applicability to real-world scenarios.

Design

The simulation design follows a previously introduced example involving the evaluation of a new therapy technique across a network of psychotherapy clinics. Within each clinic, clients were randomly assigned to either the new therapy (treatment) or an established therapy (control). The design systematically varied key factors, including within-cluster sample size, number of clusters, and within-cluster variability.

To examine scenarios with varying degrees of effects, the proportion of clusters with truly null effects was varied from 1.00 (all clusters null, representing Type I error conditions) to 0.00 (no clusters null, representing power conditions) in increments of 0.25. In conditions where the proportion of null-effect clusters was less than 1.00, the true effect in non-null clusters was also varied to assess its influence on Type I error and power rates; the effect in a non-null cluster could potentially influence rejection rates in null clusters due to shrinkage toward the mean effect.

The simulation design included conditions where all clusters had equal total sample sizes and group sizes (i.e., treatment vs. control). For each cluster, statistical significance test results were evaluated for correctness by assessing whether the null hypothesis was correctly rejected under power conditions or incorrectly rejected under error conditions. The levels of the manipulated factors are presented in [Table 1.1](#).

Procedure

Data were generated using customized R functions (R Core Team, 2024) that simulated a two-level structure, with n_j individuals nested within clusters. A binary predictor variable was used to assign group membership ($MT = 0$ for control, and $MT = 1$ for treatment) within each cluster. The outcome variable (mental health) was generated from a normal distribution with a cluster-specific effect (b_j ; the mean difference between the treatment and control groups), an intercept, and residual variance σ^2 . We used a bottom-up approach where each of the J clusters' data were independently simulated and then combined into a single dataset. This approach allowed us to explicitly generate each cluster's data from a known true effect—set to zero for null clusters (i.e., $b_j = 0$) and non-zero for nonnull clusters (i.e., $b_j = 1, 2$, or 5), based on the manipulated proportion of null clusters (π)—with individual observations within each cluster scattered around that value as a function of residual variance σ^2 , so cluster-specific treatment estimates did not correspond perfectly to the true cluster effect.

Data were analyzed using the *lme4* (Version 1.1-36; Bates et al., 2015) and *mixedup* (Version 0.4.0; Clark, 2024) packages. The Monte Carlo simulation code was organized using the *SimDesign* package (Version 2.18; Chalmers, 2016; Sigal & Chalmers, 2016). The working model specified in each iteration is the one presented in Equation 1.2 and follows the *lme4* syntax of:

```
lmer(ment_health ~ MT + (1 + MT | cluster), data = data).
```

Note that this specification includes the covariance between random effects, which allows the model to remain invariant to the coding of MT (here, $MT = 0/1$) by avoiding the coding-dependent variance constraints that arise when τ_{01} is fixed to zero. While centring MT (e.g.,

coding as $-0.5/0.5$) can improve interpretability of the intercept and may sometimes enhance numerical stability, it was not necessary here for the aims of the simulation study. Allowing τ_{01} to be freely estimated implies that the between-clinic variance may differ across treatment levels under 0/1 coding; imposing $\tau_{01} = 0$ (with appropriate centring; for example, under $-0.5/0.5$ coding) would instead constrain the between-clinic variances to be equal.

All models were estimated using restricted maximum likelihood (REML). The default optimizer in *lme4*, *nloptwrap*, was employed initially. This optimizer is an R wrapper function for the *NLOpt* optimization library (Johnson, 2024), which provides a variety of nonlinear optimization algorithms. The optimization process was terminated when all parameter estimates changed by less than 10^{-4} across successive iterations or when the maximum number of function evaluations, set to 10^5 , was reached to ensure adequate exploration of the likelihood surface and to reduce the risk of nonconvergence. Model convergence was assessed based on gradient checks and optimization diagnostics. If convergence issues were detected, alternative optimization algorithms were applied. Specifically, models were re-estimated using the Bound Optimization BY Quadratic Approximation (BOBYQA) or the Nelder-Mead algorithms to address potential estimation challenges.

Under certain conditions, particularly under the complete null where the true between-cluster variance is zero, it is reasonable that the estimated random-effects covariance matrix ($\hat{T}$) may approach singularity. In such cases, the model implies near-complete shrinkage of cluster-specific effects toward the population mean, corresponding to a degenerate covariance structure. However, this is not necessarily a modeling failure, but rather a boundary solution that is consistent with the data-generating process.

Because boundary solutions can produce numerically degenerate cluster-specific predictions, singular fits were handled algorithmically in the simulation code. When all cluster-specific slope predictions were numerically identical (indicating complete shrinkage and a degenerate solution), the replication was discarded and regenerated. When singularity warnings occurred but cluster-specific predictions remained distinct (even slightly, e.g., different only in the 3rd decimal place), the replication was retained. This handling rule was implemented to avoid degenerate solutions in which no between-cluster variability could be estimated. Consequently, estimates of familywise Type I error rate under complete null conditions are conditional on replications yielding estimable, non-degenerate random-effect predictions.

Evaluating the Performance of Testing Approaches

Although the multiplicity problem is primarily concerned with inflated Type I error rates, we evaluated both Type I error rates and statistical power to provide a more comprehensive assessment of each testing approach. These metrics are inherently connected and considering them side by side offers a clearer picture of each method's performance.

To evaluate Type I error, we measured the FWE rate, defined as the probability of at least one false positive within a set of tests. In our simulation, a “set” or “family” was defined as the group of clusters for each condition. As a result, the true null family size varied depending on the condition. Within each clinic (i.e., cluster), we conducted a two-group comparison (treatment vs. control) across all testing approaches (i.e., MLM, unadjusted tests, FWE-control, and FDR control). For each condition, we recorded the proportion of iterations in which at least one null-effect cluster yielded (incorrectly) a statistically significant result, representing the FWE rate for that condition.

To assess statistical power, we calculated average power (per-comparison power), defined as the mean probability of correctly rejecting false null hypotheses (i.e., those with a non-null population effect) across all clusters. For each iteration, we identified the proportion of non-null clusters where the test was statistically significant, then averaged these proportions across all iterations within each condition. This approach provides an overall measure of how sensitive each testing method is to detecting true effects, offering a balanced complement to the FWE results.

Results

Simulation results are detailed in [Figures 1.1](#), [1.2](#), and [1.3](#). For transparency and reproducibility, the simulation code and additional materials, including high-resolution figures, are accessible via the Open Science Framework (OSF): osf.io/gbx26.

Type I Error Rates

Complete-Null Conditions

Beginning with the condition where all clusters have true null effects ([Figure 1.1](#)), the Type I error rate estimates largely agreed with Gelman et al.'s (2012) hypothesis. The MLM approach maintained consistently low familywise Type I error rates across all sample size levels, with a mean of .12 for a smaller number of clusters (e.g., $J = 8$), which is higher than both FWE and FDR (roughly twice as high as α), but considerably lower than the unadjusted approach. For larger cluster sizes (e.g., $J = 256$), the familywise Type I error rate averaged around .017, which is noticeably lower than the rates demonstrated by FWE and FDR control procedures and substantially lower than the unadjusted approach.

Partial-Null Conditions

As the proportion of clusters with true null effects increases from 25% to 75% (top to bottom rows in [Figure 1.2](#)), we observe a slight increase in familywise Type I error rates for the MLM approach, reflected in flatter and slightly elevated error curves. When a larger proportion of clusters have null slopes, the overall heterogeneity in slopes across clusters decreases because more slopes aggregate tightly around zero rather than being spread across a wider range of values. As a result, the estimated between-cluster slope variance is lower, leading to a stronger shrinkage of clusters' slopes toward the overall, group-level mean (fixed effect). When that mean is non-zero—driven by the non-null clusters—this shrinkage can cause null clusters' estimated slopes to pull farther away from zero, thereby increasing the risk of falsely detecting an effect when none exists and contributing to faintly elevated Type I error rates. However, the differences in FWE between conditions with varying proportions of clusters with null effects are rather subtle.

Effect Size. Increased non-null effect size values (i.e., when at least 25% of clusters had non-null effect sizes) were associated with markedly lower familywise Type I error rates for the null clusters with the MLM approach. Specifically, moving from $b_{1j} = 1$ (leftmost column) through $b_{1j} = 2$ (the middle column) to $b_{1j} = 5$ (rightmost column) resulted in a progressively steeper initial decline and lower overall plateau in error rates. This pattern indicates that larger effects made it easier for this approach to avoid incorrect null rejections, as evidenced by the teal curves' rapid descent as sample size increases and stabilization at lower error values when the cluster size is large. One explanation for this finding is a higher estimated between-cluster variability when there is a mix of null and large effects; as the cluster-specific effects become more discrepant (e.g., $b_{1j} = 5$ and $b_{1j} = 0$ vs $b_{1j} = 1$ and $b_{1j} = 0$), the estimated variability is greater

and the pull of the true null effects toward the fixed effect is weaker. These findings suggest that the MLM approach is particularly well-suited for scenarios in which stronger underlying effects are present among other clusters with truly null effects.

Within-cluster Sample Size. For smaller non-null effects, $b_{1j} = 1$ and $b_{1j} = 2$, the MLM approach exhibits a fluctuating trend across smaller cluster-specific sample sizes. In particular, the familywise Type I error rates remain relatively low when the cluster-specific sample size is very small, then rise as sample size increases: at very small cluster sizes, even though there is a stronger pull towards the fixed effect, there are not enough observations to reject the null hypothesis, yielding wider CIs and thus lower false-positive rates. However, as the cluster-specific sample size increases, the CIs narrow, leading to temporary inflation in Type I errors. Eventually, at larger cluster sizes, the error rate stabilizes again, aligning more closely with general trends of improvement associated with increased sample size (except when the number of clusters is very large, which is discussed next). This trend suggests that, for smaller true effect sizes, partial gains in sample size can initially introduce greater uncertainty before providing enough evidence to reduce errors more consistently.

Number of Clusters. With a small number of clusters (e.g., $J = 8$ and $J = 16$), familywise Type I error rates are very low and close to FWE control, particularly among larger effects and sample sizes. As the number of clusters increases, Type I error rates rise: with a larger number of tests, there is a higher probability of incorrectly rejecting the null hypothesis at least once. Not only do error rates increase with more clusters, but the pattern of their decline across different within-cluster sample sizes also varies. With more clusters, the reduction in error rates as sample size increases is more gradual. Conversely, with fewer clusters, error rates start lower and decrease more sharply as the sample size goes up. In other words, although a larger within-

cluster sample size reduces Type I error rates, the effect is more pronounced with a smaller number of clusters. A larger number of clusters exhibits greater inertia, making it more challenging to reduce error rates quickly with an increased sample size within clusters.

Within-Cluster Variability. The degree of within-cluster variability had a substantial impact on familywise Type I error rates for the MLM approach. As shown in [Figure A1](#) in [Appendix A](#), lower values of within-cluster variability (e.g., $\sigma = 1$) indicated more rapid decreases in Type I error with increasing sample size, especially when effect sizes were moderate or large. In contrast, high within-cluster variability (e.g., $\sigma = 5$) showed inflated Type I error rates that remained elevated even at the largest sample size tested ($n = 500$). The familywise Type I error rates often exceeded those of the unadjusted approach, particularly when the proportion of null-effect clusters is low (e.g., 25%) and non-null effect is small. This pattern reflects the influence of σ on the degree of shrinkage, as shown in Formulas 3, 4, and 5 above; when σ is large, the reliability (λ) placed on the observed cluster-level slope decreases, leading to greater pull toward the fixed effect and reduced precision.

Power Rates

Power-Only Conditions

The power-only conditions evaluated all methods across the full set of manipulated variables under the scenario where all clusters exhibited a true non-zero effect (i.e., the proportion of clusters with null effects, $\pi = 0$). Results are presented in the top row of [Figure 1.3](#). The MLM approach consistently yielded the highest power rates across all levels of effect size, sample size, and number of clusters. This advantage is attributable to partial pooling, which allows the estimation of individual clusters' effects to “borrow strength” from others, stabilizing

estimates and increasing the chance of detecting true effects. Differences in power rates between MLM and the alternative methods were more noticeable for smaller effects (e.g., $b_{1j} = 1$ and $b_{1j} = 2$), where the probability of correctly rejecting the null hypothesis approached 1.00 less rapidly as sample size increased (so differences in power were more apparent). Although the number of clusters had a comparatively smaller influence on power than sample size or effect size, conditions with a greater number of clusters generally exhibited higher power.

Partial Null Conditions

Patterns similar to those observed in the power-only conditions emerged when only a subset of clusters exhibited true non-zero effects, as shown in the bottom three rows of [Figure 1.3](#). Once again, the MLM approach demonstrated the highest power rates compared to other methods across all levels of effect size, sample size, and number of clusters. However, as the proportion of clusters with true null effects increased from 25% to 75%, the MLM approach showed a subtle decline in power. This trend is illustrated by steeper power curves as a function of sample size under lower proportions of null clusters; that is, when fewer clusters have null effects (e.g., $\pi = 0.25$), statistical power increases more rapidly and reaches higher levels with smaller sample sizes compared to conditions with a greater proportion of null clusters (e.g., $\pi = 0.75$). The decline in power is likely driven by lower estimated between-cluster slope variance under higher proportions of null clusters, which induces stronger shrinkage of non-null cluster-level slopes toward the fixed effect (i.e., near 0). In other words, as the proportion of null clusters increases, the fixed effect estimate is pulled closer to zero, which in turn draws the non-null clusters' slopes away from their true values and reduces statistical power. However, any power rate disparities between the different proportions of null clusters are modest.

Effect Size. As expected, an increase in effect size resulted in substantially higher power rates across all approaches, with MLM approaching 100% power most rapidly. For example, when the effect size was $b_{1j} = 2$, the MLM method achieved approximately 97% power with just 8 clusters ($J = 8$) and a sample size of 50 per cluster, even when only 25% of the clusters had a true non-zero effect, compared to 75.6% power with the FWE control method. Power increased even further under conditions with a higher proportion of non-null clusters or larger effect sizes.

Within-cluster Sample Size. Similar to the expected relationship between effect size and power, increasing sample size led to higher power rates across all methods, with the MLM approach showing the most rapid gains. For smaller effect sizes (i.e., $b_{1j} = 1$ and $b_{1j} = 2$), MLM power increased steeply as sample size grew, whereas for larger effect sizes (i.e., $b_{1j} = 5$), the MLM approach reached perfect power even at relatively small cluster sizes (e.g., $n = 24$), regardless of the proportion of null clusters.

Number of Clusters. Increasing the number of clusters was associated with modest improvements in power for the MLM approach. The effect of cluster count was most noticeable for smaller effect sizes (i.e., $b_{1j} = 1$) and smaller proportions of null effects (i.e., $\pi = 0.25$), where having more clusters improved power. However, for moderate and large effect sizes, the benefit of additional clusters diminished, especially with larger sample sizes, as power approached ceiling levels across all conditions.

Within-Cluster Variability (σ). The degree of within-cluster variability had a notable effect on power for the MLM approach. As shown in [Figure A2](#) in [Appendix A](#), lower values of within-cluster variability (e.g., $\sigma = 1$) were associated with much more rapid rise in power as sample size increased with the MLM approach compared to other methods. In contrast, when within-cluster variability was high (i.e., $\sigma = 5$), the MLM approach showed slower power gains,

particularly in conditions with small non-null effects (i.e., $b_{1j} = 1$) and a high proportion of null clusters (75%).

Discussion

Researchers have long wrestled with the issue of inflated Type I error rates when conducting multiple significance tests. Despite decades of discussion, there remains very little agreement on whether, when, or how to implement control measures. In fact, the only consensus is that there is no consensus. Indeed, popular approaches like the FWE and FDR help control the risk of false positives, but they also reduce statistical power and are inconsistently applied (if at all) in psychological research (Martinez Gutierrez et al., 2025). Recent studies have proposed MLM as a model-based alternative that addresses multiplicity through partial pooling and shrinkage rather than applying post-hoc corrections. The present study systematically evaluated the extent to which MLM offers a viable solution to the multiplicity problem by balancing control of familywise Type I error and maintaining sufficient statistical power. To do so, we conducted a Monte Carlo simulation study to empirically compare MLM with common multiplicity control methods widely used in psychology.

Summary of Findings

The results offer a more nuanced picture than previous theoretical claims might suggest. Under complete null conditions (where all clusters had true null effects), MLM maintained low familywise Type I error rates across a wide range of sample sizes and cluster counts. These findings are broadly consistent with the idea that partial pooling pulls extreme estimates toward the fixed effect, thus reducing false positives. However, under partial null conditions, where only a subset of clusters had true null effects, the picture was more complex. Type I error rates were

influenced by a combination of factors, including within-cluster sample size, number of clusters, within-cluster variability, and the magnitude of non-null effects, all of which strongly and directly affect between-cluster variability. When the mean fixed effect was non-zero, but between-cluster variability was low, clusters with true null effects were pulled away from zero, leading to an inflation in false positives. Conversely, when between-cluster variability was higher (e.g., due to larger effect sizes of non-null clusters, large sample sizes, or lower within-cluster variability), this risk was mitigated and MLM continued to perform well.

In terms of power, MLM consistently outperformed the other methods, especially when effect sizes were small to moderate (i.e., when ceiling effects were not reached). Its ability to “borrow strength” across clusters yielded higher sensitivity in detecting true effects, and this advantage was most pronounced when there was an adequate sample size or sufficient between-cluster variability. Overall, MLM presents the most favourable trade-off between false positive control and the ability to detect true effects under conditions where all effects are null or when heterogeneity is high (e.g., clusters vary greatly in their effects), the number of clusters is small to moderate, and sample sizes within each cluster are large.

These findings also offer a useful first step toward evaluating a key claim made by Gelman and colleagues (2012) that partial pooling is the mechanism by which Bayesian hierarchical models reduce Type I error inflation in the presence of multiple comparisons. Although our simulation did not incorporate Bayesian methods, the use of frequentist MLM allowed us to isolate the performance of partial pooling on its own. The results suggest that, under complete-null conditions, partial pooling can indeed control familywise Type I error to a degree comparable to or even better than traditional methods like FWE or FDR correction. This lends support to Gelman et al.’s assertion that multiplicity may not require additional correction

when estimates are appropriately shrunk toward a mean effect. However, the results also show that this advantage is not universal. Under partial-null conditions, when the fixed effect is non-zero and between-cluster variability is limited, partial pooling can actually increase the likelihood of false positives in clusters with true null effects. This finding highlights that while partial pooling contributes meaningfully to error control, its effectiveness depends strongly on the underlying data structure. Fully disentangling the role of shrinkage from the influence of Bayesian priors remains an open question, but our findings provide initial evidence that partial pooling alone can serve as a limited, context-dependent remedy to the multiplicity problem.

Implications for Practice and Recommendations

For applied researchers, these findings suggest that MLM is not a blanket solution to the multiplicity problem, but it can be very effective under the right circumstances. When researchers are interested in estimating and testing multiple related effects (e.g., across sites, schools, or participant subgroups), the number of clusters is not too high (or the clusters are large), and there's reason to expect meaningful variability across these units, MLM offers a principled way to manage multiplicity without resorting to post-hoc corrections of significance tests, especially when cluster sample sizes are large. For example, say we are interested in how individuals experience divorce-related shame or stigma across a dozen or so countries or cultural groups that differ in levels of individualism and collectivism. In such cases, we would reasonably expect meaningful heterogeneity in effects, making MLM a suitable approach for estimating and testing those differences without being overly concerned about Type I error inflation. Similarly, if researchers suspect that all, or almost all, units have a near-null effect, again, MLM can be a good solution to combating the increasing probability of a false positive.

It is important to be aware of the conditions under which MLM may fall short. When clusters contain a mix of effects but overall between-cluster variability is low (e.g., most cluster estimates are close to 1 or 1.25), the fixed effect is still pulled away from zero by the non-null clusters. In these cases, the shrinkage mechanism may introduce bias by nudging the null-effect clusters toward a non-zero mean. This situation can increase the likelihood of incorrectly rejecting the null hypothesis for those clusters, especially if the number of clusters is high, cluster sizes are moderate (i.e., neither very small nor large enough to stabilize estimates), and within-cluster variability is high. So, while MLM performs well when there is meaningful between-cluster variability, it can be misleading when the fixed effect does not adequately represent the typical cluster, for example, when effects have a bimodal distribution.

In sum, MLM is a flexible modeling tool, but its performance depends heavily on the underlying data structure and modeling assumptions. For studies where cluster-specific inference is of interest and where partial pooling is appropriate, MLM offers a compelling alternative to conventional Type I error adjustment procedures, addressing the multiplicity problem without the need for post-hoc corrections.

Based on our findings, we offer the following practical recommendations:

1. Use MLM when cluster-specific effects are of interest and you expect either meaningful variation between cluster effects or that all effects are null. MLM performs best when it can draw on between-cluster heterogeneity to stabilize estimates. Be cautious when applying MLM in scenarios suspected of having two (or more) “dominating” effects (e.g., half of the clusters’ effects are null and the other half is non-null), but there is little difference between the effects (e.g., small effects in the non-null clusters). In such cases, shrinkage can backfire.

2. Do not assume MLM “solves” the multiplicity problem. Its advantages are conditional, not universal. When in doubt, simulate your scenario or conduct a sensitivity analysis.
3. Report and interpret fixed effects, random effect variability estimates, and cluster-specific estimates (when appropriate; e.g., if there are not too many clusters), including CIs and SEs. This comprehensive reporting improves transparency and helps readers comprehend the nuances of the data and have a more granular understanding of the relationships of interest. Focusing only on the fixed effects, as often occurs in psychology, can be misleading and lead researchers to overlook relevant details.

Limitations and Future Directions

Like any simulation study, ours also has limitations. We focused on balanced designs where all clusters had equal sample sizes and equal group allocation, ensuring that the slope’s fixed effect coefficient represented the pure within-cluster effect. In unbalanced designs, the fixed effect can blend within- and between-cluster effects—what Hoffman & Walters (2022) term “smushed effects”—unless the model explicitly separates them (e.g., by group-mean centring the predictor and including cluster means at Level 2). While this feature kept the design manageable, real-world data are often more complicated; for example, unequal cluster sizes, missing data, and covariates are common, and each of these aspects could influence the performance of MLM in subtle ways. We also limited our scope to two-level models with continuous outcomes, estimated using frequentist MLM with REML. Extensions to three-level models, categorical outcomes, or Bayesian estimation with informative priors remain open areas

for further research. It is also worth exploring how MLM interacts with covariates, especially in more complex designs.

Our focus was on Type I error and power, but we did not consider related issues like Type S (sign) or Type M (magnitude) errors (Gelman & Carlin, 2014; Gelman & Tuerlinckx, 2000), or broader concerns around interpretation, transparency, and model misspecification (except for the complete null conditions where between-cluster variability is zero in the population).

Another important direction for future work involves expanding this investigation to include Bayesian MLM. Although our study focused on frequentist MLM, prior proposals—most notably by Gelman et al. (2012)—argue that Bayesian MLM, specifically, offers a principled solution to the multiplicity problem. Gelman et al. (2012) suggest that when viewed from a Bayesian hierarchical modeling perspective, the problem of multiple comparisons disappears. Similarly, Sjölander and Vansteelandt (2019) argue that multiplicity is not an issue at all within the Bayesian framework. However, these claims often blend the effects of partial pooling with those of the priors, making it unclear whether the observed error control in Bayesian MLM stems from the structure of the model, the prior distributions, or their interaction. Future simulation studies could help disentangle these mechanisms by systematically varying the type and informativeness of priors (e.g., weakly informative, informative, or bimodal) while keeping the hierarchical structure constant. This investigation would clarify whether Bayesian estimation provides added benefits over frequentist MLM and under what conditions these benefits are most pronounced, offering applied researchers clearer guidance on the role of priors in addressing multiplicity.

Conclusion

This study presents initial evidence that MLM, through partial pooling, can help address the multiplicity problem under certain conditions, particularly when all effects are expected to be null or when there is meaningful heterogeneity across clusters. However, the results also indicate that MLM does not consistently resolve the multiplicity problem across more realistic partial-null scenarios. Rather than serving as a universal substitute for post-hoc adjustments, MLM offers a model-based framework whose performance depends on the configuration of effects and data structure. These findings clarify the conditions under which partial pooling influences Type I error behaviour and provide a foundation for further investigation, including Bayesian extensions that more explicitly model uncertainty and prior structure.

CHAPTER THREE

A BAYESIAN SOLUTION TO A FREQUENTIST PROBLEM? ADDRESSING THE MULTIPLICITY PROBLEM WITH BAYESIAN HIERARCHICAL MODELING (STUDY 2)

Modern empirical investigations in psychology typically involve testing multiple hypotheses at once (Martinez Gutierrez et al., 2025). Researchers are often interested in comparing numerous subgroups, conditions, or interventions simultaneously. However, conducting multiple significance tests increases the chance of drawing incorrect conclusions. The problem of multiplicity—arising from the inflation of Type I error rates due to multiple testing—has long been a methodological concern in psychological research and many other sciences (Cribbie, 2017; Frane, 2020; Hancock & Klockars, 1996).

In Study 1, we explored how multilevel modeling (MLM), a model-based alternative to traditional Type I error control procedures, can address the issue of multiplicity. Study 1 provided preliminary evidence that MLM, via partial pooling, can mitigate the multiplicity problem, especially in scenarios where all effects are null or where effects vary substantially across clusters. Although MLM is not a universal solution, it offers a strong alternative to post-hoc adjustments such as familywise error rate (FWE) control and false discovery rate (FDR) procedures when used under the right conditions. Results from Study 1 clarify how partial pooling contributes to controlling Type I error inflation, but they also raise questions about the extent to which model-based approaches can be further improved by incorporating prior information and Bayesian thinking.

Previous research has argued that, when viewed through a Bayesian hierarchical lens, the multiplicity problem “disappears entirely” (Gelman et al., 2012, p. 189), or is at least inherently addressed through the regularizing effect of partial pooling across parameters (Berry & Hochberg, 1999). Others have suggested that Bayesian hierarchical models offer a coherent inferential framework in which error rates are implicitly controlled by modeling assumptions related to prior beliefs (Chang & Berger, 2021; Sjölander & Vansteelandt, 2019). Although both perspectives may hold simultaneously, it remains unclear whether the empirical false positive behavior of Bayesian hierarchical models is primarily driven by shrinkage from partial pooling, the influence of prior distributions, or the interplay between the two, and how each component contributes to reducing inflated error rates under repeated sampling.

Despite these theoretical promises, the extent to which Bayesian multilevel modeling (BMLM) controls Type I error and retains adequate statistical power in the context of multiple tests remains an empirical question. This empirical question becomes especially complex when researchers use different types of priors—such as software defaults, weakly informative priors, or strongly informative priors—and when those priors either align or conflict with the true data-generating process. Although default or weakly informative priors are commonly used in applied psychology (Depaoli et al., 2020) and even suggested to address multiple testing by Gelman et al. (2012), their impact on multiplicity control has not been systematically evaluated in comparison to frequentist methods. Moreover, while some priors may be deliberately specified to reflect null or small effects, others may express directional or even strong expectations about the presence of an effect, raising important questions about their influence on false positive rates across a set of comparisons.

BMLM represents a natural extension of the approach evaluated in Study 1. Similar to its frequentist counterpart used in Study 1, BMLM involves partial pooling of cluster-level effects, but it also enables researchers to formally incorporate prior beliefs about model parameters and to quantify uncertainty through posterior distributions. In this study, we empirically evaluate the extent to which BMLM mitigates the multiplicity problem as previously proposed (e.g., Berry & Hochberg, 1999; Gelman et al., 2012). Our goal is to examine how partial pooling, prior information, and posterior-based inference under a hierarchical Bayesian framework *jointly* influence familywise false positive rates under repeated sampling across different conditions. In doing so, we also disentangle the respective contributions of different prior distributions and shrinkage to reducing false positive rates by contrasting Bayesian and frequentist MLM to one another and traditional multiplicity control procedures.

How Bayesians View and Approach the Multiplicity Problem

At the core of Bayesian thinking is the idea that probability reflects the strength of a belief in—or credibility of—a hypothesis, the value of a parameter, or the occurrence of an event, based on available evidence and prior knowledge (Kruschke, 2015). A key philosophical and mathematical distinction between frequentist and Bayesian perspectives is that frequentist statistics considers the parameter being estimated as a fixed (yet often unknown) quantity, whereas Bayesian statistics regard the parameter as a random variable with a distribution of possible values, defined by a probability distribution referred to as the *posterior* (Gelman et al., 2013; Kruschke, 2015; McElreath, 2020).

In Bayesian inference, parameters are considered unknown, “fluid” values portrayed by probability distributions, not fixed quantities; as new data are observed, the posterior distribution

is updated to reflect evolving beliefs (Gelman et al., 2013; Kruschke, 2015; McElreath, 2020). Rather than testing whether a parameter is exactly zero, Bayesian methods typically focus on estimating *a range of possible values* and quantifying the probabilities (i.e., credibilities or degrees of belief) associated with each. Because of this focus on continuous uncertainty using degrees of belief—as opposed to a binary decision given by the null hypothesis significance testing (NHST) approach commonly used in frequentist statistics—Bayesian frameworks do not rely on long-run (i.e., repeated sampling) Type I error rates to guide inference (Gelman et al., 2013; Gelman et al., 2012).

Instead, multiplicity can be addressed through the structure of the model itself (e.g., hierarchical or random-effects models that [partially] pool information across units; Berry & Hochberg, 1999; Gelman et al., 2012) or through the specification of prior distributions. For instance, researchers can assign priors that reflect relationships between model parameters (i.e., hyperparameters; Berry & Hochberg, 1999; Sjölander & Vansteelandt, 2019), priors that ration a common probability for each test that depends on the total number of hypotheses (Chang & Berger, 2020; Bogdan et al., 2008; Scott & Berger, 2006; Scott & Berger, 2010), or using shrinkage priors such as the spike-and-slab (Chang & Berger, 2020) and horseshoe (Piironen & Vehtari, 2017). These features shift the focus from correcting for the number of tests after the fact to modeling the generative process that gives rise to them, offering a fundamentally different approach to the problem of multiple comparisons.³ The next sections review the key theoretical

³ But this approach does not eliminate the need to guard against overinterpretation of spurious results.

foundations, modeling strategies, and debates in the Bayesian treatment of multiplicity, with a focus on multilevel frameworks.

A Bayesian Framework for Multiplicity

Sjölander and Vansteelandt (2019) apply and extend the Bayesian approach to multiplicity originally proposed by Berry and Hochberg (1999), illustrating its use with data from Bygren et al. (2014) on potential epigenetic effects. Here, we adapt the core idea to a simple hypothetical example. Suppose a psychology department chair suspects that faculty response speed to departmental emails depends on the subject-line tone, and that this may differ for tenured versus untenured faculty. The chair runs a small experiment by randomly sending two subject lines (e.g., “URGENT: IMMEDIATE ACTION REQUIRED!!!” vs. “Assistance with a Friendly Update”) to tenured and untenured faculty and records whether each faculty member replies within one week.

To connect this example to multiplicity, consider two *latent binary indicators*, one for untenured and one for tenured faculty, denoted θ_u and θ_t , where $\theta_g \in \{0, 1\}$ indicates whether a subject-line effect is present within group g (1 = effect present; 0 = no effect). These θ s are hypothesis indicators rather than observed outcomes or response probabilities. Let the observed evidence for each group be summarized by X_t and X_u , which could be a standardized effect summary (e.g., a difference in response proportions between the two subject lines) or, more generally, a statistic that captures the strength of evidence for an effect within each group.

The crucial step is to acknowledge that these parameters, θ_u and θ_t , are likely related. Intuitively, if the subject-line manipulation appears effective in one group, it becomes more plausible that the effect exists in the other group as well than that the effect is absent. In the

Bayesian framework, this dependence is represented directly through a *joint prior* distribution, $p(\theta_u, \theta_t)$ that does not factorize into $p(\theta_u)p(\theta_t)$. Combining this joint prior with the likelihood of the observed evidence yields the joint posterior:

$$\begin{aligned} p(\theta_u, \theta_t | X_u, X_t) &= \frac{p(X_u, X_t | \theta_u, \theta_t) p(\theta_u, \theta_t)}{\sum_{\theta_u, \theta_t=0,1} p(X_u, X_t | \theta_u, \theta_t) p(\theta_u, \theta_t)} \\ &= \frac{p(X_u | \theta_u) p(X_t | \theta_t) p(\theta_u, \theta_t)}{\sum_{\theta_u, \theta_t=0,1} p(X_u | \theta_u) p(X_t | \theta_t) p(\theta_u, \theta_t)}. \end{aligned} \quad (2.1)$$

Even if the observed data, X_u and X_t , are independent of one another, having observations from one faculty position gives us at least some insight into the parameter of the other because the prior links the indicators. For example, the posterior probability that an effect is present among untenured faculty is

$$p(\theta_u = 1 | X_u, X_t) = \frac{\sum_{\theta_t=0,1} p(X_u | \theta_u = 1) p(X_t | \theta_t) p(\theta_u=1, \theta_t)}{\sum_{\theta_u, \theta_t=0,1} p(X_u | \theta_u) p(X_t | \theta_t) p(\theta_u, \theta_t)}. \quad (2.2)$$

The implication for multiplicity is simple: when related hypotheses are modeled jointly, evidence bearing on one hypothesis can influence posterior beliefs about the others. In this sense, multiplicity adjustment can emerge from the model itself, because posterior inferences about any single hypothesis are made in the presence of the other hypotheses through the shared prior structure. Importantly, this is not a post-hoc correction applied after separate tests; it is a consequence of specifying dependence among the parameters (or indicators).

Hierarchical Models: Model-Based Control

Bayesian hierarchical models provide another strategy for addressing multiplicity (Berry & Hochberg, 1999; Gelman et al., 2012). Returning to our previous email-response example,

consider the case where the parameters of interest represent population-level response probabilities for each group rather than binary effect latent indicators. Let θ_u and θ_t now represent the probability of a prompt reply for untenured and tenured faculty, respectively. For instance, if 7 out of 10 of the untenured faculty answered within a week, then $\hat{\theta}_u = 0.7$.

In a hierarchical formulation, group-specific parameters are treated as exchangeable draws from a common population distribution. Extending this example further, let θ_u and θ_t be drawn from a population distribution, G . Here, G is a population distribution over group-specific response probabilities, so that each θ_g is treated as an exchangeable draw from a common $\text{Beta}(\alpha, \beta)$ distribution. Conceptually, this formulation assumes that the different personnel groups in the department (e.g., tenured faculty, untenured faculty, teaching-stream faculty, staff) are not fundamentally unrelated, but instead represent draws from a broader distribution of response tendencies. If each group contains multiple individuals (e.g., $y_{u_1}, y_{u_2}, y_{u_3}, \dots, y_{u_U}$), then this is a multilevel model in the traditional sense, with individuals (Level 1) nested within groups (Level 2). Here, each y_{g_i} represents whether a particular individual, i , in group g (e.g., Maurice Moss from IT) replies within a week ($y_{g_i} = 1$) or not ($y_{g_i} = 0$), which we can model as

$$y_{g_i} \sim \text{Bernoulli}(\theta_g), \quad \theta_g \sim \text{Beta}(\alpha, \beta), \quad (2.3)$$

where θ_g s are random effects expressing group-specific probabilities of a prompt response. The θ_g values are linked through the shared population distribution, $G \sim \text{Beta}(\alpha, \beta)$, allowing information to be shared across groups. In this hierarchical framework, the posterior distribution for one group's parameter (e.g., θ_u) depends on its own group's data ($y_{u_1}, y_{u_2}, y_{u_3}, \dots, y_{u_U}$) and also, indirectly, on data from other groups (e.g., $y_{t_1}, y_{t_2}, y_{t_3}, \dots, y_{t_T}, y_{a_1}, y_{a_2}, y_{a_3}, \dots, y_{a_A}$, etc.) because the other groups' data inform the shared population's hyperparameters, α and β . Since G

governs all θ_g s in the model, any update to our understanding of α and β estimates affects our inferences about θ_u .

This is the mechanism of partial pooling: strong evidence that tenured faculty respond promptly increases our belief that untenured faculty might do so as well and vice versa, simply because they are modeled as draws from the same distribution. This “borrowed strength” causes the posterior estimates for each θ_g to shrink or pull toward one another and toward the overall departmental rate. The force of the shrinkage depends on both the within-group sample size and the variability within and between groups, whereas the direction of the pull is the posterior mean of G , which depends on both the within- and between-group effects. Because the inferences of the random effects are partially pooled rather than treated as independent, the model inherently “softens” extreme estimates (i.e., distant from the fixed effect); if the fixed effect is null or close to zero, the number of false positives will be reduced. Thus, shrinkage toward a common departmental response rate—if it is null or at least low—can mitigate the impact of multiple comparisons without the need for post-hoc adjustments.

Bayesian Multiplicity Control (in Variable Selection)

Bayesian variable selection and model comparison procedures, such as Bayes factors and posterior inclusion probabilities (PIP), are said to automatically adjust for multiplicity under certain conditions. Scott and Berger (2010) show that both fully Bayesian and empirical Bayes approaches incorporate multiplicity control in high-dimensional regression contexts. A common misconception is that this control comes solely from the Bayesian “Ockham’s razor” effect: integrating likelihoods over higher-dimensional parameter spaces spreads prior mass more thinly, so larger models need to provide a substantially better fit to achieve the same marginal

likelihoods. Although the Bayesian Ockham's razor effect favours parsimony, it is not a multiple-testing penalty per se. Similarly, simply assigning equal prior probability to all possible models (e.g., each model has a prior probability of 1 divided by the total number of models) is akin to having a prior inclusion probability of .5 for each predictor. This setup would not control for multiplicity because it will always correspond to an expected model size which is half of the total number of predictors, $m/2$, regardless of how large m is because each predictor in the model has a 50% chance of being included Scott and Berger (2010).

True Bayesian multiplicity adjustment enters through the choice of priors that consider or influence variable selection probabilities and depends on the data (and m). When many hypotheses or predictors are evaluated simultaneously, a modern practice in Bayesian variable-selection problems is to model variable inclusion as exchangeable Bernoulli trials that share a common success probability across predictors, p ; p is the prior probability that each predictor's coefficient has a non-zero value. To illustrate this idea, let $\gamma_i \in \{0,1\}$ be a binary latent indicator of whether a variable, x_i , should be included as a predictor in the model (i.e., $\gamma_i = 1$; $H_1: b_i \neq 0$) or excluded (i.e., $\gamma_i = 0$; $H_0: b_i = 0$). Conditional on the inclusion probability p , we model $\gamma_i \sim \text{Bernoulli}(p)$, so the prior probability of any model M_γ is:

$$p(M_\gamma | p) = p^{k_\gamma} (1 - p)^{m - k_\gamma}, \quad (2.4)$$

where k_γ is the number of included variables in the model.

If we treat p as a random parameter to be estimated from the data, it will “yield an automatic multiple-testing penalty” (Scott & Berger, 2010, p. 6). p can be estimated from the data using maximum likelihood (i.e., empirical Bayes) or using fully Bayesian approaches by assigning p a prior distribution. The idea here is that, as the number of candidate variables (m)

increases while the true number of relevant ones (k) stays fixed, the posterior for p (or EBLUP prediction) will shrink near zero. For example, from a fully Bayesian approach, say we set a prior of $p \sim \text{Beta}(a, b)$, so its posterior will be:

$$p \mid \gamma \sim \text{Beta}(a + k, b + m - k), \quad (2.5)$$

for which its mean will be:

$$E[p \mid \gamma] = \frac{a+k}{a+b+m}. \quad (2.6)$$

Thus, as m increases, the expected p decreases, which means that the probability of each predictor being included in the model is smaller with a larger number of possible predictors—in other words, multiplicity control.

The hierarchical prior on p provides the mechanism by which Bayesian variable selection adapts to the scale of the problem and controls multiplicity. This approach is conceptually similar to having a very small prior probability of including any given variable in the model, such as $p \approx 1/m$ (Scott & Berger, 2006). Note that, according to Jeffreys (1961; as cited in Scott & Berger, 2010), multiplicity control also automatically takes hold when p is set more subjectively. For instance, all models with the desired, or believed to be true, number of predictors are assigned a higher prior probability than the rest, or by choosing p to capture how many variables one believes should be in the model (i.e., proportion of variables likely to be included) given m .

Shrinkage Priors

Spike-and-Slab Priors. Chang and Berger (2020) take this discussion further by showing mathematically that Bayesian variable-selection priors—especially the spike-and-

slab—have a built-in multiplicity adjustment that can also apply to mean comparisons. In their setup, each mean parameter μ_i has a mixture prior that assigns a point mass (the “spike”) at zero and a continuous distribution (the “slab”) that centers on zero, but spreads over nonzero effects, too:

$$\mu_i \sim p\delta_0 + (1 - p) \mathcal{N}(0, \sigma^2), \quad (2.7)$$

where δ_0 denotes a point mass at zero (i.e., a distribution that assigns all its mass to the value 0) and p is the probability that the mean is exactly zero (true null), $p = P(\mu_i = 0)$. Consequently, most mean estimates are forced to be close to zero, whereas only a few very strong observed means are allowed to escape from the spike at zero to provide non-zero mean estimates. Chang and Berger provide proofs of how the posterior behaves under spike-and-slab priors: in the complete null case—when all true mean differences are zero—the Bayesian posterior drives error rates toward zero as the number of mean comparisons increases.

Horseshoe Priors. Another influential shrinkage prior for high-dimensional problems is the horseshoe prior (Carvalho et al., 2010; Piironen & Vehtari, 2017). Unlike the spike-and-slab, which discretely separates coefficients into “exactly zero” and “possibly nonzero,” the horseshoe is a continuous shrinkage prior that automatically adapts to sparsity in the data. Each parameter μ_i is given a prior:

$$\mu_i \sim \mathcal{N}(0, \lambda_i^2 \tau^2), \quad \lambda_i \sim C^+(0, 1), \quad \tau \sim C^+(0, 1), \quad (2.8)$$

where $i = 1, 2, 3, \dots, m$, λ_i is a local shrinkage parameter assigned to each μ_i , τ is a global shrinkage parameter, and C^+ denotes the standard half-Cauchy distribution. The idea here is that

the global parameter τ enforces overall shrinkage toward zero, whereas the local parameters λ_i allow some coefficients to escape the shrinkage if the data strongly support nonzero effects.

From a multiplicity perspective, this structure works similarly to spike-and-slab: as the number of tests m increases but true signals remain the same, the global parameter τ pulls most effects toward zero, which can reduce false positives. At the same time, the local parameters λ_i adaptively relax the shrinkage on large effects. Bogdan et al. (2011) showed that the horseshoe prior achieves near-optimal performance for sparse true effects, making it another promising alternative. Conceptually, the horseshoe is closer to how MLM and partial pooling handle multiplicity—through continuous shrinkage of extreme estimates—than to discrete model-selection procedures like the spike-and-slab.

Bayesian Multilevel Modeling (BMLM)

Consider the same model and working example as in Study 1: a random-slopes MLM model where slopes and intercepts are allowed to covary:

$$\text{Level 1: } y_{ij} = b_{0j} + b_{1j}X_{1j} + e_{ij} \quad (2.9)$$

$$\text{Level 2a: } b_{0j} = \gamma_{00} + \delta_{0j}$$

$$\text{Level 2b: } b_{1j} = \gamma_{10} + \delta_{1j}$$

where, again, y_{ij} is the mental health measured score for client i in clinic j , b_{0j} is the mean mental health score for the control group in clinic j , b_{1j} represents the effect of treatment in clinic j , X_1 is an indicator for treatment status (i.e., 0 for control, 1 for new therapy), and e_{ij} is the difference between client i 's mental health score and their treatment group's mean in clinic j , which is distributed as $e_{ij} \sim N(0, \sigma^2)$. Here, γ_{00} represents the grand mean, which is the mean mental health

score of the control group across all clinics, and δ_{0j} is the deviation of clinic j 's control group mean from the grand mean, distributed as $\delta_{0j} \sim N(0, \tau_{00})$. Next, γ_{10} is the overall treatment effect, representing the mean difference in mental health score between the treatment and control groups across all clinics, and δ_{1j} is the deviation of clinic j 's treatment effect from the overall effect, distributed as $N(0, \tau_{11})$. Further, the random intercept and slope are allowed to covary, $Cov(\delta_{0j}, \delta_{1j}) = \tau_{01}$.

Gelman et al. (2012) illustrate shrinkage in hierarchical models using the scalar case of cluster-specific effects such that:

$$E(b_{1j} | \hat{b}_{1j}) = \frac{\frac{1}{\tau_{11}}\gamma_{10} + \frac{1}{v_{1j}}\hat{b}_{1j}}{\frac{1}{\tau_{11}} + \frac{1}{v_{1j}}}, \quad sd(b_{1j} | \hat{b}_{1j}) = \frac{1}{\sqrt{\frac{1}{\tau_{11}} + \frac{1}{v_{1j}}}}, \quad (2.10)$$

where $\hat{b}_{1j}$ is the unpooled effect in clinic j and v_{1j} is its sampling variance. The equations above show that the posterior mean for each clinic effect is a weighted mean of the overall effect, γ_{10} , and the clinic's raw estimate, $\hat{b}_{1j}$, where the weights are determined by the relative precision of the two sources of information (see also Equations 1.7 and 1.8 in Study 1). Clinics with small samples (and large v_{1j}) are pulled more strongly toward the population mean, while sites with large samples remain close to their raw estimates. This is the formal expression of shrinkage as described in Gelman et al. (2012, pp. 195–196). This scalar formulation applies to the simple case of estimating a single cluster-level parameter.

In the present random-slopes model with correlated random effects, shrinkage is determined by the full covariance structure encoded in T and the sampling covariance matrix V_j

(Equations 2.11–2.13), rather than a single scalar weight. This logic extends naturally to a random-slope model with correlated random effects. For clinic j , let:

$$b_j = \begin{bmatrix} b_{0j} \\ b_{1j} \end{bmatrix} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix} + \begin{bmatrix} \delta_{0j} \\ \delta_{1j} \end{bmatrix}, \quad (2.11)$$

with:

$$\begin{bmatrix} \delta_{0j} \\ \delta_{1j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, T \right), \quad T = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}. \quad (2.12)$$

Now, let $\hat{b} = (\hat{b}_0, \hat{b}_1)$ denote the OLS intercept and slope estimates of clinic j , respectively, with sampling covariance matrix V_j . By standard conjugate updating (Raudenbush & Bryk, 2002; Gelman & Hill, 2007), the posterior distribution for the clinic-level coefficients is:

$$b_j \mid \hat{b}_j, \mu_b, T \sim MVN(m_j, S_j), \quad (2.13)$$

with:

$$S_j = (T^{-1} + V_j^{-1})^{-1}, \quad m_j = S_j(T^{-1}\mu_b + V_j^{-1}\hat{b}_j),$$

where $\mu_b = (\gamma_{00}, \gamma_{10})^T$. Here, m_j is the posterior means vector for the intercept (first component) and slope (second component) in clinic j , and S_j is the posterior variance-covariance matrix. This multivariate structure mirrors the scalar formula of Gelman et al. (2012): each clinic's estimates are shrunk toward the grand means, μ_b , with the amount of pooling determined by the relative precision of the clinic data (V_j) versus the covariance structure in T . However, because T is a full covariance matrix, shrinkage occurs jointly across intercepts and slopes: the slope estimates depend on the entire covariance structure in T , not solely on the marginal slope variance τ_{11} .

Note that these posterior mean and covariance equations are equivalent to Equations 1.7

and 1.8 in Study 1, describing the empirical Best Linear Unbiased Predictor (EBLUP) expression. The fully Bayesian analogue has the same conditional form, but rather than plugging in fixed values for the hyperparameters, it integrates over their posterior distribution. In full Bayes, we do not condition on fixed hyperparameters. Instead, we treat them as random variables with priors. Following Raudenbush and Bryk (2002), we let $p_1(b \mid \gamma, T)$ denote the (conditional) prior distribution of the random effects given the hyperparameters and $p_2(\gamma, \sigma^2, T)$ the “true” prior distribution over these hyperparameters. The joint posterior distribution of all unknowns, given the observed data $\mathcal{D}$, can then be written as

$$p(b, \gamma, \sigma^2, T \mid \mathcal{D}) \propto f(\mathcal{D} \mid b, \sigma^2) p_1(b \mid \gamma, T) p_2(\gamma, \sigma^2, T), \quad (2.14)$$

where $f(\mathcal{D} \mid b, \sigma^2)$ is the conditional density of the observed data (i.e., the likelihood function) given the random effects and residual variance. From this joint posterior, inferences about any parameter of interest are obtained by integrating (marginalizing) over the other parameters. For example, the posterior distribution of the random effects, b_j , for cluster j is:

$$p(b_j \mid \mathcal{D}) = \int p(b_j \mid \widehat{b}_j, \gamma, T, \sigma^2) p(\gamma, T, \sigma^2 \mid \mathcal{D}) d\gamma dT d\sigma^2, \quad (2.15)$$

which averages the posterior of b_j conditional on the (assumed fixed) hyperparameters and OLS estimates of the intercept and slopes (i.e., data summary) for cluster j (this component can be thought of as “what cluster j ’s effect would look like if we knew the true hyperparameter values”) over the posterior uncertainty in those parameters (the “price” for acting as if we knew the true hyperparameter values). The corresponding posterior mean of b_j is then:

$$E[b_j \mid \mathcal{D}] = \int m_j p(\gamma, T, \sigma^2 \mid \mathcal{D}) d\gamma dT d\sigma^2 \quad (2.16)$$

where m_j (the conditional posterior mean of b_j given the hyperparameters) is defined above. Note that m_j is identical to Equation 1.7 in Study 1:

$$m_j = E[b_j \mid \mathcal{D}, \gamma, T, \sigma^2] = \lambda_j \hat{b}_j + (I - \lambda_j) \gamma, \quad (2.17)$$

with:

$$\lambda_j = \left(\frac{1}{\sigma^2} X_j^\top X_j + T^{-1} \right)^{-1} \frac{1}{\sigma^2} X_j^\top X_j. \quad (2.18)$$

Thus, this expression shows that the posterior mean of each clinic's effect is a shrinkage-weighted mean of the clinic's OLS estimate $\hat{b}$ and the overall mean γ , with the weights determined by the precision of the clinic data (V_j) relative to the between-clinic variance (T).

In summary, E/BLUP and fully Bayesian shrinkage share the same algebra at the conditional level, but full Bayes propagates uncertainty in the hyperparameters instead of treating them as known. The scalar expressions presented by Gelman et al. (2012) are thus a special case of this more general multivariate result. When focusing on the slope component, b_{1j} , the posterior mean can resemble a weighted-average form under simplifying conditions (e.g., diagonal T , independent random effects, and suitably structured V_j). However, in the general random-slopes model with correlated random effects, shrinkage is determined by the full covariance structure and does not reduce to a single scalar weight. In practice, the multivariate form captures the richer structure of intercept–slope covariance across clinics, while preserving the same principle: posterior estimates reflect a compromise between clinic-specific information and the population distribution, with greater shrinkage when data are less precise or between-clinic variability is small.

Of course, these posterior estimations are analytically convenient because we use conjugate priors (a normal–normal setup) here, which yields closed-form posteriors for the cluster-specific effects. This convenience arises because the normal likelihood for the cluster-level estimates combines algebraically with the assumption that the cluster intercepts and slopes are drawn from a common normal distribution governed by γ and T . But once we move away from this conjugate structure—for example, by using horseshoe or spike-and-slab priors on the hyperparameters—we no longer have the luxury of having closed-form solutions, and the posteriors must be approximated with Markov Chain Monte Carlo (MCMC) sampling.

To date, most of the formal results have been worked out under the “complete null” or joint null hypothesis, where all true cluster effects are zero. In such a case, shrinkage behaves as advertised: the cluster estimates are pulled toward the grand mean (around zero), producing a kind of multiplicity control byproduct. But, in psychology, it is less likely that *all* effects tested are truly zero, no matter how small and practically negligible the effects are. Perhaps more realistically, some clusters show genuine effects (even if they are infinitesimally small), whereas others do not. We know surprisingly little about how BMLM behaves in these partial-null situations—where null and non-null effects coexist—and even less about how prior choices change the balance between estimates, precision, and Type I errors in these more sensible situations. To address this issue, the present study evaluates how BMLMs perform under conditions that better reflect real research settings (i.e., partial-null situations), and how different prior specifications shape the resulting inferences.

A Note on Statistical Significance

In Study 1, we used BLUP estimates and their precision to construct confidence bands. To determine whether a cluster-specific effect was statistically significant at a nominal Type I error $\alpha = .05$, we evaluated whether the 95% confidence interval excludes zero (“rejecting the null hypothesis”) or does not (“fail to reject the null”). In Bayesian statistics, confidence bands are expressed as credible intervals. The credible interval is bounded by the lower and upper pre-specified percentiles of the parameter’s posterior distribution. For instance, a 95% credible interval includes all values between the 2.5th and 97.5th percentiles of the posterior. Although Bayesian inference does not rely on null hypothesis significance testing in the frequentist sense, simulation studies often evaluate the repeated-sampling behaviour of Bayesian procedures using frequentist operating characteristics. We therefore calculated the proportion of replications in which a 95% credible interval for a cluster-specific slope excluded zero. When the true population slope was zero, this proportion was interpreted as an empirical false positive rate, a Type I error, for the purpose of this study.

Current Research

The present study builds on Study 1 by incorporating BMLM into the same simulation framework used to evaluate multiplicity control. Specifically, we examine how different prior distributions for the fixed effect and its hyperparameters influence the familywise Type I error rate and statistical power when conducting simultaneous tests across multiple clusters. The priors for the fixed effect include flat priors, skeptical (weakly informative) priors, confident (informative) priors, and more complex prior structures such as the horseshoe and spike-and-slab

distributions. Each prior is applied across a similar range of design conditions used in Study 1, including variation in sample size, number of clusters, and proportion of true null hypotheses.

By comparing BMLM directly with frequentist MLM and standard multiplicity-control methods, this study examines whether Bayesian approaches provide reliable inference in the context of many simultaneous tests. The goal is not only to evaluate performance (e.g., multiplicity control) but also to demonstrate how different prior choices affect the Type I error rate. This work is meant to be both methodological and practical. Together, Study 1 and Study 2 provide a fuller picture of how both frequentist and Bayesian strategies can be used to think about the problem of multiplicity in psychological research.

Method

Monte Carlo Simulation Study

Objective

This simulation study extends Study 1 by adding BMLM as another approach to multiplicity. This study focuses on how alternative prior distributions for the fixed effect and its hyperparameters affect familywise Type I error and statistical power when testing across multiple clusters. This second study directly compares frequentist MLM, BMLM, and standard multiplicity adjustments, all within the same population models.

Design

The simulation setup followed the same applied example as in Study 1: a network of psychotherapy clinics where clients were randomly assigned to either a new therapy or a control condition. As before, we varied core design aspects such as the number of clusters, within-cluster

sample size, and the proportion of clusters with true null effects. The resulting conditions ranged from scenarios where all clusters were null (complete null conditions) to mixtures of null and non-null clusters (partial null conditions) and cases where no clusters were null (power-only conditions). For non-null clusters, the magnitude of the true effect was also varied to examine how shrinkage toward the overall mean influenced both false positives and power.

The design of the current study was slightly reduced from the design of Study 1. Because we examined nine different prior specifications in each condition (alongside MLM, FWE, FDR, and unadjusted mean comparisons), the full set of design conditions would have been computationally impractical. For example, certain conditions (e.g., 500 individuals in every 256 clusters) would have required several hours of run time per iteration (i.e., several hours for each of 1,000 iterations), even while using a parallel, multi-core high-performance computing (HPC) service. We therefore reduced the number of conditions, balancing feasibility with the ability to observe the key performance patterns across priors.

As in Study 1, statistical significance was evaluated at the cluster level: rejections in non-null clusters contributed to power whereas rejections in true-null clusters were counted as Type I errors. The factors manipulated in the design are summarized in [Table 2.1](#). The different priors and their specifications are presented in [Figure 2.1](#).

Procedure

Data generation followed the same structure as in Study 1, with n_j individuals nested within each of J clusters. Within each cluster, a binary variable indicated treatment status (MT = 0 for control, and MT = 1 for treatment). The outcome variable (mental health) was drawn from a normal distribution with a cluster-specific treatment effect (b_{1j} ; the mean difference between the

treatment and control groups), an intercept (set at 0), and residual variance σ^2 (set at 1). Using a bottom-up approach, each cluster's data were independently and randomly generated and then combined into a single dataset. This procedure allowed us to determine the true cluster effects explicitly—set to zero for null clusters (i.e., $b_{1j} = 0$) and non-zero for non-null clusters (i.e., $b_{1j} = 1$ or 5)—depending on the manipulated null proportion (π).

We extended the Study 1 framework by fitting BMLMs alongside frequentist MLMs and traditional multiplicity-control methods. Frequentist MLMs were estimated with the *lme4* package using the `lmer` function (Version 1.1-36; Bates et al., 2015) and the *mixedup* package with the `extract_random_coefs` function (Version 0.4.0; Clark, 2024). BMLMs were estimated using the *brms* package with the `brm` function (Version 2.22.00; Bürkner, 2017) and back-end sampling via *cmdstanr* (Version 0.8.0; Gabry et al., 2025).

For Bayesian models, a range of prior distributions was applied to the fixed effect and hyperparameters (see [Figure 2.1](#)). Each prior was tested independently; that is, no prior was tested with any other custom priors beyond the defaults implemented in *brms*. By default, *brms* assigns weakly informative or skeptical priors: a half-Student-*t* prior with 3 degrees of freedom, location 0, and scale 2.5 prior, half- $t(3, 0, 2.5)$, on standard deviations and residual error (σ), a uniform Lewandowski–Kurowicka–Joe, LKJ(1), prior on correlation matrices, a Student-*t* prior with 3 degrees of freedom, location 0.3, and scale 2.5 prior, $t(3, 0.3, 2.5)$, on the intercept, and a flat (improper uniform) prior on the fixed effect. We intentionally avoided “mixing and matching” priors to isolate the influence of each specific prior and parameter on model behaviour and inference. Also, crossing all priors would have produced a large number of prior

permutations and thus models per condition, which would have made the simulation even more computationally challenging.

It is worth noting that the horseshoe and spike-and-slab priors were applied here in a configuration that differs from their most common use case. Both priors were originally developed for high-dimensional settings where many parameters are estimated simultaneously, and most are expected to be zero—contexts where the horseshoe’s global shrinkage parameter can calibrate itself by surveying the entire collection of coefficients, and the spike-and-slab can assign inclusion probabilities across many parameters simultaneously. In the present study, however, each prior was applied to a single fixed-effect parameter, γ_{10} , representing the overall treatment effect. The practical function of both priors in this context was therefore to exert strong downward pressure on γ_{10} , pulling the overall treatment effect estimate toward zero and thereby encouraging shrinkage of cluster-specific slopes in that direction. In this sense, they operated less as their intended sense and more as highly aggressive null-favouring priors on a scalar parameter. Their intended use was to evaluate whether this aggressive regularization of the fixed effect bears any advantage for multiplicity control relative to simpler confident priors, and their performance should be interpreted in that light.

Models were run with four MCMC chains and 1500 total iterations per chain, including 500 burn-in iterations. Convergence was monitored using the Gelman–Rubin diagnostic $\hat{R}$ (which compares within- and between-chain variance and approaches 1 under convergence), effective sample sizes, and trace plots before the simulation for each model and condition. During the simulation, exclusion criteria were $\hat{R} > 1.1$, consistent with commonly used thresholds in large-scale simulation studies. To manage computation, simulations were

distributed across multiple HPCs (*Nibi* and *Narval*, hosted on Digital Research Alliance Canada). The full simulation code and HPC specifications are available on the Open Science Framework (OSF): osf.io/gbx26.

Evaluating the Performance of Testing Approaches

As in Study 1, performance was evaluated in terms of both familywise Type I error and statistical power. For Type I error, we defined the family of tests as the set of clusters with true null effects in each condition. The familywise error rate was calculated as the probability of at least one false rejection within this set. For power, we computed average per-comparison power, defined as the mean probability of correctly rejecting false nulls (i.e., clusters with non-null population effects). For each iteration, we recorded the proportion of non-null clusters with significant results, then averaged these proportions across replications to provide an overall measure of sensitivity to detect true effects while viewed alongside Type I error rates.

Results

Simulation results are presented in [Figures 2.2 through 2.5](#). The simulation code and additional materials, including high-resolution figures, are accessible via OSF: osf.io/gbx26.

Type I Error Rates

Complete-Null Conditions

In the condition where all clusters were generated under the null ([Figure 2.2](#)), the familywise Type I error rates aligned with Gelman et al.'s (2012) shrinkage hypothesis. For most prior specifications, the BMLM approach maintained consistently low error rates across sample size levels, averaging well below the nominal, $\alpha = .05$. However, not all priors performed

equally. One exception was the prior that assumed substantial between-cluster variability in slopes (τ_{11}), which yielded inflated error rates that are similar to those obtained by the unadjusted testing approach.

This exception arises because the prior that strongly favours (very informative) large slope heterogeneity is at odds with what is implied by the data and essentially counteracts shrinkage: cluster-specific slopes are pulled toward the overall mean (which is at or very near zero) to a lesser extent and are “encouraged” to vary more. When the data-generating process contains only null effects, this muffled shrinkage—coupled with narrow credible intervals—increases the chance that noise in individual clusters is mistaken for a true effect. In other words, the prior’s strong expectation of high between-cluster variability works against the multiplicity-controlling byproduct of shrinkage, inflating Type I error rates to levels resembling the unadjusted approach. To illustrate the phenomenon underlying these findings, we generated a random dataset under one of the complete-null conditions and plotted the cluster-specific slopes from the MLM and two BMLM models ([Figure 2.4](#), top row). The MLM and BMLM with default priors produce slope estimates tightly centered around zero with very narrow confidence bands, whereas the BMLM with a confident high-variability prior yields more dispersed estimates, with point estimates and error bars varying more freely around the fixed effect.

Partial-Null Conditions

Results from partial-null conditions, where there is a mix of both null and non-null clusters to varying degrees, are presented in [Figure 2.3](#). To maintain readability, results from the traditional approaches (FWE, FDR, and unadjusted comparisons) are not displayed in this figure but can be accessed on OSF and in [Figure 1.2](#) of Study 1.

Proportion of Null Clusters. As the proportion of clusters generated under the null increased from 25% to 75% (top to bottom rows in [Figure 2.3](#)), we observe a slight rise in familywise Type I error rates at each cluster-count condition. This pattern reflects the expected increase in the probability of having at least one Type I error occurring as the number of tests (i.e., clusters with true null effects) increases.

Effect Size. Non-null effect magnitude had a pronounced influence on MLM error rates, but much less so on BMLM. When non-null effects were small ($b_{1j} = 1$), the familywise Type I error rate for MLM declined only gradually as sample size increased. By contrast, when non-null effects were large ($b_{1j} = 5$), MLM error rates dropped sharply with increasing cluster size and stabilized at substantially lower levels. BMLM, on the other hand, showed relatively consistent error rates regardless of effect size.

In contrast to the frequentist MLM, where familywise Type I error rates fall with stronger effects, BMLM error rates remained rather high and stable. This difference likely occurred because τ_{11} estimates are more sensitive to the dispersion of slopes: when clusters with large effects are present, the estimated τ_{11} inflates, weakening shrinkage and limiting how much null clusters are pulled toward a nonzero mean. With BMLM, however, posterior estimates of τ_{11} are (at least partially) regularized by priors, which limits how much the variance component can inflate with larger effects. As a result, shrinkage remains stronger and null clusters continue to gravitate toward the non-zero overall mean, even at larger effect sizes.

Within-Cluster Sample Size. Similar to effect size, BMLM showed little sensitivity to within-cluster sample size, especially when effects were large. This finding likely reflects the role of priors in determining shrinkage. As n_j increases, raw estimates of cluster slopes (i.e.,

mean differences) carry more weight, affecting the estimated variance component (τ_{11}) and reducing shrinkage in frequentist MLM. In BMLM, however, shrinkage has a stronger inertia due to the influence of the priors on T and σ^2 , making it less responsive to increased sample size. As a result, posterior means of cluster-specific effects remain slightly more shrunk toward the overall mean in BMLM than in frequentist MLM.

Another factor that may explain the consistently higher familywise Type I error rates in BMLM compared to frequentist MLM is the narrower credible intervals of cluster-specific effects. When credible intervals are very narrow—particularly when the fixed effect is farther away from zero—the chance of excluding zero increases, raising the probability of a Type I error. In other words, even if the cluster-specific slope estimates are near zero, their intervals may be too tight to capture zero. This pattern is depicted in the bottom row of [Figure 2.4](#), where BMLM intervals are noticeably narrower than those from MLM, despite the point estimates being very similar.

Consistently high familywise error rates across within-cluster sample size levels also appear under the default *brms* priors. Although weakly regularizing, these priors still exert shrinkage toward smaller variance values, which can lead to narrower credible intervals when data are limited: the *brms* τ_{11} default prior is a half-Student- $t(3, 0, 2.5)$, and the intercept–slope correlation is assigned an LKJ prior, which under the default settings corresponds to a uniform prior over correlation matrices. Thus, the incorporation of priors, even vaguely informative ones, can counteract the beneficial effect of increasing the within-cluster sample size on shrinkage through the slope variance component. As a result, posterior estimates of τ_{11} might not inflate as strongly as in the frequentist case, which likely contributes to the stability of Type I error rates

across sample sizes. Moreover, the narrow credible intervals, which are present even under default priors, further contribute to the high error rates (see bottom row in [Figure 2.4](#)).

Number of Clusters. Conversely, the number of clusters (J) had a much stronger impact than sample or effect sizes. As the number of clusters increases, more tests are being performed, leading to a greater chance of observing at least one Type I error. Also, with more clusters, the posterior of τ_{11} is estimated more precisely, which can amplify shrinkage effects when the global mean is nonzero.

Prior Specifications. BMLM yielded nearly identical familywise Type I error rates across the different priors. There are a few likely explanations for this result. First, with moderate to large sample sizes or number of clusters, the likelihood dominated the priors, so posterior estimates of cluster-specific effects were similar across different priors. Similarly, in conditions that naturally produced high τ_{11} estimates—such as when effect sizes were large and clusters were clearly heterogeneous (e.g., clusters were either $b_{1j} = 5$ or $b_{1j} = 0$)—even the most aggressive, high-variability prior added little beyond what the data already implied. In such cases (as illustrated in the bottom row of [Figure 2.4](#)), the prior was aligned with the likelihood, and its influence appeared nearly indistinguishable from that of weaker or default priors under these conditions.

Second, shrinkage and Bayesian estimation produced relatively narrow credible intervals. When the overall mean was shifted away from zero due to non-null clusters, these tight intervals increased the chance that at least one null cluster's interval excluded zero, which inflated familywise errors. Together, the dominance of the likelihood and the inherently narrow BMLM intervals explain why priors showed little difference in error rates and why error rates remain

high in partial-null conditions. We further illustrate these explanations at the bottom of [Figure 2.4](#), where, even with relatively small sample and cluster sizes, the difference between the default and the strongest prior is negligible.

Finally, recall that cluster-specific slopes are modeled as exchangeable draws from a population distribution, rather than as fixed parameters without a higher-level distribution. Priors are specified on the hyperparameters of that population distribution, but not directly on the slopes. Hence, the priors' influence on cluster slopes and their credible intervals is only by proxy. As a result, the prior assumptions we specify in this study had a naturally bounded impact on error rates, which might also explain the similarity across specifications.

Notable differences between prior specifications emerged only under the most data-limited scenarios (e.g., $J = 12$, $n_j = 10$ or 20 , and $b_{1j} = 1$). In these conditions, the likelihood provided relatively weaker information, allowing priors to exert greater influence on posterior estimates. The confident high-variability prior, which places more mass on large between-cluster variance, inflated slope heterogeneity and widened credible intervals. This prior therefore produced lower familywise error rates when the proportion of null clusters was small ($\pi = 0.25$), because the estimated fixed effect remained closer to zero and higher error rates when the proportion of null clusters was large ($\pi = 0.75$) because the estimated fixed effect shifted further away from zero. These differences show that in small-sample, low-effect size contexts, prior assumptions about τ_{11} can meaningfully shape inference, whereas in larger or stronger data conditions, the consequence of using different priors for this approach is virtually non-existent.

Power Rates

Power rates for the small effect size condition ($b_{1j} = 1$) are shown in [Figure 2.5](#). Power results from large effects ($b_{1j} = 5$) are omitted because all methods reached a ceiling effect with power at 1.00 across all sample sizes and cluster counts. The primary differences occurred only when the effects and samples were small: unadjusted, FDR, and FWE had the lowest power, whereas the power of MLM and most BMLM priors, which are virtually identical, retained very high rates that quickly reached 1.00 in within-cluster sample sizes of $n_j = 50$ or larger. With small samples in the power-only conditions ($\pi = 0$; top row of [Figure 2.5](#)), the BMLM with a confident high-variability prior showed modest power reductions compared to MLM and other BMLM specifications, though it still outperformed all traditional approaches (top-right panel in [Figure 2.5](#)). These lower power rates likely reflect a misalignment between the prior and the data: the true between-cluster variability was zero (all true slopes are of the same magnitude) and so the high-variability prior inflated slope variability and widened credible intervals, increasing the chance of including zero. Nonetheless, power rates using this prior are still high (~ 0.80) even with very small sample sizes, and reached 1.00 by $n_j = 50$.

Cluster count had only a limited influence on power, with more clusters providing slightly higher power. Overall, both MLM and most BMLMs maintained near-perfect power across different null proportions, with only minor declines under higher proportions of null clusters, small effects, and cluster sizes below $n_j = 50$. These lower power rates reflect stronger shrinkage and wider credible or confidence intervals in small samples; because the overall mean was closer to zero when effects were small and many clusters were null, non-null slopes were more likely to include zero, which led to lower detection rates of true effects.

Discussion

The notion that Bayesian methods offer a fundamental solution to the multiplicity problem has been around for decades. Many proponents (e.g., Berry & Hochberg, 1999; Gelman et al., 2012; Sjölander & Vansteelandt, 2019) argue that Bayesian hierarchical modeling reframes the multiplicity problem by integrating information across parameters through partial pooling, rather than by applying post-hoc error-rate adjustments. In this view, shrinkage induced by priors and the hierarchical structure can regularize extreme estimates and mitigate spurious findings, reducing the perceived need for post-hoc multiplicity adjustments. Yet, most demonstrations are theoretical and restricted to complete-null conditions where all true cluster effects are zero. Far less is known about how these claims hold in partial-null settings—which are likely more realistic in psychological research—or how prior choices interact with the shrinkage mechanism of partial pooling.

This study had three aims: (1) to attach an empirical backing to the theoretical propositions of addressing the multiplicity problem with Bayesian hierarchical modeling, (2) to disentangle the contributions of shrinkage through partial pooling from those of prior specification and Bayesian estimation, and (3) to compare the performance of BMLM, MLM, and traditional approaches beyond a complete-null case and across mixed scenarios where null and non-null effects coexist.

Summary of Findings

Consistent with theoretical expectations, BMLM performed exceptionally well under complete-null conditions and outperformed all other approaches, whereas MLM performed, on average, similarly to FWE and FDR. Almost all BMLM models yielded familywise Type I error

rates well below the nominal level, supporting the idea that shrinkage toward a null or near-null population mean can, in fact, protect against spurious findings almost completely. The only exception occurred when a prior heavily favoured strong between-cluster variability, which slightly weakened pooling and inflated the rate of false positives to levels resembling unadjusted mean comparisons.

In partial-null conditions, however, the picture was less favourable for BMLM. Across a wide range of priors, BMLM maintained relatively high and stable familywise error rates, in contrast to frequentist MLM, where error rates declined noticeably lower than other approaches with larger effects or sample sizes. This stability in BMLM error rates reflects the interaction of shrinkage, priors, and Bayesian estimation: because posterior estimates of variance components—which directly affect the degree of shrinkage—are also determined by their priors, shrinkage was much less sensitive to changes in sample size, and null clusters remained drawn toward a nonzero fixed effect.

In addition, in most of our simulation conditions, BMLM often produced tighter credible intervals which were much less likely to include zero than MLM confidence intervals, even if the slope estimate only slightly differed from zero. These narrow intervals, paradoxically, worked against the error-controlling benefits of shrinkage. Partial pooling shrinks slope estimates toward the overall mean, but if that mean is nonzero, even a slight pull can make the estimate of a true null cluster statistically significant when its interval barely misses zero. Put simply, in partial-null conditions, narrow credible intervals impede the benefits of shrinkage to reducing familywise Type I error rates: if shrinkage draws null clusters toward a nonzero mean, and intervals are too narrow, false rejections become more likely.

Power rates showed ceiling effects with large effects: all approaches, including MLM, BMLM, and traditional corrections, achieved perfect detection. With small effects, both MLM and BMLM reached perfect or near-perfect power once within-cluster sizes reached $n_j = 50$. In smaller samples, power remained high and virtually identical across the two approaches, with one exception: under no-null conditions, a BMLM with a strongly high-variance prior yielded slightly lower power than other prior specifications and MLM, though it still outperformed conventional multiplicity corrections and no adjustment.

Overall, both BMLM and MLM showed similar, very high power rates across all conditions. With respect to familywise Type I error rates, however, the two approaches showed opposite patterns: BMLM had very low error rates in complete-null conditions but high rates in partial-null conditions. MLM, on the other hand, kept error rates low in partial-null conditions when sample sizes or effect sizes were large, but in complete-null conditions, MLM performed about the same as traditional corrections. For example, in complete-null conditions with a low cluster count, MLM produced slightly higher error rates than FDR or FWE (around .12), whereas with a high cluster count, its rates dropped and were closer to .025.

Beyond Error Rates: Effect Sizes, Precision, and Just a Touch of Context

Indeed, BMLM and MLM both showed high familywise Type I error rates in many partial-null conditions, but these results need context. First, the familywise error rate is the probability of making *at least one* false rejection across a set of tests. And while observing a single Type I error still counts towards the familywise rate, that was often exactly the case; in the partial-null conditions, MLM and BMLM usually exhibited a very small number of Type I errors. Thus, while familywise error rates are high, the *per-comparison error rate*—the

probability of rejecting a true null by chance—was nearly zero for MLM and at the nominal level for BMLM. These results are presented in [Appendix B](#).

Second, although both BMLM and MLM often incorrectly rejected the null hypothesis in partial-null conditions, the actual bias in slope estimates was minor; that is, even when confidence or credible intervals excluded zero, the corresponding point estimates and posterior means were still very close to the true value of zero. This finding is a good example of why it is wise to focus on the magnitude of the effect and its uncertainty rather than relying solely on statistical significance tests. Confidence bands that hardly miss zero may be flagged as “significant,” but if the estimates themselves are practically indistinguishable from null, the substantive conclusion should be that little or no effect exists. The emphasis should be put on quantifying the effects and their precision; looking at the estimates or posterior means and their confidence or credible intervals gives a more informative picture of what the model is telling us than reducing our conclusions to a mere binary decision.

A further point concerns how BMLM should be evaluated in relation to familywise Type I error rates. By design, Bayesian inference is not concerned with long-run (i.e., repeated sampling) error rates in the same way as frequentist NHST. Instead, Bayesian statistics focuses on posterior distributions that quantify uncertainty about parameters; Bayesian inference provides the probability that a clinic-specific effect exceeds (or does not exceed) a particular threshold (e.g., the minimally meaningful effect size of treatment) given the observed data. If one is worried about making a false positive—say, whether the treatment in clinic 42 is actually effective—the relevant evidence is the posterior probability of a clinically meaningful effect in clinic 42, conditional on the full model and the data. Importantly, this probability is interpreted in light of the full set of clinics analyzed. For example, if 100 clinics were examined in total and

only clinic 42 appears to show a meaningful effect, that context is part of the inferential picture; the joint posterior distribution reflects the shared structure across clinics and the uncertainty introduced by considering them simultaneously. The relevant question, then, is not how often such intervals would exclude zero (or a clinically meaningful effect) across repeated samples, but rather what is the posterior probability that the effect meaningfully differs from zero given the data (which includes the remaining 99 clinics that do not show a meaningful effect). In that sense, Bayesian inference does not ignore multiplicity; instead, it makes both the number of comparisons and the uncertainty surrounding each effect the grounds on which the inference from the posterior distribution stands. The reader is therefore not “shielded” with post-hoc correction from the researcher’s degrees of freedom concerning multiplicity, but rather has access to the full picture of estimates and must interpret any single clinic in the context of all others examined.

From this perspective, it is not completely unexpected that BMLM does not always lend itself well to frequentist concepts like familywise Type I error rate. Still, using credible intervals as a stand-in for significance testing—as we did here—makes the comparison possible, and what emerges is that BMLM can “look bad” under partial-null conditions simply because its narrow intervals increase the chance of excluding zero. Ultimately, high familywise error rates in partial-null conditions with BMLM are not necessarily a failure of the method, but rather a byproduct of applying a frequentist decision rule (reject/not reject) to a Bayesian framework, or looking at a Bayesian model through a frequentist’s lens. In practice, researchers using BMLM should place less weight on whether an interval excludes zero and more weight on the size and uncertainty of the effect, which is where Bayesian estimation is most informative and what it is designed for.

More generally, familywise Type I error is a convenient and informative benchmark because it easily captures well the heart of the multiplicity problem, the inflation of false positives that occurs as the number of simultaneous tests increases. However, error management is more than just deflating that ballooning rate. Incorrect conclusions are inevitable; they are a natural, and perhaps even necessary, part of the scientific process—no single procedure or approach can prevent them. Instead, a successful method provides a theoretically sound and justifiable framework for managing those errors. Such methods draw information from parallel tests and the broader context, balance error and power, and offer useful, interpretable insights about the parameters of interest. From this perspective, familywise Type I error is just one lens through which we view performance rather than the ultimate or sole criterion of success. The value of a method lies not only in the numerical control of, or matching, some arbitrary error rate but also in its theoretical coherence and its ability to yield meaningful and trustworthy inferences. In that sense, both MLM and BMLM offer principled, model-based ways of regulating false discoveries through partial pooling and the incorporation of prior beliefs—approaches that neither simply “fix” p values and confidence intervals mechanically nor suppress error inflation artificially, but instead integrate information across tests and deliver more defensible and scientifically useful conclusions.

Another perspective for interpreting these findings is to reconsider where the responsibility for addressing multiplicity should lie. The Bayesian framework does not eliminate the problem so much as it shifts the responsibility to the reader. Consider the treatment of priors. Priors are not simply “correct” or “incorrect”—they reflect assumptions—and each reader is free to evaluate whether those assumptions are reasonable for their purposes. Nonetheless, priors can be both a strength and a liability. Well-chosen priors can improve estimation and stabilize

inference, but poorly chosen ones can just as easily pull results away from the truth. Either way, prior assumptions can meaningfully shape conclusions. For that reason, prior selection should be intentional, transparent, and carefully justified, and it is often recommended to be accompanied by prior predictive checks and sensitivity analyses (Alter et al., 2025; Kruschke, 2015, 2021).

Similarly, perhaps there is no “right” and “wrong” treatment of multiple testing. Instead of blindly applying adjustments by default, researchers should consider their options deliberately and meticulously, with clear and explicit justification. Researchers should also report sensitivity analyses using different adjustments to evaluate how results change under different multiplicity treatments. Researchers can also perform “multiplicity predictive checks,” where they can simulate data and test whether their adjustment yields expected and sensible error–power trade-offs for their study design and assumptions.

Perhaps even a better resolution is to recognize that the multiplicity problem does not have to be the researcher’s burden alone; rather than the researcher solving the problem for the reader, results should be presented transparently so that readers are free to apply their own adjustments (if they so choose) to the set of hypotheses most relevant to their questions or interests. Reporting estimates, confidence or credible intervals, and posterior distributions or unadjusted p values with sufficient precision allows readers to form their own hypotheses and apply whichever correction, if at all, and conclusions they deem appropriate. Thus, we should keep in mind that multiplicity is not just a statistical problem that falls on the analyst, but also a philosophical and interpretive one which should be shared with the reader. Importantly, careful and transparent reporting is the bedrock of any principled solution, Bayesian or frequentist.

Limitations and Future Directions

Several caveats, design choices, and constraints should be noted when interpreting these findings. Similar to Study 1, we focused on two-level models with continuous outcomes and balanced designs (equal cluster sizes and equal group allocation), which ensured that the fixed slope represented the pure within-cluster effect. Many applications depart from these assumptions. Unequal cluster sizes, missing data, and covariates are common, and each can affect shrinkage. We also did not vary some other structural features: random intercept–slope correlations, within-cluster variance, and number of levels (three-level or crossed designs). These modeling characteristics, too, warrant a systematic evaluation.

As briefly discussed earlier, our evaluation metric was deliberately frequentist: we determined whether a credible interval included 0 as a decision rule to compute familywise Type I error and power within BMLM. This metric makes comparisons to Study 1 and to standard controls (FWE, FDR) transparent, but it is not how Bayesian inference is designed to operate. Future work may wish to include Bayesian-centric criteria such as posterior sign and magnitude errors (Type S/M errors), posterior regions of practical equivalence (ROPE) .

We varied several prior types on the fixed-effect and slope variance parameters, including shrinkage, the *brms* defaults, skeptical, and confident priors. We did not, however, specify priors for other parameters in the model, such as the random-effects correlation or the residual variance parameters; these were estimated with the default weakly informative settings. Furthermore, we did not intentionally manipulate the prior–data agreement. For example, in some conditions, the priors aligned with the data-generating process (e.g., high τ_{11} under strong heterogeneity), which likely curbed how different the posteriors could be. Future work should more deliberately vary the alignment of priors with the data to test BMLM more fully. We also tested each prior on its

own for feasibility, but did not cross all prior types with one another. Some combinations—for instance, pairing an extreme τ_{11} prior with a γ_{10} shrinkage prior like the horseshoe—might have produced different results. However, including all combinations would have expanded the design substantially, which was not computationally feasible. Exploring these combinations, as well as the impact of correlation priors, remains a useful direction for future work.

Similarly, computation abilities constrained our design to some extent. Each condition fit nine BMLM variants to 1,000 replications to keep the study feasible. These replication counts yield stable patterns, but a higher number of replications would have more precisely converged on the true error or power rate. Still, we trust that these estimates are valid and reliable: the patterns for MLM and traditional methods closely paralleled those from Study 1, which was based on 5,000 replications, and the BMLM models were themselves fit with reasonably long MCMC runs. Together, these features provide confidence that our estimates are sufficiently accurate to capture the main contrasts among approaches.

A related limitation concerns how null clusters were generated in partial-null conditions. In the present design, each null cluster was assigned a true treatment effect of zero—a deliberate choice that allowed us to unambiguously identify which clusters were truly null in the population and, in turn, correctly classify false positive rejections.⁴ This design does, however, represent a structurally distinct generative process from the unimodal normal distribution that both frequentist MLM and BMLM assume governs all cluster-specific effects. How much this difference between the true data-generating mechanism and model assumptions affects performance remains an open question. For instance, heavy-tailed random-effect distributions—

⁴ However, individual observations within those clusters were still subject to random sampling variability, so estimated treatment effects were not perfectly null; they scattered around zero as a function of within-cluster sample size and residual variance.

such as Student-*t* random effects—could potentially offer some robustness by being less aggressive in shrinking extreme cluster estimates, though they would not directly model the two-component structure. It is also worth noting that this limitation applies equally to frequentist MLM in Study 1.

Perhaps a more principled approach would be to apply shrinkage priors such as the horseshoe or spike-and-slab directly at the level of the random effects rather than the fixed effect, as done in this study. Doing so would allow the model to adaptively identify which clusters belong to a true null population. Indeed, this different modeling approach was not pursued in the present study, in part because it departs from the standard BMLM framework advocated by Gelman et al. (2012), whose multiplicity argument rests on the hierarchical structure, even with weakly informative priors, rather than explicit sparsity-inducing priors on the random effects. More fundamentally, however, applying shrinkage priors at the random-effect level shifts the inferential question from estimating the magnitude of cluster-specific effects—the primary goal in most applied psychological research—to classifying which clusters belong to a distinct, null population. This reframing may be well-suited for multiplicity control as an end in itself, but it sits uneasily with the broader aim of psychological research, where controlling Type I error is a safeguard rather than the target, and where describing the size and uncertainty of effects is usually more informative than rendering a verdict on whether they are perfectly zero. That said, this remains a compelling direction for future work, particularly in contexts where sparsity is a genuine and theoretically motivated expectation.

Conclusion

This study is one of the first attempts to empirically evaluate the extent to which Bayesian multilevel modeling can address the multiplicity problem. The picture that emerged is

mixed. Most BMLM models did extremely well in complete-null conditions, keeping false positives close to zero and much lower than unadjusted or standard correction methods, while still retaining high power. But, the story is different in partial-null conditions: shrinkage, together with very narrow credible intervals, raised BMLM familywise error rates much higher than MLM, FWE, and FDR. The takeaway is not that BMLM “solves” the multiplicity problem, but that it changes how we think about it. Instead of relying on frequentist-style reject/not-reject rules, we should focus on effect sizes and their uncertainty. BMLM is a useful and flexible tool, but it certainly cannot unburden us from all the fundamental issues shared with frequentist analyses. Its strength is in shifting the conversation away from binary decisions and error rates toward a more transparent way of presenting and discussing results, where multiplicity is treated more as an issue of interpretation and subjectivity, which is shared with the reader, rather than a decision that must be determined by the researcher with a single solution.

CHAPTER FOUR

GENERAL DISCUSSION

The present dissertation examined the multiplicity problem from a modeling perspective rather than a purely procedural one. Across two simulation studies, the goal was to evaluate the claim that multilevel modeling—frequentist and Bayesian—can mitigate inflation in Type I error rates when multiple related parameters are estimated simultaneously (e.g., Berry & Hochberg, 1999; Gelman et al., 2012). Rather than assuming that MLM or BMLM inherently resolve multiplicity concerns, the studies assessed how error rates and power change under varying design conditions, including differences in sample size, effect size, and prior specification. The results place important boundaries on these claims: MLM and BMLM can reduce false positive rates under certain structural conditions, but their effectiveness largely depends on the nature of the comparisons being made and the assumptions built into the model.

Study 1 demonstrated that frequentist MLM can mitigate error inflation through partial pooling—the sharing of information across clusters and tests—which acts as a natural form of regularization. Under many research conditions—such as in complete null or partial-null conditions with large sample sizes or effect sizes—MLM achieved familywise Type I error rates close to or even lower than the nominal levels while maintaining greater power than traditional correction methods. This finding empirically supports a phenomenon that was previously proposed theoretically: multiplicity control can emerge organically from modeling rather than from external adjustment rules applied to p values and confidence intervals. But, these results also reveal that this model-based approach has considerable boundaries, too, and how certain combinations of conditions—the ratio of true to null effects, the variability within and between

clusters, cluster size and, number of clusters—jointly determine the degree of shrinkage which influences how effectively multilevel models contain error inflation.

Study 2 extended this investigation into Bayesian territory. By incorporating prior information and Bayesian estimation, BMLM proved to be a more intricate mechanism of controlling multiplicity. Simulation results implied that priors could mediate how design factors influenced shrinkage and credible interval width, which in turn affected familywise Type I error and power rates. Under complete-null conditions, most prior specifications maintained familywise Type I error well below the nominal rates and outperformed both MLM and conventional adjustments. Under partial-null conditions, however, BMLM showed relatively stable but high error rates across sample sizes and effect magnitudes, which is the result of both prior-driven regulation of shrinkage and very narrow credible intervals that continued to pull null effects toward nonzero values even with a large cluster sample size. Prior specification had little effect in most cases under partial-null conditions; only in data-limited settings did a confident, high-variability prior on the between-slope variability meaningfully affect outcomes. These findings illustrate both the promise and limits of BMLM: although the hierarchical model can improve multiplicity control, its effectiveness depends mostly on the alignment of the prior specifications with the true parameter values.

The two studies offer converging evidence that multiplicity can, in some settings, be approached through modeling rather than mechanical adjustment. Multilevel frameworks—both frequentist and Bayesian—allow researchers to model similar or related hypotheses jointly and draw inferences that account for shared information among tests. To this extent, Type I error management becomes a property of the model itself, organically emerging from how information is structured and shared across tests rather than imposed externally as a corrective afterthought.

However, this should not be interpreted as a general solution. The extent to which partial pooling regulates false positives depends critically on the configuration of effects, the degree of heterogeneity, sample size, and prior assumptions. Under partial-null conditions or with strongly informative variance priors, model-based approaches may attenuate error inflation, leave it unchanged, or even exacerbate it. Thus, these multilevel frameworks do not eliminate the multiplicity problem; they shift it into the modeling features and assumptions. Researchers considering such approaches should treat them as provisional, assumption-sensitive strategies and accompany them with careful diagnostics, sensitivity analyses, and transparent reporting.

Overall, this research evaluates the extent to which MLM and BMLM can mitigate multiplicity-related error inflation under specific structural conditions. Across both studies, MLM and BMLM changed the balance between false positives and power by sharing information across parameters through partial pooling and, in the Bayesian case, also through prior specification. The extent of this change varied across simulation conditions: error behaviour depended on the magnitude of the true effects, the proportion of null parameters, and—in some of the Bayesian model conditions—the strength and form of the prior distributions. Accordingly, multilevel modeling should not be viewed as a general solution to the multiplicity problem, but rather as a modeling strategy whose operating characteristics depend on the assumptions embedded in the design and analysis. Ultimately, the usefulness of any inferential framework depends as much on transparency as on performance: results should be reported clearly enough for readers to evaluate the analyst's assumptions and decisions and share the responsibility of interpreting evidence.

REFERENCES

- Alter, U., Too, M. A. & Cribbie, R. A. (2025). Navigating the Bayes maze: The psychologist's guide to Bayesian statistics, a hands-on tutorial with R code. *International Journal of Psychology*, 60: e13271. <https://doi.org/10.1002/ijop.13271>
- Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health*, 86(5), 726–728. <https://doi.org/10.2105/AJPH.86.5.726>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bell, A., Holman, D., & Jones, K. (2019). Using Shrinkage in Multilevel Models to Understand Intersectionality: A Simulation Study and a Guide for Best Practice. *Methodology*, 15(2), 88–96. <https://doi.org/10.1027/1614-2241/a000167>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82(1–2), 215–227. [https://doi.org/10.1016/S0378-3758\(99\)00044-0](https://doi.org/10.1016/S0378-3758(99)00044-0)
- Bogdan, M., Ghosh, J. K., & Tokdar, S. T. (2008). *A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing*. In Beyond parametrics in

- interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen (Vol. 1, pp. 211-231). Institute of Mathematical Statistics. 10.1214/193940307000000158
- Bogdan, M., Chakrabarti, A., Frommlet, F., & Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 39(3), 1551–1579. <http://www.jstor.org/stable/23033607>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Bygren, L. O., Tinghög, P., Carstensen, J., Edvinsson, S., Kaati, G., Pembrey, M. E., & Sjöström, M. (2014). Change in paternal grandmothers early food supply influenced cardiovascular mortality of the female grandchildren. *BMC genetics*, 15(1), 12. <https://doi.org/10.1186/1471-2156-15-12>
- Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465–480. <https://doi.org/10.1093/biomet/asq017>
- Chalmers R, & Adkins M (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, 16(4), 248–280. doi:10.20982/tqmp.16.4.p248
- Chang, S., & Berger, J. O. (2020). Frequentist properties of Bayesian multiplicity control for multiple testing of normal means. *Sankhyā: The Indian Journal of Statistics, Series A* (2008), 82(2), 310–329. <https://www.jstor.org/stable/48735020>

- Chang, S., & Berger, J. O. (2021). Comparison of Bayesian and frequentist multiplicity correction for testing mutually exclusive hypotheses under data dependence. *Bayesian Analysis*, 16(1), 111–128. <https://doi.org/10.1214/20-BA1196>
- Clark, M. (2022). *mixedup: Miscellaneous functions for mixed models* (R package version 0.4.0). <https://m-clark.github.io/mixedup>
- Cribbie, R. A. (2003). Multiple comparison procedures: The p-value adjustment fallacy. *Canadian Journal of Experimental Psychology*, 57(3), 138–145. <https://doi.org/10.1037/h0087425>
- Cribbie, R. A. (2017). Multiplicity control, school uniforms, and other perplexing debates. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, 49(3), 159–165. <https://doi.org/10.1037/cbs0000075>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Cumming, G., & Calin-Jageman, R. J. (2016). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.
- Depaoli, S., Winter, S. D., & Visser, M. (2020). The importance of prior sensitivity analysis in Bayesian statistics: Demonstrations using an interactive Shiny app. *Frontiers in psychology*, 11, 608045. <https://doi.org/10.3389/fpsyg.2020.608045>
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293), 52–64. <https://doi.org/10.1080/01621459.1961.10482090>

- Evans, C. R., Williams, D. R., Onnela, J.-P., & Subramanian, S. V. (2018). A multilevel approach to modeling health inequalities at the intersection of multiple social identities. *Social Science & Medicine*, 203, 64–73. <https://doi.org/10.1016/j.socscimed.2017.11.011>
- Frane, A.V. (2020). Misguided Opposition to Multiplicity Adjustment Remains a Problem. *Journal of Modern Applied Statistical Methods*, 18, 28. <https://doi.org/10.1002/pst.1584>
- Gabry J, Češnovar R, Johnson A, Bröder S (2025). *cmdstanr: R interface to 'cmdStan'*. R package version 0.8.0, <https://discourse.mc-stan.org>, <https://mc-stan.org/cmdstanr/>
- Games, P. A. (1971). Multiple comparisons of means. *American Educational Research Journal*, 8(3), 531–565. <https://doi.org/10.2307/1161937>
- Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on psychological science: a journal of the Association for Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.
- Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189–211. <https://doi.org/10.1080/19345747.2011.618213>
- Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3), 373–390. <https://doi.org/10.1007/s001800000040>

- Green, M. A., Evans, C. R., & Subramanian, S. V. (2017). Can intersectionality theory enrich population health research? *Social Science & Medicine*, 178, 214–216.
<https://doi.org/10.1016/j.socscimed.2017.02.029>
- Hancock, G. R., & Klockars, A. J. (1996). The quest for α : Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research*, 66(3), 269–306. <https://doi.org/10.3102/00346543066003269>
- Hoffman, L., & Walters, R. W. (2022). Catching up on multilevel modeling. *Annual review of psychology*, 73(1), 659-689. <https://doi.org/10.1146/annurev-psych-020821-103525>
- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine* 2(8): e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
- Johnson, S. G. (2024). *The NLOpt nonlinear-optimization package*. <https://nlopt.readthedocs.io>
- Jones, K., Johnston, R., & Manley, D. (2016). Uncovering interactions in multivariate contingency tables: a multi-level modelling exploratory approach. *Methodological Innovations*, 9. <https://doi.org/10.1177/2059799116672874>
- Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparison wise Type I error control. *Psychological Methods*, 4(1), 58–69. <https://doi.org/10.1037/1082-989X.4.1.58>
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

- Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, 5(10), 1282–1291. <https://doi.org/10.1038/s41562-021-01177-7>
- Martinez Gutierrez, N. DeSouza, J., Alter, U., Bell, S., Beribisky, N., Farmus, L., Tu, E., & Cribbie, R. A. (2025). Multiplicity Control in Psychological Research: A Systematic Review of Applied Practices. [Accepted]. *Canadian Psychology / Psychologie canadienne*. <https://osf.io/ne79j/>
- Merlo, J. (2018). Multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) within an intersectional framework. *Social Science & Medicine*, 203, 74–80. <https://doi.org/10.1016/j.socscimed.2017.12.026>
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.
- Open Science Collaboration (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, 316(7139), 1236–1238. <https://doi.org/10.1136/bmj.316.7139.1236>
- Petry, N. M., Alessi, S. M., & Ledgerwood, D. M. (2012). A randomized trial of contingency management delivered by community therapists. *Journal of consulting and clinical psychology*, 80(2), 286–298. <https://doi.org/10.1037/a0026826>

- Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2), 5018–5051.
<https://doi.org/10.1214/17-EJS1337SI>
- Poole, C. (1991). Multiple comparisons? No problem! *Epidemiology*, 2(4), 241–243. Retrieved from <https://www.jstor.org/stable/20065673>
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.
- Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1), 43–46. <https://doi.org/10.1097/00001648-199001000-00010>
- Sandell, R., Svensson, M., Nilsson, T., Johansson, H., Viborg, G., & Perrin, S. (2015). The POSE study - panic control treatment versus panic-focused psychodynamic psychotherapy under randomized and self-selection conditions: study protocol for a randomized controlled trial. *Trials*, 16, 130. <https://doi.org/10.1186/s13063-015-0656-7>
- Saville, D. J. (1990). Multiple Comparison Procedures: The Practical Solution. *The American Statistician*, 44(2), 174–180. <https://doi.org/10.1080/00031305.1990.10475712>
- van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1), 1–26.
<https://doi.org/10.1038/s43586-020-00001-2>

- Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7), 2144–2162.
<https://doi.org/10.1016/j.jspi.2005.08.031>
- Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5), 2587–2619.
<https://doi.org/10.1214/10-AOS792>
- Sigal, M. J., & Chalmers, R. P. (2016). Play It Again: Teaching Statistics With Monte Carlo Simulation. *Journal of Statistics Education*, 24(3), 136–156.
<https://doi.org/10.1080/10691898.2016.1246953>
- Sjölander, A., & Vansteelandt, S. (2019). Frequentist versus Bayesian approaches to multiple testing. *European Journal of Epidemiology*, 34(9), 809–821.
<https://doi.org/10.1007/s10654-019-00517-2>
- Snijders, T. A. B., & Berkhof, J. (2008). Diagnostic checks for multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 141–175). Springer.
https://doi.org/10.1007/978-0-387-73186-5_5
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE Publications.
- Smith, C. E., & Cribbie, R. A. (2013). Multiplicity control in structural equation modeling: Incorporating parameter dependencies. *Structural Equation Modeling*, 20(1), 79–85.
<https://doi.org/10.1080/10705511.2013.742385>
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

- Wagstaff, A., van Doorslaer, E., Burger, R. (2019). SMS nudges as a tool to reduce tuberculosis treatment delay and pretreatment loss to follow-up. A randomized controlled trial. *PLOS ONE* 14(6): e0218527. <https://doi.org/10.1371/journal.pone.0218527>
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604. <https://doi.org/10.1037/0003-066X.54.8.594>
- Wilson, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons. *Psychological Bulletin*, 59(4), 296–300. <https://doi.org/10.1037/h0040447>

TABLES

Table 1.1

Summary of Manipulated Factors and Simulation Conditions in Study 1

Manipulated factors	Notation	Values
Within-cluster sample size	n_j	6, 12, 24, 50, 100, 250, 500
Number of clusters	J	8, 16, 32, 64, 128, 256
Effect size	b_j	0, 1, 2, 5
Proportion of clusters with true null effects	π	0 (power-only condition), 0.25, 0.5, 0.75, 1 (complete null condition)
Testing approach	Approach	J independent mean comparisons (unadjusted, Bonferroni, and FDR), MLM

Note. n_j represents the sample size for cluster j out of a total of J clusters. All clusters have the same sample size, n_j . Additionally, treatment and control group sizes are equal within each cluster, with each group comprising half of n_j (i.e., $n_j/2$).

Table 2.1

Summary of Manipulated Factors and Simulation Conditions in Study 2

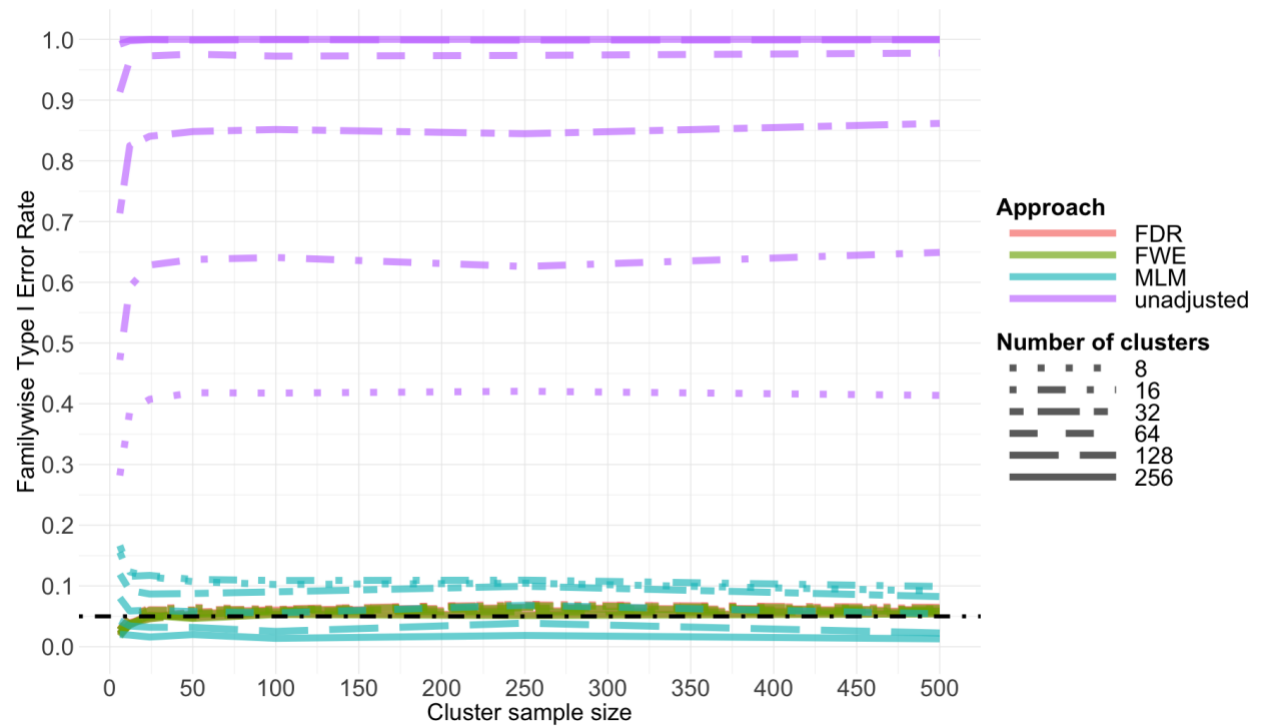
Manipulated factors	Notation	Values
Within-cluster sample size	n_j	10, 20, 50, 100, 250
Number of clusters	J	12, 48, 128
Effect size	b_j	0, 1, 5
Proportion of clusters with true null effects	π	0 (power-only condition), 0.25, 0.5, 0.75, 1 (complete null condition)
Testing approach	Approach	J independent mean comparisons (unadjusted, Bonferroni, and FDR), MLM, BMLM

Note. n_j represents the sample size for cluster j out of a total of J clusters. All clusters have the same sample size, n_j . Additionally, treatment and control group sizes are equal within each cluster, with each group comprising half of n_j (i.e., $n_j/2$).

FIGURES

Figure 1.1

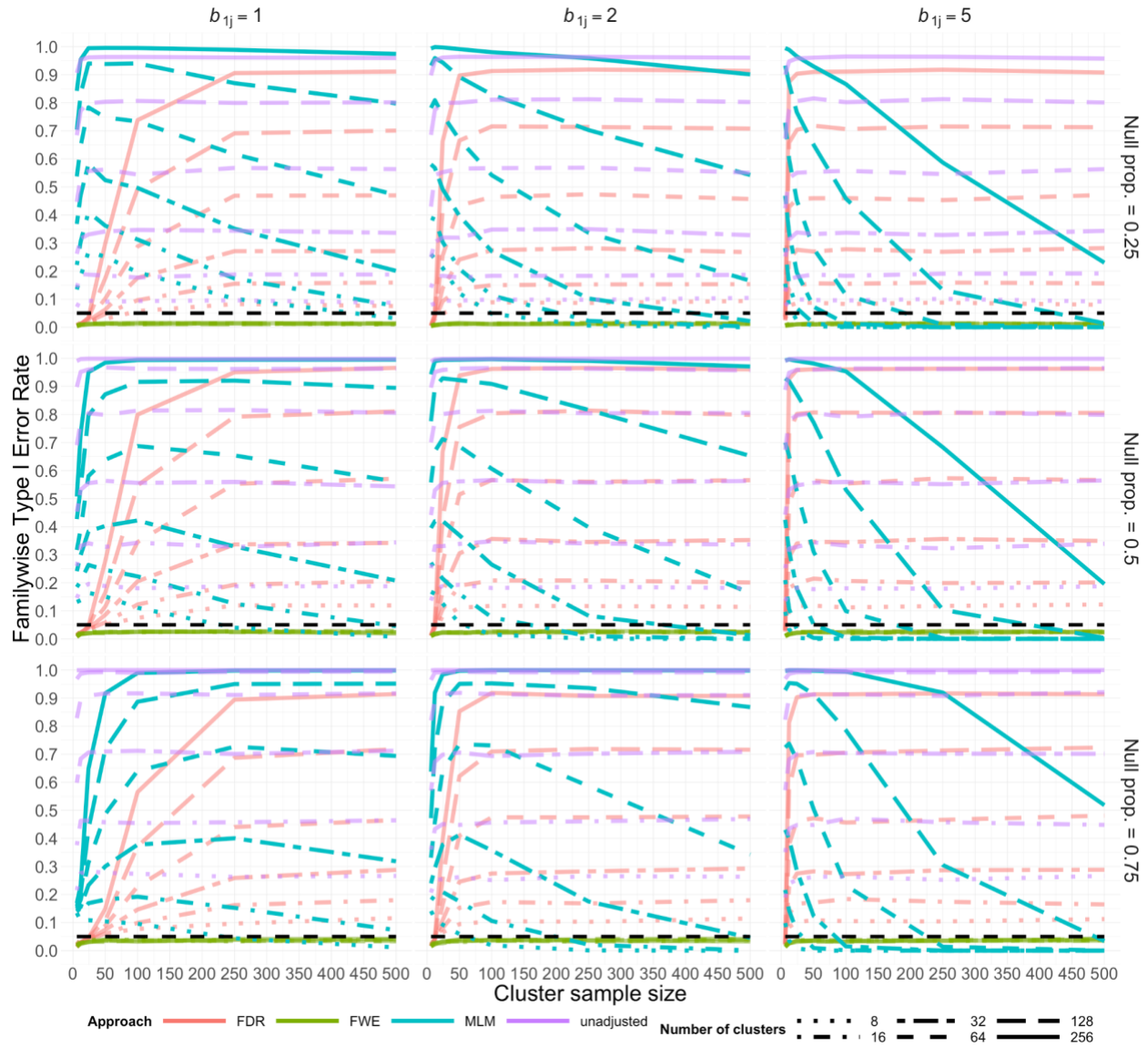
Familywise Type I Error Rates in Complete-Null Conditions in Study 1



Note. Results reflect familywise Type I error rates based on 5,000 simulations conducted under complete-null conditions, where all effects are null in the population (proportion of null effects = 1.00). Four methods for addressing Type I error inflation due to multiple significance tests are compared: false discovery rate (FDR), familywise error control (FWE), multilevel modeling (MLM), and unadjusted mean comparisons. Each method is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 8 (dotted) to 256 (solid). The horizontal black line indicates the nominal Type I error rate, $\alpha = .05$.

Figure 1.2

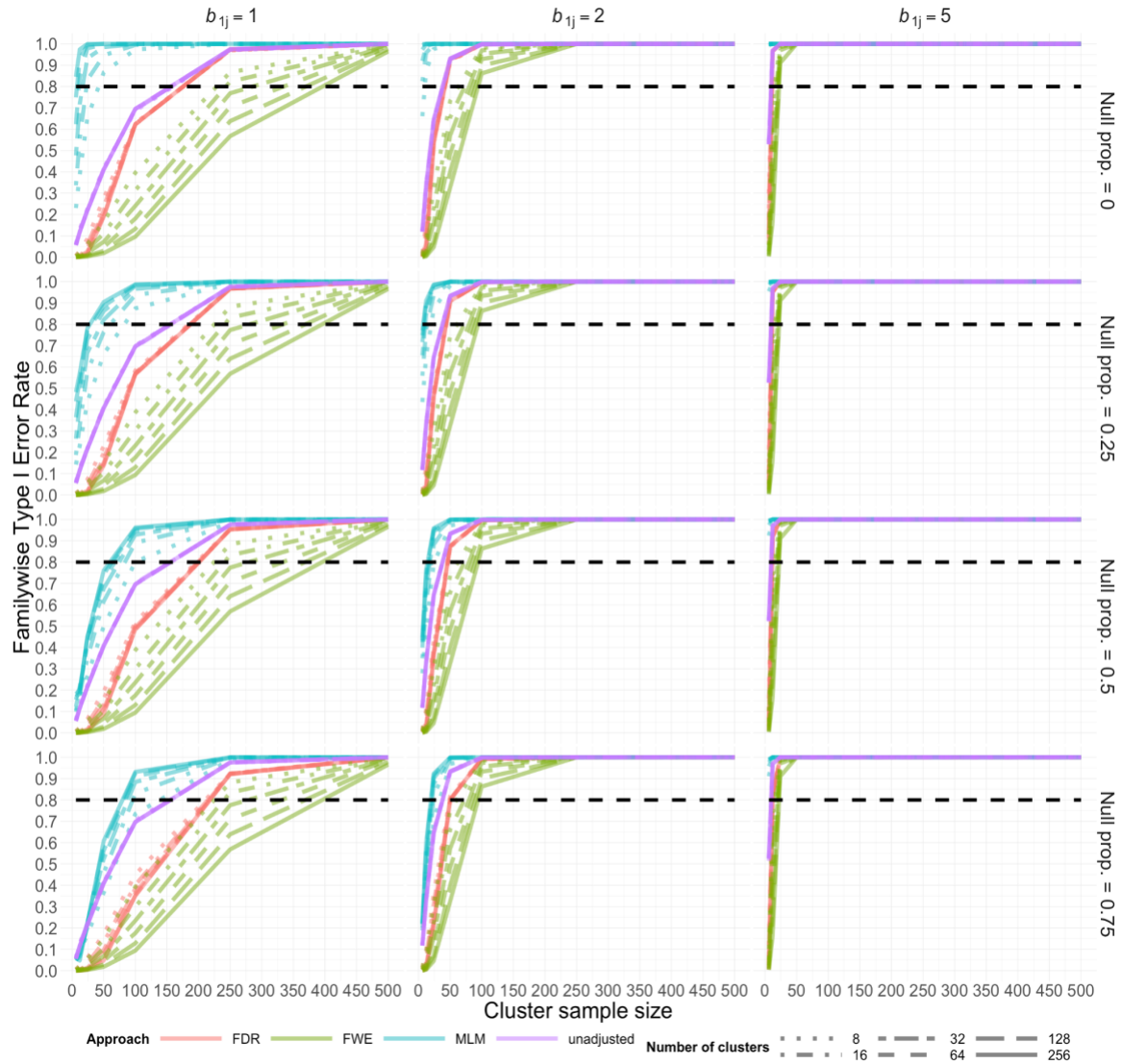
Familywise Type I Error Rates in Partial-Null Conditions in Study 1



Note. Results reflect familywise Type I error rates based on 5,000 simulations conducted under partial-null conditions, where null effects in the population are 0.25 (top row), 0.5 (middle row), or 0.75 (bottom row). Columns represent different true, population effects in non-null clusters/tests. Four methods for addressing Type I error inflation due to multiple significance tests are compared: false discovery rate (FDR), familywise error control (FWE), multilevel modeling (MLM), and unadjusted mean comparisons. Each method is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 8 (dotted) to 256 (solid). The horizontal black line indicates the nominal Type I error rate, $\alpha = .05$.

Figure 1.3

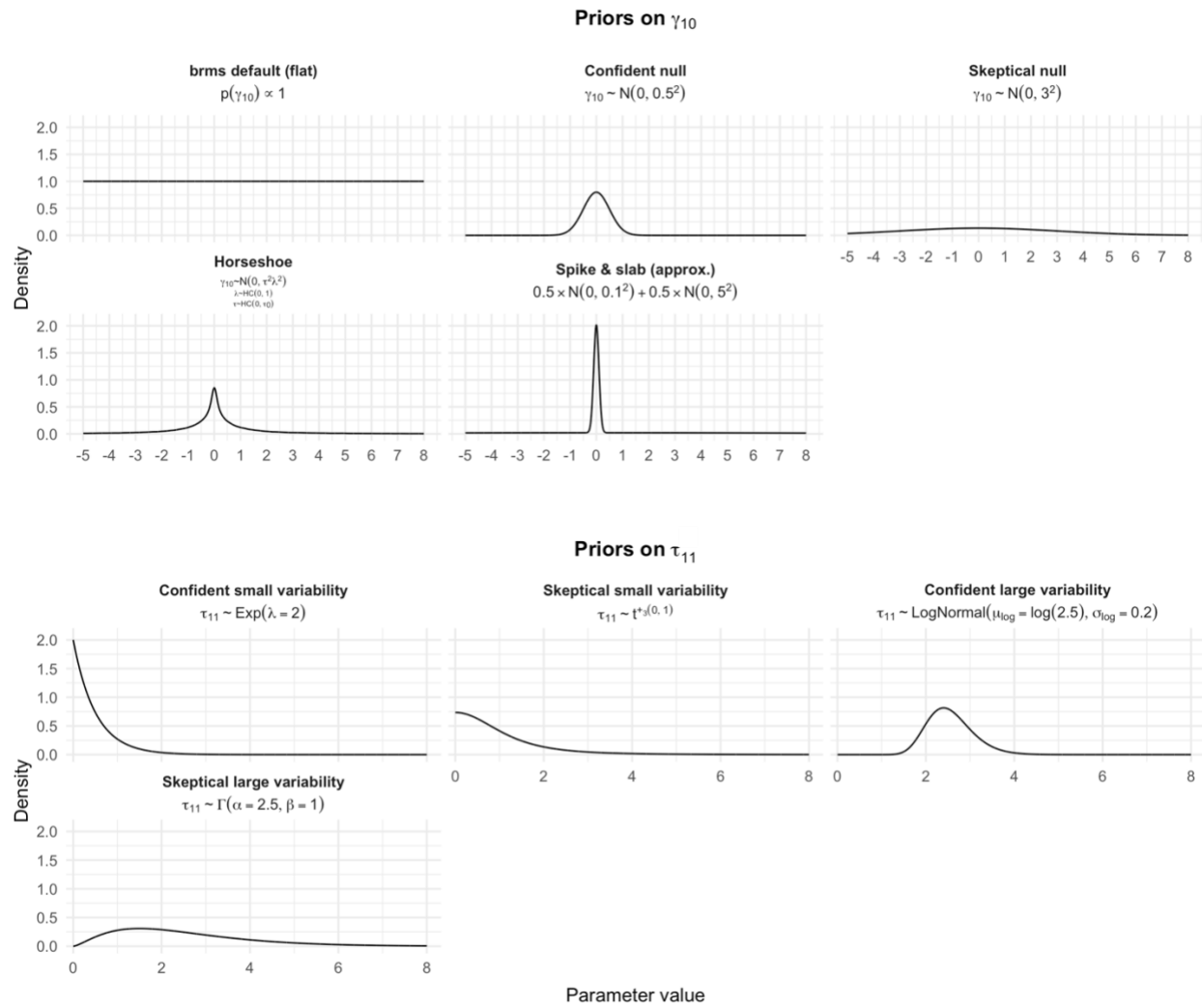
Power Rates in Study 1



Note. Results reflect power rates based on 5,000 simulations conducted under power-only conditions (top row), where there are no null effects, and under partial-null conditions, where null effects in the population are 0.25 (second row from the top), 0.5 (second row from the bottom), or 0.75 (bottom row). Columns represent different true population effects in non-null clusters/tests. Four methods for addressing Type I error inflation due to multiple significance tests are compared: false discovery rate (FDR), familywise error control (FWE), multilevel modeling (MLM), and unadjusted mean comparisons. Each method is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 8 (dotted) to 256 (solid). The horizontal black line indicates a common minimum power expectation in psychology, $1 - \beta = .80$.

Figure 2.1

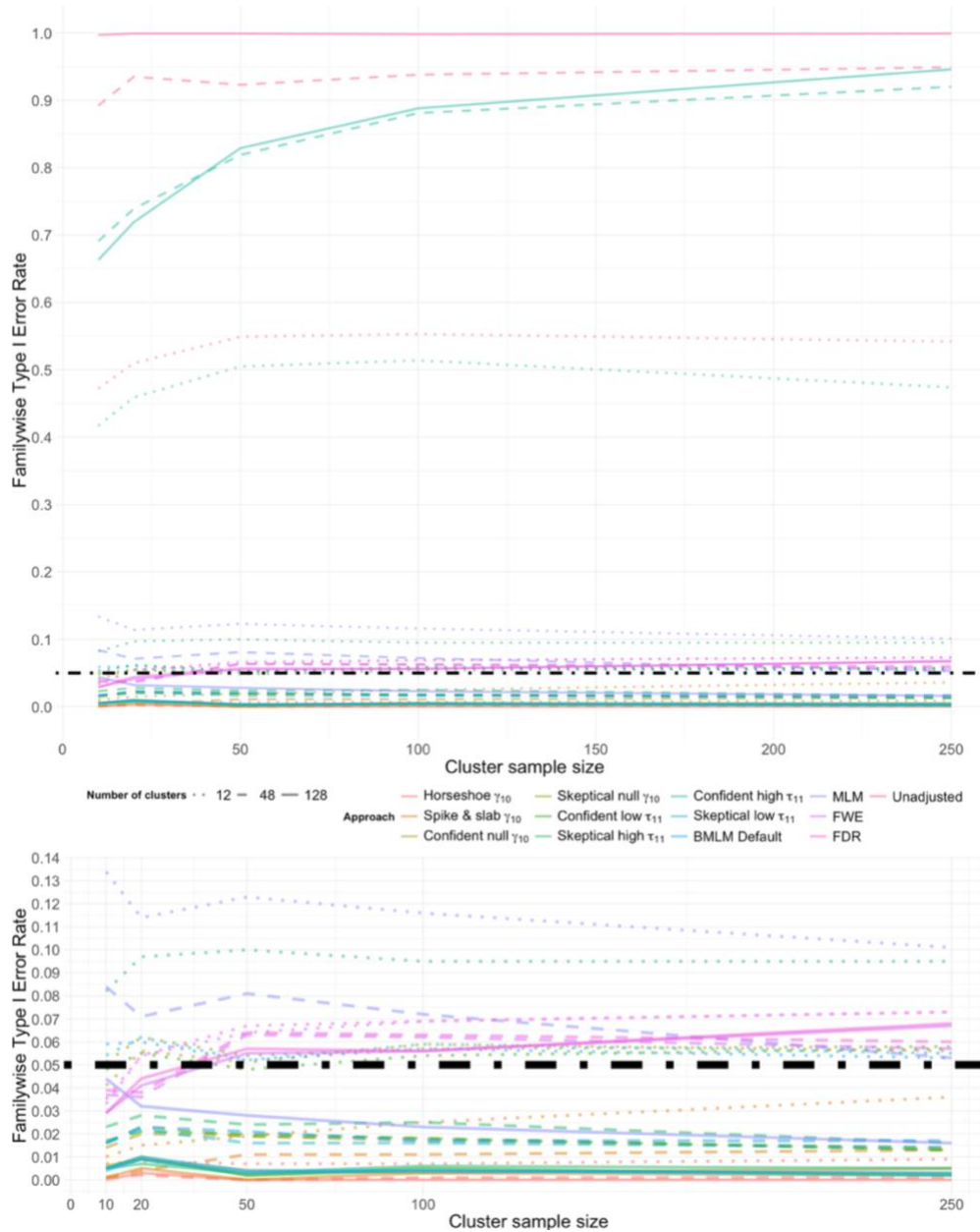
Simulation Prior Specifications in Study 2



Note. Prior distributions tested. Top two rows: priors on the fixed effect (γ_{10}). Bottom two rows: priors on the random slope SD (τ_{11}).

Figure 2.2

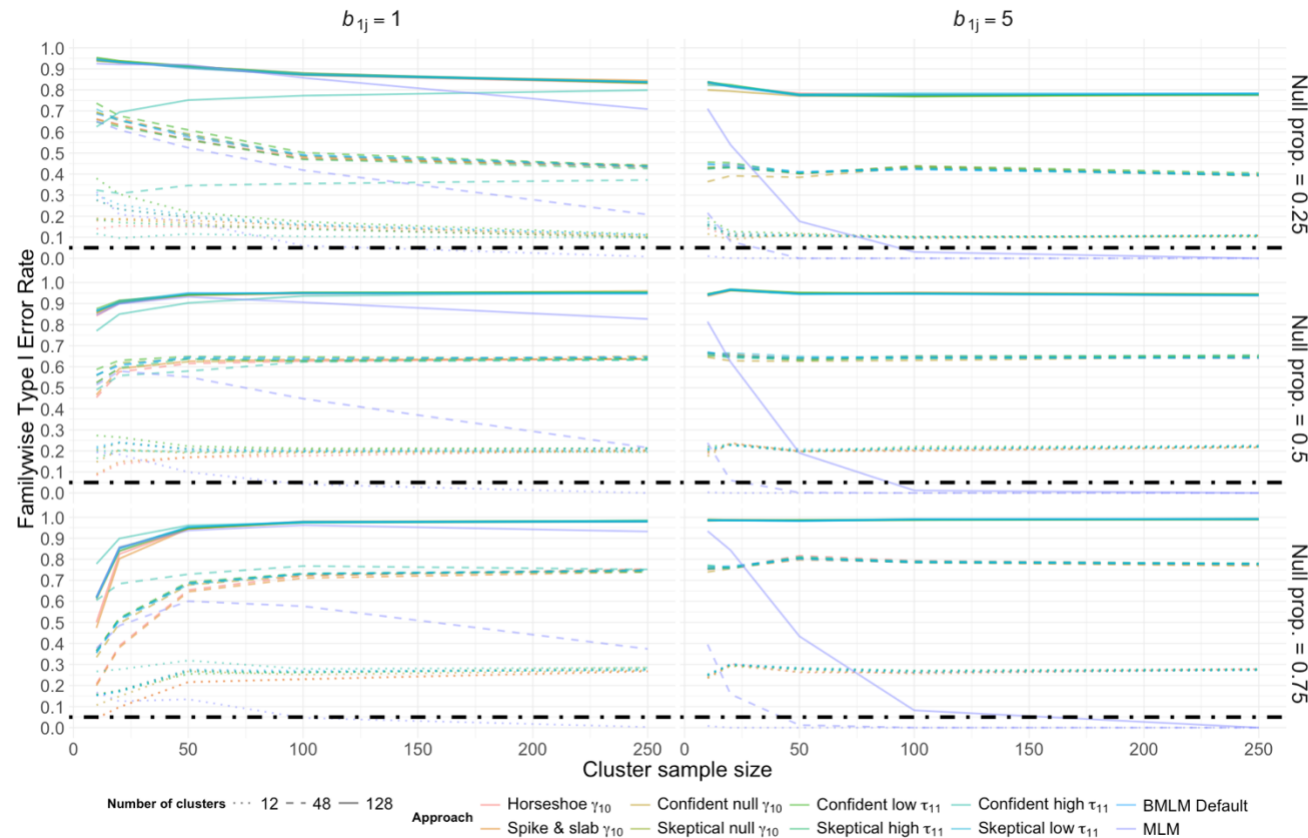
Familywise Type I Error Rates in Complete-Null Conditions in Study 2



Note. Results reflect familywise Type I error rates based on 1,000 simulations conducted under complete-null conditions, where all effects are null in the population (proportion of null effects = 1.00). The top panel displays the full range of results, whereas the bottom panel provides a close-up view of error rates near the nominal Type I error rate ($\alpha = .05$) to highlight differences among approaches in the region where most methods perform. Four methods for addressing Type I error inflation due to multiple significance tests are compared: unadjusted mean comparisons, false discovery rate (FDR), familywise error control (FWE), multilevel modeling (MLM), and numerous Bayesian multilevel modeling (BMLM) using nine different prior specifications on either the fixed-effect parameter, γ_{10} , or the between-cluster slope variability, τ_{11}^2 , which are presented in Figure 1. Each method is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 8 (dotted) to 128 (solid). The horizontal black line indicates the nominal Type I error rate, $\alpha = .05$.

Figure 2.3

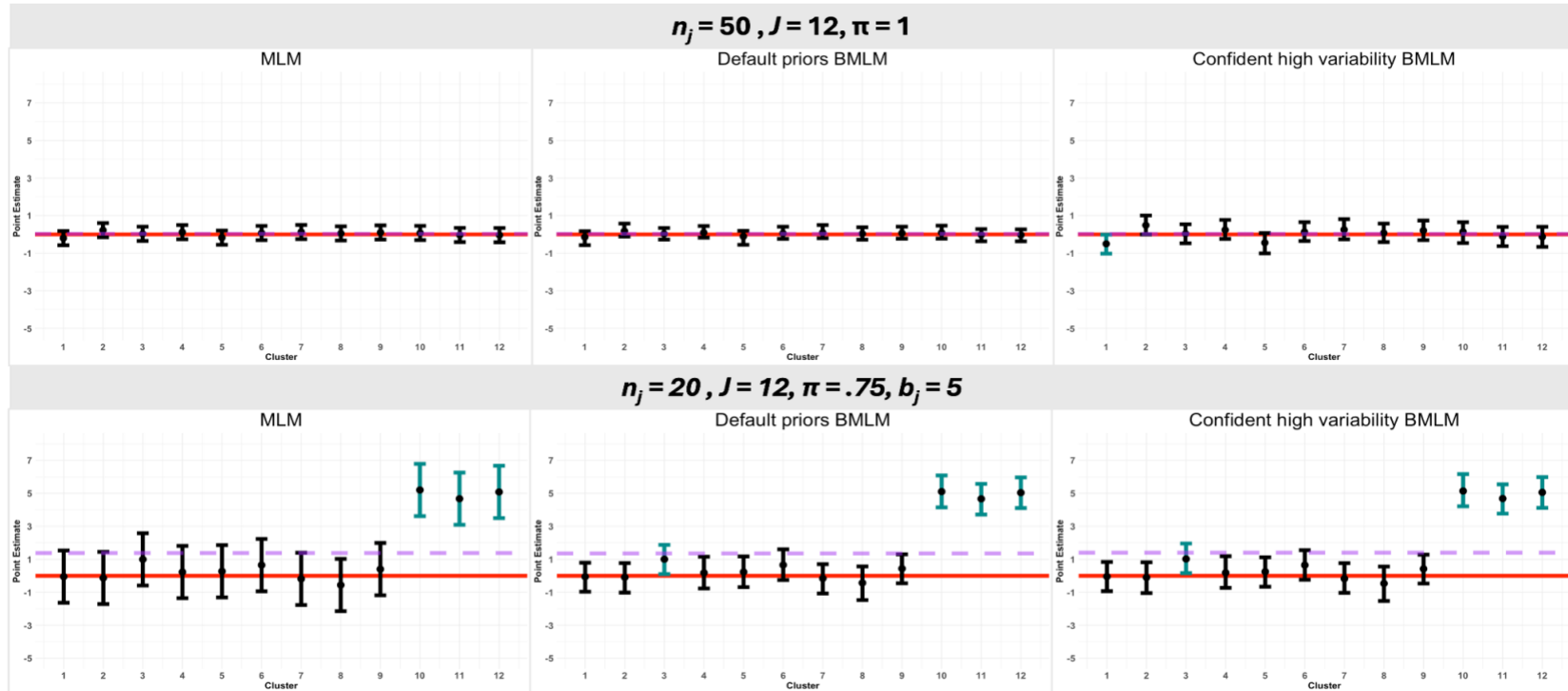
Familywise Type I Error Rates of MLM and BMLM with Different Prior Specifications in Study 2



Note. Results reflect familywise Type I error rates based on 1,000 simulations conducted under partial-null conditions, where null effects in the population are 0.25 (top row), 0.5 (middle row), or 0.75 (bottom row). Columns represent different true, population effects in non-null clusters/tests. Both frequentist multilevel modeling (MLM) and Bayesian multilevel modeling (BMLM) approaches are presented for addressing Type I error inflation due to multiple significance tests and compared. The BMLM approach included 9 models, each with a different prior assumption (as depicted in Figure 1). Each unique model is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 12 (dotted) to 128 (solid). The horizontal black line indicates the nominal Type I error rate, $\alpha = .05$.

Figure 2.4

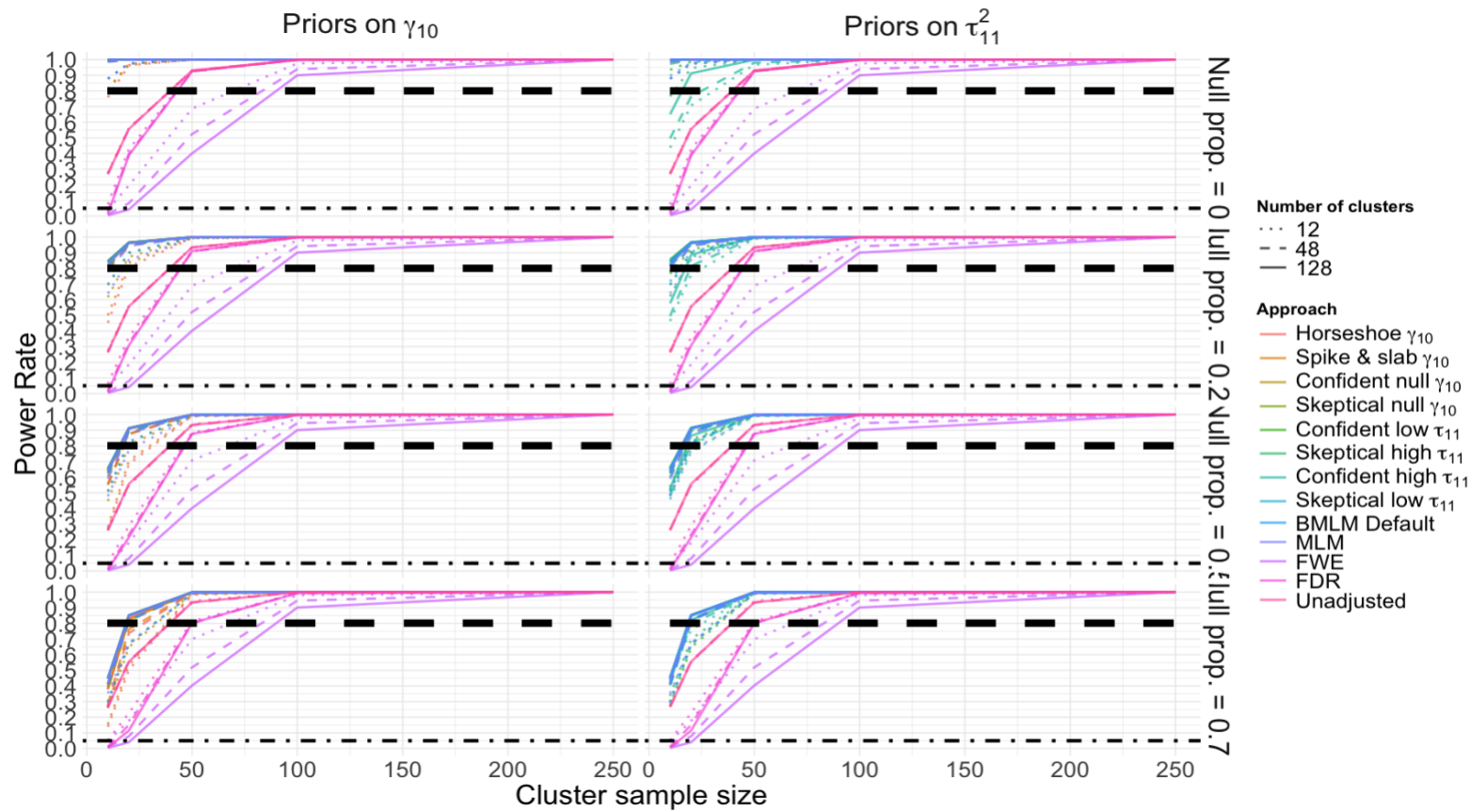
Illustrative Example of Cluster Effects Using MLM and Different Priors



Note. Cluster-specific slope estimates (MLM) or posterior means (BMLM) of a randomly generated dataset from a complete-null condition (top row) and partial-null condition (bottom row). Error bars represent the 95% confidence or credible intervals. Black error bars include zero and thus reflect failure to reject the null hypothesis $H_0: b_{1j}=0$ whereas teal error bars exclude zero and thus express statistically a significant result. The black circle at centre of each error bar illustrates the cluster-specific slope estimate.

Figure 2.5

Power Rates in Small Effect Sizes ($b_{1j} = 1$) Conditions in Study 2

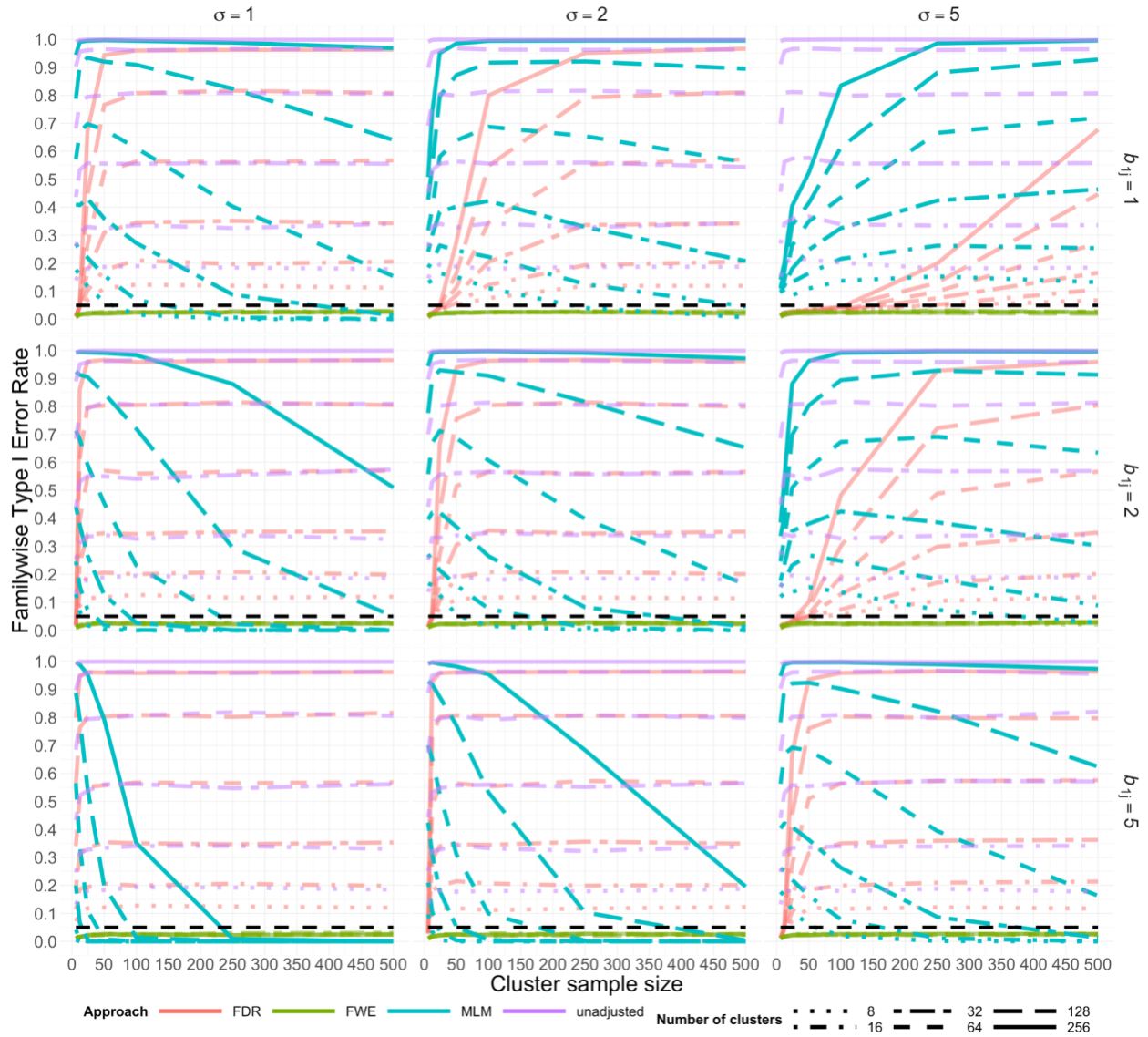


Note. Results reflect power rates based on 1,000 simulations conducted under power-only conditions (top row), where there are no null effects, and under partial-null conditions, where null effects in the population are 0.25 (second row from the top), 0.5 (second row from the bottom), or 0.75 (bottom row). Results presented are only from $b_{1j} = 1$ conditions. Columns represent different prior types, either on the fixed effect parameter, γ_{10} , or the between-cluster slope variance component, τ_{11}^2 , with false discovery rate (FDR), familywise error control (FWE), multilevel modeling (MLM), and unadjusted mean comparisons presented in both columns for convenience. Each method is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 12 (dotted) to 128 (solid). The horizontal black line indicates a common minimum power expectation in psychology, $1 - \beta = .80$.

APPENDIX A

Figure A1

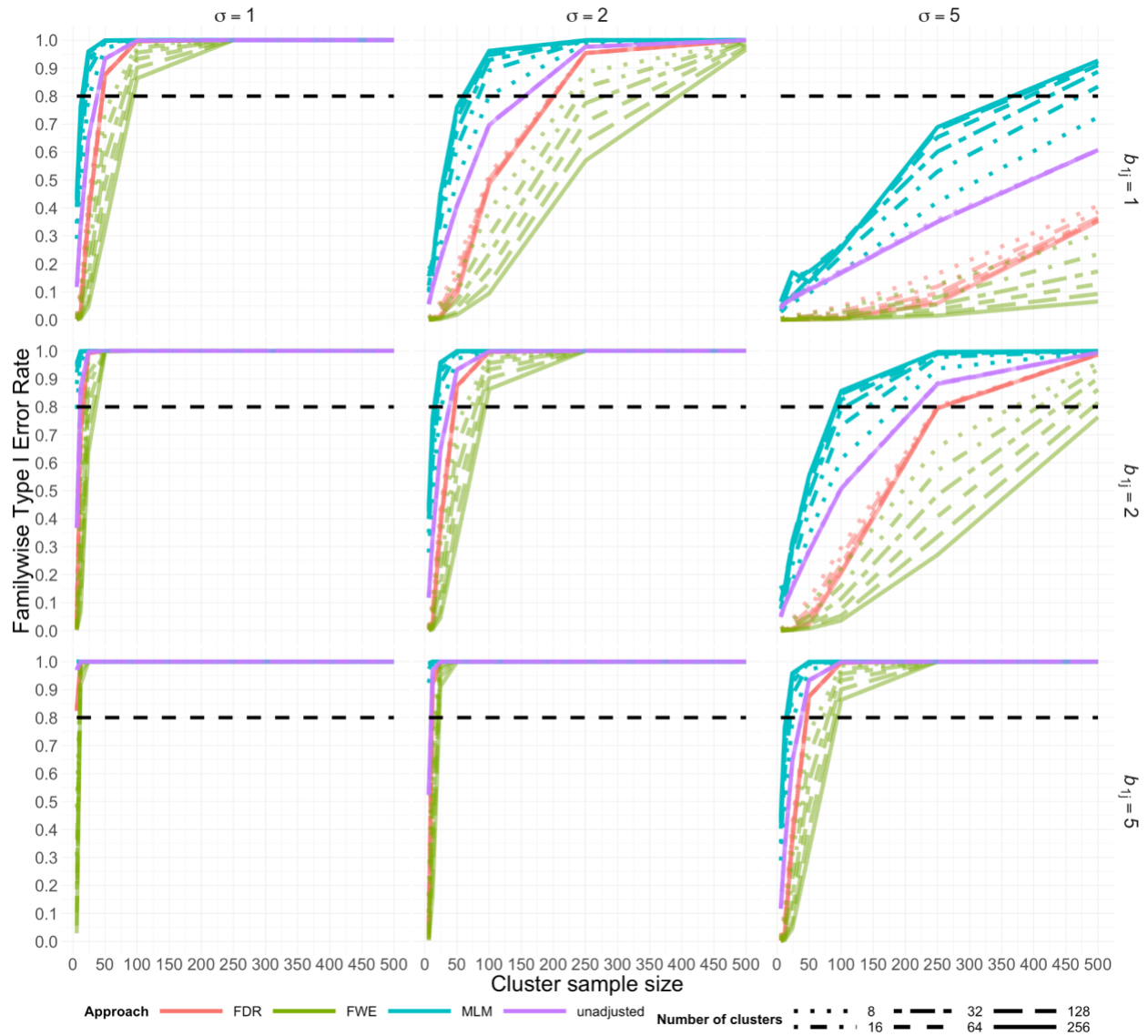
Familywise Type I Error Rates Across Within-Cluster Variability and Effect Size Levels in Half-Null Conditions (Proportion of Null Clusters is .5) in Study 1



Note. Results reflect familywise Type I error rates based on 5,000 simulations conducted under partial-null conditions, where null effects in the population are all 0.5. Columns represent different true, population within-cluster variability. Each row reflects a different effect size for the 50% of cluster with non-null effects. Four methods for addressing Type I error inflation due to multiple significance tests are compared: false discovery rate (FDR), familywise error control (FWE), multilevel modeling (MLM), and unadjusted mean comparisons. Each method is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 8 (dotted) to 256 (solid). The horizontal black line indicates the nominal Type I error rate, $\alpha = .05$.

Figure A2

Power Rates Across Within-Cluster Variability and Effect Size Levels in Half-Null Conditions (Proportion of Null Clusters is .5) in Study 1

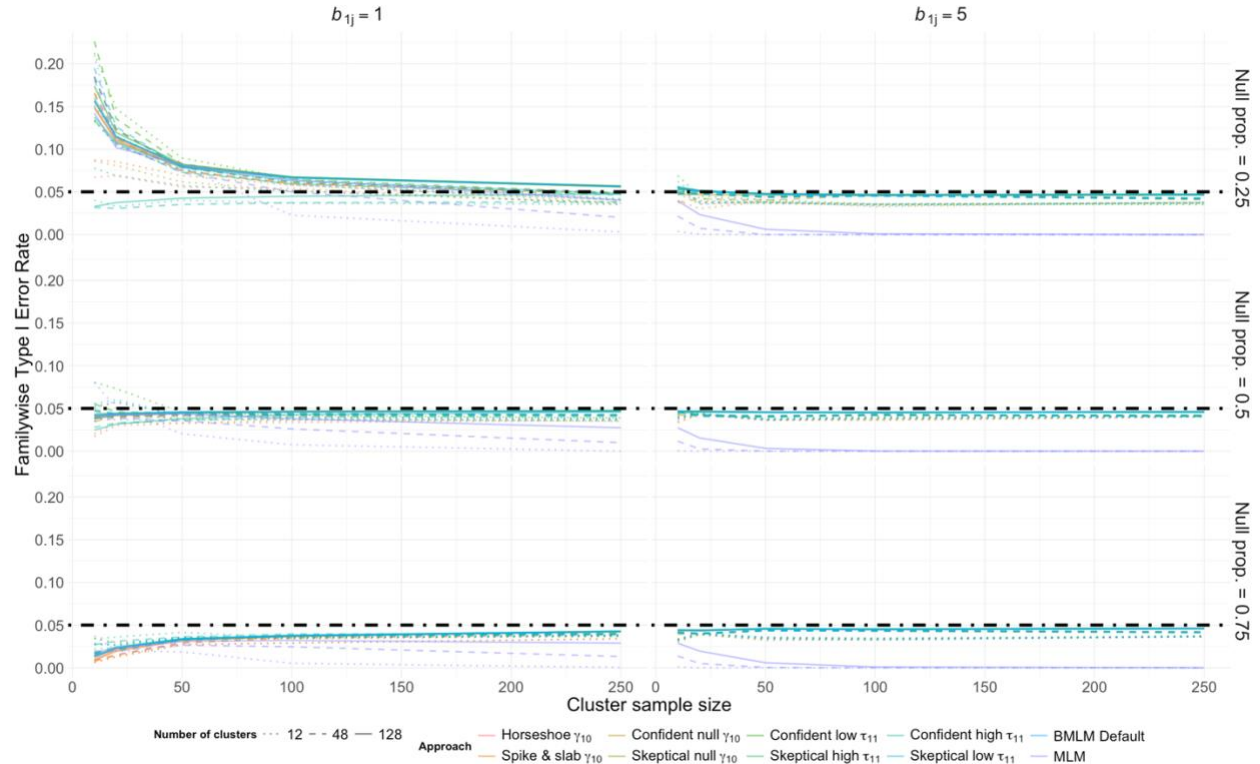


Note. Results reflect power rates based on 5,000 simulations conducted under partial-null conditions where null effects in the population are 0.5. Columns reflect different levels of within-cluster variability, whereas rows represent different true population effects in non-null clusters/tests. Four methods for addressing Type I error inflation due to multiple significance tests are compared: false discovery rate (FDR), familywise error control (FWE), multilevel modeling (MLM), and unadjusted mean comparisons. Each method is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 8 (dotted) to 256 (solid). The horizontal black line indicates a common minimum power expectation in psychology, $1 - \beta = .80$.

APPENDIX B

Figure B1

Per-Comparison Type I Error Rates of MLM and BMLM with Different Prior Specifications in Study 2



Note. Results reflect per-comparison Type I error rates (the probability of making a Type I error for a single hypothesis test) based on 1,000 simulations conducted under partial-null conditions, where null effects in the population are 0.25 (top row), 0.5 (middle row), or 0.75 (bottom row). Columns represent different true, population effects in non-null clusters/tests. Both frequentist multilevel modeling (MLM) and Bayesian multilevel modeling (BMLM) approaches are presented for addressing Type I error inflation due to multiple significance tests and compared. The BMLM approach included 9 models, each with a different prior assumption (as depicted in Figure 1). Each unique model is represented by a different colour. Line types denote the number of simultaneous tests (or clusters), increasing in density from 12 (dotted) to 128 (solid). The horizontal black line indicates the nominal Type I error rate, $\alpha = .05$.