The Dissertation Committee for Megha Joshi
Certifies that this is the approved version of the following Dissertation:

Cluster Wild Bootstrapping to Handle Dependent Effect Sizes in Meta-Analysis with Small Number of Studies

Committee:

Susan N. Beretvas, Supervisor

James E. Pustejovsky, Co-Supervisor

Seung W. Choi

Tiffany A. Whittaker

Elizabeth Tipton

Cluster Wild Bootstrapping to Handle Dependent Effect Sizes in Meta-Analysis with Small Number of Studies

by

Megha Joshi

Dissertation

Presented to the Faculty of the Graduate School
of The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2021

For Tweetie, Toshie, Yori, and Oreo

Acknowledgments

I am thankful for my two supervisors, Tasha and James. Tasha, you inspire me so much! You are an amazing teacher and mentor, and also a great leader. I am so grateful to you for always standing up for student concerns and for always showing compassion. James, I miss you! You have been an amazing mentor and have been so supportive of my work. Thank you for teaching me so much about research and coding and causal inference and meta-analysis. Thank you also for always believing in me! That has meant so much!

I am also thankful for my committee members. Dr. Choi, thanks for teaching me psychometrics. Your lectures were so in-depth and wonderful. I miss your jokes during classes. Dr. Whittaker, thanks for leading this department through a rough transition year. It must have been so difficult. Thanks for always being there for us. Thank you also for your teaching. I learned foundational statistical methods in your class that have helped me in my research and teaching till now. Dr. Tipton, thank you for your work! Your research is the basis of my dissertation.

Thank you Tweetie, Toshie, Yori, and Oreo. You have gotten me through many rough days. Thanks for sleeping on my laptop and books. Thank you for jumping over my head when I am trying to sleep. I always appreciate your love and playfulness.

Thank you to my family. I love you. Thank you to my parents for making the very difficult decision to leave their family and immigrate halfway across the world and start a new life here. I know life has been very difficult for you. I hope I can make you proud. Thank you, Dad, for your feedback on my dissertation drafts and for explaining many theorems to me over the years. Thank you, my brother Jay, for sharing your cats with me. Thank you also for being in Austin. You made my grad school years much more bearable and less lonely. Thank you also for all your help with running my simulation.

Thank you, Danny. I cannot imagine going through graduate school without you. I am grateful that we got to work on homework together, play video-games, work out over Skype, and volunteer to get out the vote. Bethany and RAZ, thank you for always looking out for me! Bethany, thanks for all the game nights and work-out sessions. Thank you, RAZ, for being my neighbor! Getting to know you both over the past few years has been so wonderful! Thank you, Jihyun, for helping me brainstorm

ideas on missing data and simulation codes. Thank you, Man, for hot pot! I miss you so much! Thank you also for all the great feedback on my work. Thank you to Young Ri, Suhwa, and Sook Hyun for helping me get through so many classes. It has been so great to hang out with you all. Thank you, Pierce! It has been a lot of fun trying to figure out different code and functions with you! Thank you, Molly, for all the dinner and movie nights! Thank you to Gleb, Melissa, Rich, Danny S, Chris, and Kejin for your advice on surviving graduate school, qualifying, dissertations, and career search! Thank you, Sangdon and Luping, for all your help and advice and friendship. This journey would not have been as fun without you all!

Thank you to Danny, Man, Young Ri, and Sangdon for your feedback on my dissertation!

Thank you to Lisa for your friendship! You've been there for so much! I am so grateful you sat next me in class that one day so many years ago. Thank you, Susu, for being my wonderful stats TA in undergrad. Thanks also for telling me about quantitative psychology. I would have never started this journey without you. Thank you, SangSuk, for always encouraging me to keep going. You were the first person to teach me how to analyze data and to code. Thank you for your mentoring and your friendship.

Thank you, Saleem. I love you so much! You have brought so much positivity and love into my life. Thank you also for helping me run my simulation. Thank you always for your kindness and thoughtfulness!

Abstract

Cluster Wild Bootstrapping to Handle Dependent Effect Sizes in Meta-Analysis with Small Number of Studies

Megha Joshi, Ph.D.

The University of Texas at Austin, 2021

Supervisors: Susan N. Beretvas and James E. Pustejovsky

Meta-regression models work under the assumption that there is only one effect size estimate per study and that the estimates are independent. However, meta-analytic studies in education and social sciences often contain multiple effect size estimates per primary study, leading to dependence in the estimates. Furthermore, meta-analytic studies can include effect sizes from multiple studies conducted by the same lab or investigator, which can also create dependence in the effect sizes. The increasingly popular method to handle dependence, robust variance estimation (RVE), can result in inflated Type I error rates when the number of studies is small. Small sample correction methods for RVE have been shown to control Type I error rates adequately but have been shown to be possibly conservative, especially for tests of multiple-contrast hypotheses. In this dissertation, I examined an alternative method, cluster wild bootstrapping, which has been examined in the econometrics literature but has not been examined under a meta-analytic framework. The results from my simulation studies showed that cluster wild bootstrapping maintained adequate Type I error rates and provided more power than the small sample correction methods that have been proposed in the meta-analytic literature thus far. I have also created an R package that implements cluster wild bootstrapping for meta-analysis.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Scientific researchers tend to produce literature on the same topic either to replicate or extend prior studies or due to a lack of awareness of prior evidence (Hedges & Cooper, 2009). Results across studies tend to vary, even when researchers try to replicate studies, due to differences in sample characteristics, research designs, analytic strategies or sampling error (Hedges & Cooper, 2009). Meta-analysis is a set of statistical techniques for synthesizing results from multiple primary studies on a common topic. Meta-analysis can be used to synthesize effect estimates from randomized or quasi-experimental studies, and correlations between variables from descriptive studies (Swanson et al., 2003). The three major goals of meta-analysis include: (1) summarizing effect size estimates across studies, (2) characterizing, and (3) explaining the variability in the effect sizes (Hedges et al., 2010).

Tanner-Smith and Lipsey (2015) is an example of a published meta-analysis. The study evaluated the effectiveness of brief alcohol interventions for adolescents and young adults. The authors included 185 study samples in their analysis and reported the overall average effect size estimate, which identified whether brief alcohol interventions reduced consumption and alcohol related problems. The results showed that the interventions led to statistically significant reductions in the outcomes. The authors also tested whether the intervention effects persisted over time and whether the effects varied by demographic characteristics of the participants, and intervention length and format. The results showed that the effects persisted up to one year after the interventions and did not vary across demographic characteristics of the participants, nor by intervention length and format.

Because meta-analysis involves synthesizing evidence from multiple primary studies, the results from a meta-analysis can have meaningful implications on policy evaluations in terms of funding interventions or policies, and targeting interventions towards certain demographics. The magnitude and direction of the pooled effect size can inform whether an intervention resulted in desired change in the outcome. Meta-regression can be used to test whether the effect sizes vary by certain characteristics of the studies or the samples. For example, Tanner-Smith and Lipsey (2015) showed that the brief alcohol interventions were effective, providing evidence for funding such

interventions. The effects of the interventions did not differ in terms of demographics indicating that the interventions need not be modified to target certain demographics.

Typical methods to conduct meta-analysis—pooling effect sizes or analyzing moderating effects with meta-regression—work under the assumption that the effect size estimates are independent. However, primary studies often report multiple estimates of effect sizes. For example, Tanner-Smith and Lipsey (2015) included studies that reported multiple correlated measures of the outcome variables, and thus had multiple dependent effect size estimates per study. Dependence can occur through two broad structures: correlated effects and hierarchical effects. Correlated effects typically occur due to primary studies collecting multiple correlated measures of an outcome, repeated measures of the outcome data, or comparing multiple treatment groups to the same control group (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015). Hierarchical effects can occur when primary meta-analytic studies include multiple experiments conducted by the same laboratory or in the same region creating dependence in the effect size parameters (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015).

Researchers may be inclined to ignore dependence and use methods that assume that each effect size estimate is independent. However, doing so can result in inaccurate standard errors and therefore, hypothesis tests with incorrect Type I error rates, and confidence intervals with incorrect coverage levels (Becker, 2000). Ad-hoc methods include averaging effect sizes by study or selecting an effect size randomly per study (Hedges et al., 2010). These methods result in loss of information and are not suitable for studying within-study variation in effect sizes (Hedges et al., 2010). A method called shifting the unit-of-analysis involves running meta-analytic models for different subsets of the data (Cooper, 1998). However, this strategy is not useful if a researcher wants to summarize effects across the subsets or study differential effects (Becker, 2000).

The ideal solution for handling dependence would be to use a multivariate model (Becker, 2000; Hedges et al., 2010; Raudenbush et al., 1988). This approach explicitly models dependence among the effect sizes (Becker, 2000; Hedges et al., 2010; Tipton, 2015). However, multivariate meta-analysis requires knowledge of correlations or covariances between pairs of effect size estimates within each study, which are often difficult to obtain from primary studies (Olkin & Gleser, 2009).

To handle dependence without knowing the covariance structure between effect size estimates, Hedges et al. (2010) proposed the use of robust variance estimation (RVE). RVE involves estimating the variances for the meta-regression model's coefficients using sandwich estimators that are valid even when the covariance structure is unknown or mis-specified (Hedges et al., 2010; Tipton, 2015). RVE is increasingly being used in applied meta-analyses (Tipton, 2015). However, the performance characteristics of RVE are asymptotic in that a large number of clusters or studies is required to provide accurate standard error estimates (Tipton, 2015). If the number of studies in a meta-analysis is small, RVE, as originally proposed by Hedges et al. (2010), can result in downwardly biased standard errors and inflation of Type I error rates (Hedges et al., 2010; Tipton, 2015).

Tipton (2015) evaluated several small sample corrections for RVE for tests of single coefficients: one based on a degrees of freedom correction (CR1, Cluster Robust Type 1) provided by Hedges et al. (2010); one based on a bias reduced linearization method (CR2) proposed by McCaffrey et al. (2001); and one based on the jack-knife technique (CR3) that is a generalization of the heteroskedasticity consistent HC3 type estimator proposed by MacKinnon and White (1985). Simulation results from Tipton (2015) showed that the methods, when used without further corrections, resulted in Type I error rate inflation. Tipton (2015) suggested using the Sattherthwaite degrees of freedom along with the adjustment methods. The simulation results showed that the CR2 adjustment method with the Satterthwaite degrees of freedom resulted in close to nominal Type I error rates. Moreover, Tipton (2015) showed that small sample size itself was not the only important factor that could influence the performance of RVE. The distribution of the covariates—for example, imbalanced categories or outliers in covariates—can also influence the performance of RVE. The CR2 adjustment method and the Satterthwaite degrees of freedom can account for imbalance and outliers.

Tipton and Pustejovsky (2015) extended Tipton (2015) and introduced small sample corrections for multiple-contrast hypotheses. In meta-analysis, multiple-contrast hypotheses can be important parts of the research aims as analysts may want to learn whether effects are same across different research designs, different populations, or different outcome measures. Tipton and Pustejovsky (2015) evaluated several methods based on eigen-decomposition and on the Hotelling's T^2 distribution. The authors also evaluated the Naive F-test that corresponds to the degrees of freedom

correction method proposed by Hedges et al. (2010) as a baseline comparison method. The results showed that the Naive F-test did not maintain Type I error rates adequately. Based on the results of their simulation study, the authors recommended a method (HTZ, Hotelling's T^2 Zhang) which approximates the test statistic using the Hotelling's T^2 distribution with degrees of freedom proposed by Zhang (2012, 2013). The results from Tipton and Pustejovsky (2015) showed that the HTZ test had Type I error rates closest to the nominal rate of 0.05. However, the estimated Type I error rates of the HTZ test were below the nominal rate, indicating that the test may possibly be conservative. Tipton and Pustejovsky (2015) did not directly evaluate power. Alternative methods that maintain adequate Type I error rates while providing better power compared to the HTZ test need to be examined.

In this dissertation, I examined an alternative method, cluster wild bootstrapping, which has been examined in the econometrics literature to handle dependence when the number of clusters is small (Cameron et al., 2008). General bootstrapping is a computational technique to estimate unknown terms like standard errors, confidence intervals, and p-values (Boos et al., 2003; Cameron et al., 2008). Bootstrapping involves re-sampling from the original data a number of times to create an empirical distribution which is used in place of the distribution of an estimate or test statistic (Boos et al., 2003). Bootstrapping can provide accurate estimates of uncertainty when other methods fail (Boos et al., 2003; MacKinnon, 2009).

Cluster wild bootstrapping involves sampling transformed residuals (Cameron et al., 2008). The residuals are calculated based on the model corresponding to the null hypothesis fit on the original dataset (Cameron et al., 2008; MacKinnon, 2009). The residuals are then multiplied by randomly assigned weights that are constant within each cluster (Cameron et al., 2008). These transformed residuals are used to generate a new outcome variable, which is used to calculate a test statistic for each bootstrap replication (Cameron et al., 2008). The steps of sampling transformed residuals and estimating test statistic based on the new outcome are repeated multiple times. The p-value is then calculated as the proportion of times the bootstrap test statistic is more extreme than the test statistic from the full model fit on the original dataset (Cameron et al., 2008). Cluster wild boostrapping has been shown to adequately control Type I error rates for clustered data with small number of clusters in regression analyses (Cameron et al., 2008; MacKinnon & Webb, 2018). However, it has not been

evaluated methodologically under a meta-analytic framework.

I conducted two simulation studies to examine whether cluster wild bootstrapping improves upon the performance of the HTZ test and the Naive F-test in terms of Type I error rates and power, for tests of single meta-regression coefficients and of multiple-contrast hypotheses. I included the Naive F-test as a baseline comparison method. As a part of my dissertation, I also built an R package, called `wildmeta`, that implements the cluster wild bootstrapping algorithm specifically for meta-analysis (Joshi et al., 2020). The results of my studies can provide guidance for applied meta-analytic researchers on which method to use to handle dependent effect sizes when the number of studies is small.

Chapter 2

Literature Review

2.1 Meta-Analysis

Meta-analysis involves summarizing across the effect size estimates reported in or calculated from information reported in primary studies. Effect sizes are quantitative measures of relationships among variables (Hedges, 2008). They are more comparable measures of the relationship between variables than p-values from statistical tests as the p-values are dependent on the test statistics, which are dependent on sample sizes (Hedges, 2008). A small p-value does not necessarily indicate a large effect (Hedges, 2008). On the other hand, effect sizes depend on population parameters rather than on sample sizes and thus, are comparable across studies (Hedges, 2008). Common measures of effect size include standardized mean differences (SMDs), correlations, difference in proportions, and odds ratios (Hedges, 2008; Hedges & Cooper, 2009). The SMD, a common measure of effect size for summarizing intervention research, captures the magnitude and direction of the standardized difference in the outcome variable between a treatment and a comparison group (Borenstein & Hedges, 2019; Tipton, 2015). For example, Tanner-Smith and Lipsey (2015) analyzed the SMDs between people who received brief alcohol interventions compared to those who did not on measures related to alcohol consumption and alcohol related problems.

The major goals of meta-analysis are: (1) pooling effect size estimates across studies, (2) characterizing, and (3) explaining variability in the effect sizes. The three goals of meta-analysis are discussed below.

2.1.1 Pooling Effect Size Estimates

The first major goal of meta-analysis is to summarize effect size estimates across multiple studies to estimate the average effect of an intervention or the average measure of the relationship between two variables (Konstantopoulos & Hedges, 2019). Let m denote the number of studies in a meta-analysis, with each study contributing one effect size estimate, T_i for $i = 1, ..., m$. Below, let w_i indicate some general weight.

Effect size estimates can be pooled as follows (Konstantopoulos & Hedges, 2019):

$$\hat{\mu} = \frac{\sum_{i=1}^{m} w_i T_i}{\sum_{i=1}^{m} w_i} \tag{2.1}$$

One way to pool effect size estimates is by weighing them by the inverse of their variance estimates; these weights denote the precision of the estimated effect sizes (Viechtbauer, 2007). The calculation of the inverse variance weights depends on certain assumptions which are discussed below.

Common Effect and Fixed Effects Models

Rice et al. (2018) outlined different assumptions that can be made when pooling effect size estimates. One assumption is the identical parameters assumption underlying the common effect model. This assumption states that one true effect size underlies all of the studies (Rice et al., 2018). An alternative assumption is the independent parameters assumption underlying the fixed effects model. This assumption treats the set of studies in a meta-analysis as all of the studies in the population of interest (Rice et al., 2018). The inferences derived from a fixed effects analysis would be valid for the specific set of studies included in the analysis (Rice et al., 2018). When using the common effect and fixed effects models, the inverse variance weights can be calculated as $w_i = 1/\hat{\sigma_i}^2$, where $\hat{\sigma}_i^2$ denotes the sampling error in the estimation of the effect sizes (Konstantopoulos & Hedges, 2019).

Random Effects Model

Unlike the fixed effects model, the random effects model treats the set of studies in a meta-analysis as a sample of all possible studies in the population of interest (Higgins et al., 2009; Konstantopoulos & Hedges, 2019). The variance of the effect sizes between studies is denoted by τ^2 (Higgins et al., 2009; Konstantopoulos & Hedges, 2019). The pooled effect is still a weighted average of the effect size estimates. However, the inverse variance weights account for between study variance, τ^2, as well as the sampling error. When using the random effects model, the inverse variance weights can be calculated as the inverse of the sum of the two variance components, $w_i = 1/(\hat{\tau}^2 + \hat{\sigma_i}^2)$, where $\hat{\tau}^2$ is some estimate of the between-study variance (Konstantopoulos & Hedges, 2019).

2.1.2 Characterizing Variability

Pooling effect size estimates provides an average estimate of the effect of an intervention or the relationship between some variables. However, typically, researchers are also interested in the existence of variation in the effect sizes (Konstantopoulos & Hedges, 2019). To test whether a common population effect size underlies effect size estimates in a meta-analysis, the Cochran's Q statistic can be used (Cochran, 1954). It characterizes the sum of squared differences between individual study effects and the pooled effect across studies (Konstantopoulos & Hedges, 2019). The individual effects and the pooled effect are weighted by the inverse variance weights. The Q statistic is calculated as (Konstantopoulos & Hedges, 2019):

$$Q = \sum_{i=1}^{m} w_i T_i^2 - \frac{\left(\sum_{i=1}^{m} w_i T_i\right)^2}{\sum_{i=1}^{m} w_i} \tag{2.2}$$

The statistic is then compared to a χ^2 distribution with $m - 1$ degrees of freedom (Konstantopoulos & Hedges, 2019). A significant test result would indicate that it is unlikely that a common population effect size underlies all the effect sizes or, in other words, that the effect sizes likely vary in the population.

Another statistic that characterizes variability in the effect sizes is I^2 (Higgins & Thompson, 2002). It is a descriptive statistic that denotes the percentage or proportion of variance in the observed effect size estimates that is due to variation in the true effect sizes (Borenstein et al., 2017). Higgins and Thompson (2002) provided the following formula for calculating I^2:

$$I^2 = \frac{Q - (m - 1)}{Q} \times 100\% \tag{2.3}$$

Borenstein et al. (2017) contended that I^2 is not an absolute measure of heterogeneity as it cannot be used to derive the range of effect sizes in different populations.

Additionally, $\hat{\tau}^2$ is a descriptive statistic that denotes the estimated variation in the true effects or, as Viechtbauer (2007) described it, the estimated variance of the random variable producing the true effect sizes. There are various estimators suggested in the meta-analytic literature for τ^2 (Viechtbauer, 2010). Estimators include the DerSimonian-Laird estimator (DerSimonian & Laird, 1986), the Hedges estimator (Hedges & Olkin, 2014; Raudenbush, 2009), and the maximum-likelihood and

restricted maximum-likelihood estimators (Raudenbush, 2009; Viechtbauer, 2005). For a detailed comparison of the estimators, please see Veroniki et al. (2016).

2.1.3 Explaining Variability

In addition to pooling effect size estimates and characterizing variability in effect sizes, meta-analysts often want to examine what factors explain or are associated with the variability. For example, the major questions in Tanner-Smith and Lipsey (2015) included whether the effects of the brief alcohol interventions differed for different demographic groups and whether the effects differed based on the length and format of the interventions. Identifying answers to such questions can clarify whether an intervention is effective for groups of interest, and whether the intervention should be developed further to be more effective for target populations under relevant conditions.

Meta-Regression Model

To explain variability in the effect sizes, a meta-regression model is generally used. Let T_i denote effect size estimate i, p denote the number of regression parameters, $x_{i1}, ..., x_{i,p-1}$ denote a set of moderator values associated with effect size estimate i, $\beta_0, ..., \beta_{p-1}$ denote a vector of regression coefficients, and ϵ_i denote the error term (Konstantopoulos & Hedges, 2019). Moderators can include variables like average age or percentage female of the study sample. The meta-regression model can be written as follows (Hedges et al., 2010; Tanner-Smith et al., 2016):

$$T_i = \beta_0 + x_{i1}\beta_1 + , ..., + x_{i,p-1}\beta_{p-1} + \epsilon_i \tag{2.4}$$

A random effects weighted least squares meta-regression model can be used to estimate the parameters in Equation 2.4 (Tanner-Smith et al., 2016). An intercept-only model can be used to estimate the overall average effect size. A statistically significant test of a regression coefficient for a moderator from Equation 2.4 would indicate that the effects of a treatment or intervention depend on the level of the moderator—that the moderator explains statistically significant variation in the effect sizes. For example, Tanner-Smith and Lipsey (2015) found that intervention length did not significantly explain variability in the effect sizes.

2.2 Dependence

The models described so far maintain the assumption that there is only one effect size estimate per study and that the studies are independent (Hedges et al., 2010). However, in applied meta-analysis, each study can yield more than one effect size estimate from the same sample of subjects or multiple studies can be clustered in some way resulting in dependence (Hedges et al., 2010; Tipton & Pustejovsky, 2015). For example, Tanner-Smith and Lipsey (2015) included multiple effect size estimates per study because the primary studies evaluated more than one measure of the two outcomes of interest, alcohol consumption and alcohol related problem, on the same sample.

Hedges (2019b) noted that an effect size estimate i from study j, T_{ij}, can be partitioned into the true effect size, θ_{ij}, and the sampling error, ϵ_{ij}:

$$T_{ij} = \theta_{ij} + \epsilon_{ij} \tag{2.5}$$

Meta-analytic data containing dependent effect size estimates typically follow two broad structures: (1) correlated effects, and (2) hierarchical effects (Hedges et al., 2010). Dependence in the correlated effects structure occurs in effect size estimates through the sampling error terms, ϵ_{ij} (Hedges et al., 2010). On the other hand, dependence in the hierarchical effects structure occurs in the true effect sizes, θ_{ij} (Hedges et al., 2010). The two structures are discussed below.

2.2.1 Correlated Effects

Correlated effects occur when the effect size estimates are dependent. This dependence structure occurs when the same primary study collects: (1) multiple correlated measures of the outcome, (2) repeated measures of the outcome, (3) outcome measures when multiple treatment groups are compared to the same control group or multiple control groups are compared to the same treatment group, or (4) multiple correlations from the same sample (Becker, 2000; Hedges et al., 2010; Olkin & Gleser, 2009). Meta-analytic data with correlated effects type dependency would contain multiple effect size estimates per study, with each effect size estimate associated with a particular outcome measure or comparison. Dependence occurs when multiple mea-

sures are collected from the same sample, so that a study with a higher than average effect size estimate on one outcome measure, for example, will tend to have the same on other outcome measures (Hedges et al., 2010).

An example of correlated effects structure is the data analyzed by Tanner-Smith and Lipsey (2015). Studies included in Tanner-Smith and Lipsey (2015) reported multiple correlated measures of the outcome variables. For example, alcohol consumption was measured by frequency of consumption, quantity consumed, and blood alcohol concentration. Alcohol-related problems were measured by risky sexual behavior, relationship problems, and driving under influence (DUI) or driving while intoxicated (DWI) convictions.

Another example of correlated effects structure is the data analyzed by Sala et al. (2018). The authors conducted a meta-analysis to study whether video game training enhances cognitive ability. The meta-analysis included results from randomized controlled trials that compared one treatment group (active video game players) to multiple comparison groups (non video game players, and non-active video game players) yielding multiple dependent effect size estimates per study. These examples illustrate that a correlated effects data structure contains multiple effect size estimates per study that are dependent because they are measured on the same sample.

2.2.2 Hierarchical Effects

Hierarchical effects structures can occur when independent samples are nested within a larger group (Tanner-Smith et al., 2016). For example, studies can be clustered by the same lab, investigator or region, where the methods, protocols, and personnel used for conducting the studies can be similar (Tanner-Smith et al., 2016). Meta-analytic data with hierarchical effects type dependency would contain one or more effect sizes per study but with studies nested within labs or researchers or regions. Different labs can have different protocols—for example, one lab may have more consistent materials and carefully tailored interventions compared to other labs—that may result in different effects. Knowing something about the lab can tell meta-analysts about the true effect size, even though different samples may have been used by different studies conducted by the lab.

An example of hierarchical effects structure is the data analyzed by Thompson et al. (2017). The authors conducted a meta-analysis studying whether alcohol de-

creases experimentally induced pain. The authors noted that several studies in the meta-analysis were conducted by the same laboratory which may have used similar methodology to conduct the different experiments. This example illustrates that a hierarchical effects data structure contains effect sizes across studies that are dependent due to the nesting of the studies within a larger group.

2.3 Methods to Handle Dependence

One strategy to handle dependence is to ignore it, assume the effect size estimates are independent, and proceed with running the meta-regression model (Hedges et al., 2010). Hedges et al. (2010) noted that this procedure might perform well if the number of studies with dependent effect size estimates is small. However, generally that may not be the case and this approach will lead to incorrect standard errors and incorrect inferences from hypothesis tests (Hedges et al., 2010).

2.3.1 Ad-Hoc Methods

Ad-hoc methods for handling dependence include randomly selecting one effect size estimate per study or averaging effect size estimates for each study (Becker, 2000; Tanner-Smith et al., 2016). These methods can yield independent effect size estimates but can be problematic when there is within-study variation in the effect sizes or in the moderator variables. Deleting or averaging effect size estimates can cause loss of potentially necessary information (Tanner-Smith et al., 2016).

Another ad-hoc procedure is the shifting the unit-of-analysis approach, which involves creating subsets of effect size estimates (Cooper, 1998). For example, a researcher can separate the effect size estimates for each outcome measure and then conduct univariate meta-analysis per outcome measure. The shifting the unit-of-analysis approach prevents loss of information (Scammacca et al., 2014). However, it requires multiple meta-analytic models for different subsets of the data (Scammacca et al., 2014). This approach can result in low power due to multiplicity, or reduction in the number of studies. Moreover, this strategy is not useful if a researcher wants to make comparisons across the subsets or study differential effects (Hedges et al., 2010). For example, the researcher may want to calculate the average effect size estimate across all the outcome measures and examine whether the magnitude of the effect is

bigger for certain outcome measures (Hedges et al., 2010).

2.3.2 Multivariate Methods

The ideal solution to handle dependence is to run multivariate meta-analysis which explicitly models the dependencies between effect sizes (Becker, 2000; Hedges et al., 2010; Raudenbush et al., 1988). Multivariate meta-analysis requires the knowledge of the covariance structure between effect size estimates (Hedges et al., 2010). However, the covariance structure is difficult to derive from the information provided in primary studies (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015). Therefore, although the multivariate method is the most ideal solution, it is oftentimes infeasible to use in practice (Hedges et al., 2010).

2.3.3 Robust Variance Estimation

Hedges et al. (2010) proposed another procedure to handle dependence, robust variance estimation (RVE), which does not require the exact covariance structure between the effect size estimates. Instead, the variances of the meta-regression coefficients are estimated with sandwich estimators using observed residuals (Hedges et al., 2010; Tipton, 2015). RVE can be used whether the underlying meta-analytic data structure is correlated effects or hierarchical effects (Hedges et al., 2010).

In the context where there are multiple effect size estimates per study, the meta-regression model from Equation 2.4 can be written as follows. Let $\mathbf{T}_j$ denote a $k_j \times 1$ vector of effect size estimates from study j, $\mathbf{X}_j$ denote a $k_j \times p$ matrix of covariates, $\boldsymbol{\beta}$ denote a $p \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}_j$ denote a $k_j \times 1$ vector of errors with mean of 0 and covariance matrix $\boldsymbol{\Psi}_j$ for studies $j = 1, .., m$ (Hedges et al., 2010; Tipton & Pustejovsky, 2015). The meta-regression model can be written as follows (Hedges et al., 2010; Tipton & Pustejovsky, 2015):

$$\mathbf{T}_j = \mathbf{X}_j\boldsymbol{\beta} + \boldsymbol{\epsilon}_j \tag{2.6}$$

$\mathbf{X}_j$ can include categorical and quantitative variables (Tipton & Pustejovsky, 2015). The covariates can be study-level variables, ones that vary between studies like average age of the sample, or effect size-level variables, ones that vary within studies like the outcome measure used (Tipton & Pustejovsky, 2015).

Let $\mathbf{W}$ denote a block diagonal matrix of weights with components $W_1, ..., W_m$ and let $\mathbf{M} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ (Tipton & Pustejovsky, 2015). The weighted least squares estimate of $\boldsymbol{\beta}$ in Equation 2.6 can be calculated as (Tipton & Pustejovsky, 2015):

$$\mathbf{b} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j'\mathbf{W}_j\mathbf{T}_j\right) \tag{2.7}$$

The exact variance of $\mathbf{b}$ is:

$$\text{Var}(\mathbf{b}) = \mathbf{M}\left[\sum_{j=1}^{m} \mathbf{X}_j{}'\mathbf{W}_j\boldsymbol{\Psi}_j\mathbf{W}_j\mathbf{X}_j\right]\mathbf{M} \tag{2.8}$$

If $\boldsymbol{\Psi}_j$ is fully known, $\boldsymbol{\Psi}_j^{-1}$ can be used as $\mathbf{W}_j$ (Tipton & Pustejovsky, 2015). In such a case, Var($\mathbf{b}$) reduces to $\mathbf{M}$ (Tipton & Pustejovsky, 2015). However, the covariance structure between effect size estimates, $\boldsymbol{\Psi}_j$, can rarely be calculated from information reported in primary studies (Hedges et al., 2010; Tipton & Pustejovsky, 2015). Therefore, the direct calculation of Equation 2.8 is not feasible in practice.

The RVE estimator of the variance of $\mathbf{b}$ does not require the knowledge of $\boldsymbol{\Psi}_j$. The RVE estimator is as follows:

$$\mathbf{V}^{\mathbf{R}} = \mathbf{M}\left[\sum_{j=1}^{m} \mathbf{X}_j'\mathbf{W}_j\mathbf{A}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{A}_j\mathbf{W}_j\mathbf{X}_j\right]\mathbf{M} \tag{2.9}$$

Here, $\mathbf{e}_j$ is the vector of residuals for study j, where $\mathbf{e}_j = \mathbf{T}_j - \mathbf{X}_j\mathbf{b}$ (Tipton & Pustejovsky, 2015). Furthermore, $\mathbf{A}_j$ denotes a $k_j \times k_j$ adjustment matrix (Tipton, 2015). In the formulation of RVE initially proposed by Hedges et al. (2010), $\mathbf{A}_j$, in Equation 2.9, is an identity matrix (Hedges et al., 2010; Tipton & Pustejovsky, 2015). The original RVE is called CR0 (Cameron & Miller, 2015; Pustejovsky & Tipton, 2018). Hedges et al. (2010), Tipton (2015), and Tipton and Pustejovsky (2015) have proposed several other adjustment matrices for the calculation of RVE; they are explained further below.

Weights in RVE

Hedges et al. (2010) suggested two approximate formulations of the inverse variance weights based on the correlated and hierarchical effects working models. Exact

inverse variance weights cannot be calculated as the covariance structure between effect sizes is unknown (Hedges et al., 2010; Tipton & Pustejovsky, 2015). The approximate inverse variance weights are based on working models with simplified assumptions that may not reflect the structure of actual meta-analytic data. However, when calculating RVE, using the approximate weights can improve efficiency of the variance estimates compared to deriving the estimates using any other type of weights (Hedges et al., 2010; Tipton & Pustejovsky, 2015). Because the use of the weights only impact efficiency, working models that include simplified assumptions can be used without impacting statistical inferences (Hedges et al., 2010; Tanner-Smith et al., 2016; Tipton & Pustejovsky, 2015). Actual meta-analytic data can include both correlated and hierarchical effects structures simultaneously (Tanner-Smith et al., 2016). In such a scenario, Tanner-Smith et al. (2016) recommended using the working model based on the most prevalent structure. The two working models and approximate weights calculations are described below.

Correlated Effects Working Model

Correlated effects structure occurs when the error terms are dependent (Hedges et al., 2010). Below, let T_{ij} denote the effect size estimate i in study j, θ_{ij} denote the true effect size i in study j, and e_{ij} denote the sampling error term that is normally distributed with mean of 0 and variance of σ^2_{ij}. Under the correlated effects working model:

$$T_{ij} = \theta_{ij} + e_{ij} \quad \text{with} \quad e_{ij} \sim N(0, \sigma^2_{ij}) \tag{2.10}$$

The correlation between two error terms, h and i, in study j is assumed to be $\text{corr}(e_{hj}, e_{ij}) = \rho$ (Hedges et al., 2010). This assumption of constant correlation between pairs of error terms is a strong simplifying assumption underlying the correlated effects working model. In actual meta-analytic data, the correlations between two sampling error terms may differ for each pair.

Let γ_j denote the average effect size for study j, μ denote the overall average effect size across all the studies and v_j denote the between-study sampling error—the study level random effect—that is normally distributed with mean of 0 and variance of τ^2. In the correlated effects working model,

$$\theta_{ij} = \gamma_j \tag{2.11}$$

and

$$\gamma_j = \mu + v_j \quad \text{with} \quad v_j \sim N(0, \tau^2) \tag{2.12}$$

The assumption underlying the model here is that the true effect sizes are identical within studies, and only vary between studies. The assumption again is a strong assumption that may not reflect the actual structure of meta-analytic data. In correlated effects data structure, the true effect sizes may vary both within and between studies.

Equations 2.10, 2.11 and 2.12 imply that:

$$T_{ij} = \mu + v_j + e_{ij} \tag{2.13}$$

Equation 2.13 implies that the effect sizes have the following marginal variance-covariance matrix (Hedges et al., 2010; Tipton & Pustejovsky, 2015):

$$\mathbf{\Psi}_{cj} = \tau^2 \mathbf{J}_j + \rho\sigma_j^2 \left(\mathbf{J}_j - \mathbf{I}_j\right) + \sigma_j^2 \mathbf{I}_j \tag{2.14}$$

Here, $\mathbf{\Psi}_{cj}$ denotes the covariance matrix according to the correlated effects working model for study j, τ^2 denotes the between-study variance in the average effect sizes, $\mathbf{J}_j$ denotes a $k_j \times k_j$ matrix of 1's, $\mathbf{I}_j$ denotes a $k_j \times k_j$ identity matrix, ρ denotes the correlation between pairs of sampling error terms assumed to be constant within and across studies, and σ_j^2 denotes the within-study sampling variance for study j assumed to be constant within studies (Hedges et al., 2010; Tipton & Pustejovsky, 2015). Hedges et al. (2010) provided a method of moments estimator for τ^2. For the correlated effects model, Hedges et al. (2010) proposed an approximation of the inverse variance weights calculated as (Tipton & Pustejovsky, 2015):

$$\mathbf{W}_j = \frac{1}{k_j \left(\sigma_j^2 + \tau^2\right)} \mathbf{I}_j \tag{2.15}$$

Weight matrices calculated based on Equation 2.15 are then used to estimate $\boldsymbol{\beta}$ based on Equation 2.7, with the variance of $\mathbf{b}$ estimated based on Equation 2.9.

Hierarchical Effects Working Model

Hierarchical effects structure occurs when the true effect sizes are dependent (Hedges et al., 2010). Under the hierarchical effects working model (Konstantopoulos, 2011):

$$T_{ij} = \theta_{ij} + e_{ij} \quad \text{with} \quad e_{ij} \sim N(0, \sigma_{ij}^2) \tag{2.16}$$

The correlation between two sampling errors from the same study is assumed to be $\text{corr}(e_{hj}, e_{ij}) = 0$ (Konstantopoulos, 2011). This assumption may not reflect the actual data structure.

Under the working model:

$$\theta_{ij} = \gamma_j + u_{ij} \quad \text{with} \quad u_{ij} \sim N(0, \omega^2) \tag{2.17}$$

and

$$\gamma_j = \mu + v_j \quad \text{with} \quad v_j \sim N(0, \tau^2) \tag{2.18}$$

Note that unlike the correlated effects model, the hierarchical effects model has an error term u_{ij} associated with the average effect size parameter for study j. The error term is normally distributed with a mean of 0 and within-study variance of the effect sizes of ω^2. In the hierarchical effects model, level 1 is the sampling error, level 2 is the within-study error and level 3 is the between-study error (Tanner-Smith et al., 2016).

Equations 2.16, 2.17, and 2.18 imply that (Konstantopoulos, 2011):

$$T_{ij} = \mu + v_j + u_{ij} + e_{ij} \tag{2.19}$$

Equation 2.19 implies that the effect sizes have the following marginal variance-covariance matrix (Hedges et al., 2010; Tipton & Pustejovsky, 2015):

$$\mathbf{\Psi}_{hj} = \tau^2 \mathbf{J}_j + \omega^2 \mathbf{I}_j + \mathbf{V}_j \tag{2.20}$$

Here, $\mathbf{\Psi}_{hj}$ denotes the covariance matrix according to the hierarchical effects working model for study j, ω^2 denotes the within-study variance in the true effect sizes, and $\mathbf{V}_j$ denotes the $k_j \times k_j$ diagonal matrix of sampling error variances for study j (Tanner-Smith et al., 2016; Tipton & Pustejovsky, 2015). Hedges et al. (2010)

provided method of moments estimators for τ^2 and ω^2. For the hierarchical effects model, Hedges et al. (2010) proposed an approximation of the inverse variance weights calculated as (Tipton & Pustejovsky, 2015):

$$\mathbf{W}_j = diag(w_{1j}, ..., w_{kj}) \tag{2.21}$$

where

$$w_{ij} = \frac{1}{\left(\sigma_{ij}^2 + \omega^2 + \tau^2\right)} \tag{2.22}$$

Weight matrices calculated based on Equation 2.21 are then used to estimate $\boldsymbol{\beta}$ based on Equation 2.7, with the variance of $\mathbf{b}$ estimated based on Equation 2.9.

2.3.4 Large Sample Hypothesis Testing

Consider the null hypothesis for the test of a single meta-regression coefficient, $\mathrm{H}_0 : \beta_s = 0$. Let b_s denote the s^{th} item in $\mathbf{b}$, and V_{ss}^R denote the s^{th} diagonal element of $\mathbf{V^R}$. The Wald test statistic is formulated as (Tipton & Pustejovsky, 2015):

$$t_s = \frac{b_s}{\sqrt{V_{ss}^R}} \tag{2.23}$$

When the number of studies is large, the t_s statistic follows a normal distribution (Hedges et al., 2010; Tanner-Smith et al., 2016). A z-test can be conducted to determine whether the meta-regression coefficient is different from zero. A statistically significant result would provide evidence that the null hypothesis is likely not true.

In addition to tests of single coefficients, meta-analysts are also often interested in tests of multiple-contrast hypotheses—e.g., comparison of nested models and the moderating effect of a categorical variable with multiple levels. An example of multiple-contrast hypothesis test includes the analysis conducted by Bediou et al. (2018) in a meta-analysis examining whether action video games impact perceptual, attentional and cognitive skills. The authors conducted a multiple-contrast hypothesis test to examine whether the effects of action video games were same across different cognitive domains. The different domains included perception, top-down attention, spatial cognition, multi-tasking, and verbal cognition.

Consider the null hypothesis $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{C}$ denotes a $q \times p$ contrast matrix

and $\mathbf{c}$ denotes a $q \times 1$ vector (Tipton & Pustejovsky, 2015). Here, q denotes the number of constraints. For example, meta-analysts might want to examine the equality of regression coefficients where $\mathrm{H}_0 : \beta_1 = \beta_2 = ... = \beta_{p-1}$ (Tipton & Pustejovsky, 2015). To evaluate the null hypothesis, a Q statistic can be calculated as follows (Tipton & Pustejovsky, 2015):

$$Q = (\mathbf{Cb} - \mathbf{c})' \left(\mathbf{C}\mathbf{V}^{\mathbf{R}}\mathbf{C}'\right)^{-1} (\mathbf{Cb} - \mathbf{c}) \tag{2.24}$$

The Q statistic follows a χ^2 distribution with q degrees of freedom when the number of studies is adequately large (Tipton & Pustejovsky, 2015). The Q statistic can be converted to an F statistic where $F = Q/q$ (Tipton & Pustejovsky, 2015). The F statistic follows an F distribution with q and infinity degrees of freedom when the number of studies is large (Tipton & Pustejovsky, 2015). A statistically significant result would indicate that the null hypothesis is rejected—that the contrasts specified may not hold.

2.3.5 Asymptotic Characteristic of CR0 Type RVE

In the CR0 type RVE, the true covariance matrix, $\mathbf{\Psi}_j$, is estimated by $\mathbf{A}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{A}_j$, where $\mathbf{A}_j$ is a $k_j \times k_j$ identity matrix. Although $\mathbf{A}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{A}_j$ is a poor estimate of $\mathbf{\Psi}_j$, because of the weak law of large numbers, as the number of studies increases, $\mathbf{V}^{\mathbf{R}}$ converges to Var($\mathbf{b}$) (Hedges et al., 2010; Tipton, 2015). Therefore, the performance characteristics of the CR0 RVE are asymptotic in that the method requires a large number of clusters or studies to provide accurate standard errors (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015). Simulation studies have shown that if the number of studies is small, the CR0 RVE can result in downwardly biased standard errors and inflation of Type I error rates for tests of single coefficients as well as for tests of multiple-contrast hypotheses (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015).

2.3.6 Small Sample Corrections for Tests of Single Coefficients

To try to improve upon the effects of the asymptotic characteristic of the CR0 type RVE, Hedges et al. (2010) and Tipton (2015) examined several corrections to RVE for

tests of single coefficients. Hedges et al. (2010) suggested the following adjustment matrix (CR1) to be used when calculating RVE:

$$\mathbf{A}_j = \sqrt{m/(m-p)}\mathbf{I}_j \tag{2.25}$$

Here, m denotes the number of studies, p denotes the number of coefficients estimated, and $\mathbf{I}_j$ denotes a $k_j \times k_j$ identity matrix (Hedges et al., 2010; Tipton & Pustejovsky, 2015). The t-statistic for the test of a single coefficient can then be compared to a t-distribution with $m-p$ degrees of freedom (Hedges et al., 2010; Tipton & Pustejovsky, 2015). However, Hedges et al. (2010) and Tipton (2015) have shown that even with the adjustment, Type I error rates can be inflated when the number of studies is small. Tipton (2015) and Tipton and Pustejovsky (2015) noted that over half of published meta-analyses in education and social sciences contain fewer than 40 primary studies; over half of published meta-analysis have small number of studies. Therefore, small sample corrections for RVE need to be examined.

Tipton (2015) proposed and evaluated further methods to improve small-sample performance of RVE for single coefficient meta-regression t-tests. The methods evaluated by Tipton (2015) included the degrees of freedom correction (CR1) proposed by Hedges et al. (2010), a bias reduced linearization method (CR2) proposed by McCaffrey et al. (2001), and a jack-knife estimator (CR3). The three methods are generalizations of the heteroskedasticity-consistent variance estimator types HC1, HC2, and HC3 respectively from MacKinnon and White (1985). Each of these methods uses a different adjustment matrix when calculating RVE in Equation 2.9.

Tipton (2015) proposed using the Satterthwaite correction for degrees of freedom along with using the adjustment matrices. The Satterthwaite degrees of freedom are approximated as follows (Tipton, 2015; Tipton & Pustejovsky, 2015):

$$\nu_s = \frac{2\mathrm{E}(V_{ss}^R)^2}{\mathrm{Var}(V_{ss}^R)} \tag{2.26}$$

The distribution of V_{ss}^R is approximated using a multiple of a χ^2 distribution. Assume that $aV_{ss}^R \sim \chi^2$ for some constant a. Based on the properties of the χ^2 distribution, $\mathrm{E}(aV_{ss}^R) = \nu_s$ and $\mathrm{Var}(aV_{ss}^R) = 2\nu_s$. Because a is a constant, $a = \nu_s/\mathrm{E}(V_{ss}^R)$ and $2\nu_s = \mathrm{Var}(aV_{ss}^R) = a^2\mathrm{Var}(V_{ss}^R) = \nu_s^2 \frac{\mathrm{Var}(V_{ss}^R)}{[\mathrm{E}(V_{ss}^R)]^2}$. In practice, $\mathrm{E}(V_{ss}^R)$ and $\mathrm{Var}(V_{ss}^R)$ are usually unknown but can be approximated based on the working model.

Table 2.1

Tipton (2015): Data Generating Conditions

Conditions	Study 1 Values	Study 2 Values
Number of studies (m)	10, 20, 40	20
Number of effect sizes per study (k_j)	1, 2, 5, 10, varied	10, varied
Between-study heterogeneity (I^2)	0, 0.33, 0.5	0, 0.33, 0.5
Correlation between outcomes (ρ)	0, 0.5, 0.8	0, 0.5, 0.8
Sample size per study (N_j)	40	40, varied

Tipton (2015) ran two simulation studies to examine the small sample correction methods. The author generated standardized mean difference estimates derived from primary studies that compared two groups on multiple correlated outcomes. Tipton (2015) generated eight covariates, three of which were study level. Some of the covariates were generated to be highly imbalanced or non-normal. The number of studies (m), the number of effect sizes per study (k_j), between-study heterogeneity (τ^2), the within-study correlation between outcomes (ρ), and the sample size per study (N_j) were varied as shown in Table 2.1. In the table, the τ^2 values are translated to I^2 values. The author included conditions where the number of effect sizes per study and the sample size per study varied by study.

Tipton (2015) compared the CR1, CR2, and CR3 correction methods, with and without the Satterthwaite degrees of freedom, in terms of Type I error rates, power, and degrees of freedom. In Study 1, Tipton (2015) ran models that included one covariate at a time. Tipton (2015) examined power in Study 1. To examine power, Tipton (2015) generated standardized mean differences with the regression coefficient for the moderators set to 0.10, 0.20, and 0.40 for the dichotomous variables and 0.05, 0.10, and 0.20 for the continuous variables. In Study 2, Tipton (2015) ran different models including different sets of covariates with each set including four of the covariates. Tipton (2015) used $m - p$ degrees of freedom in conditions where the Satterthwaite correction for degrees of freedom was not used. Tipton (2015) estimated the test statistics using fixed effects weights disregarding the correct working model to gain computational efficiency. The number of simulation replications was 10,000.

Simulation results from Tipton (2015) showed that all the methods when used without any further correction performed poorly, resulting in Type I error rate inflation. However, all methods performed well when combined with the Satterthwaite

correction when the degrees of freedom were greater than or equal to four (Tipton, 2015). When the degrees of freedom were less than four, Type I error inflation occurred (Tipton, 2015). The CR2 and CR3 methods combined with the Satterthwaite degrees of freedom showed better Type I error rate control than the CR1 method combined with the Satterthwaite degrees of freedom. Tipton (2015) showed that the CR3 method with the Satterthwaite degrees of freedom was more conservative than the CR2 method with the Satterthwaite degrees of freedom.

The CR2 adjustment corresponds directly to the HC2 adjustment (Tanner-Smith et al., 2016). Below, let $\mathbf{H}_j$ denote the hat matrix with $\mathbf{H}_j = \mathbf{X_j} \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}_j'\mathbf{W}_j$ (Tanner-Smith et al., 2016). The CR2 adjustment matrix for the correlated effects working model is as follows (Tipton, 2015):

$$\mathbf{A}_j^c = \left(\mathbf{I} - \mathbf{H}_j\right)^{-1/2} \tag{2.27}$$

The CR2 adjustment matrix for the hierarchical effects working model is as follows (Tipton, 2015):

$$\mathbf{A}_j^h = \mathbf{W}_j^{-1/2}[\mathbf{W}_j^{-1/2} \left(\mathbf{I} - \mathbf{H}_j\right) \mathbf{W}_j^{-3/2}]^{-1/2}\mathbf{W}_j^{-1/2} \tag{2.28}$$

Tipton and Pustejovsky (2015) noted that when the working model is correct, using the CR2 adjustment matrix when estimating $\mathbf{V^R}$ provides an exactly unbiased estimate of the variance of $\mathbf{b}$.

Tipton (2015) noted that small sample size was not the only factor that affected Type I error rates of RVE. Tipton (2015) showed through simulations that imbalance and leverage in moderators can also influence the performance of RVE. Using the CR2 adjustment matrices when calculating RVE accounts for leverage (Tanner-Smith et al., 2016; Tipton, 2015). Using the Satterthwaite degrees of freedom can further account for leverage as the Satterthwaite degrees of freedom formula incorporates the noisiness in RVE due to high leverage (Tipton, 2015). Tipton (2015) showed that the Satterthwaite degrees of freedom tended to be smaller for covariates with large imbalances and high leverage. Tipton (2015) suggested that it may be important to use the Satterthwaite degrees of freedom even if the total number of studies is large.

2.3.7 Small Sample Corrections for Tests of Multiple-Contrast Hypotheses

Tipton and Pustejovsky (2015) extended the methods developed by Tipton (2015) to F-tests of multiple-contrast hypotheses. For multiple-contrast hypotheses tests, Tipton and Pustejovsky (2015) first considered a degrees of freedom correction (CR1) similar to the correction for single coefficient tests suggested by Hedges et al. (2010). The test statistic can be calculated as $F = Q/q$, which is compared to an F distribution with q and $m - p$ degrees of freedom. The Q statistic is calculated using the CR1 adjustment matrices. Tipton and Pustejovsky (2015) called this test the Naive F-test.

In the development of small sample corrections for multiple-contrast hypotheses tests, Tipton and Pustejovsky (2015) used the CR2 adjustment matrices suggested by Tipton (2015). The authors used the approximate inverse variance weights suggested by Hedges et al. (2010). Tipton and Pustejovsky (2015) implemented several strategies to approximate the sampling distribution of the Q statistic. The authors reviewed literature on analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), and generalized estimating equations (GEEs). In each of these areas, small sample corrections for RVE had been examined.

Let $\mathbf{D}$ denote the random matrix $\left(\mathbf{C}\mathbf{V}^{\mathbf{R}}\mathbf{C}'\right)^{-1}$ from Equation 2.24. The first strategy uses a spectral decomposition of $\mathbf{D}$, estimating Q as a sum of independent univariate random variables (Tipton & Pustejovsky, 2015). The second strategy proposed by Tipton and Pustejovsky (2015) approximates the sampling distribution of $\mathbf{D}$ using the Wishart distribution leading to a Hotelling's T^2 distribution based test statistic. The Hotelling's T^2 distribution is a multivariate distribution that is proportional to the F distribution (Hotelling, 1931; Tipton & Pustejovsky, 2015). The Q statistic follows a Hotelling's T^2 distribution with degrees of freedom η (Tipton & Pustejovsky, 2015):

$$\frac{\eta - q + 1}{\eta q} Q \quad \sim F(q, \eta - q + 1) \tag{2.29}$$

To estimate η, Tipton and Pustejovsky (2015) examined three different approaches.

Tipton and Pustejovsky (2015) ran a simulation study to evaluate the different correction methods. The simulation study design of Tipton and Pustejovsky (2015)

Table 2.2

Tipton and Pustejovsky (2015): Data Generating Conditions

Conditions	Values
Number of studies (m)	10, 15, 20, 30, 40, 60, 80, 100
Between-study heterogeneity (I^2)	0.00, 0.33, 0.50, 0.75
Correlation between outcomes (ρ)	0.00, 0.50, 0.80

followed that of Tipton (2015). Tipton and Pustejovsky (2015) generated correlated standardized mean differences based on summary statistics. The authors generated effect size estimates derived from primary studies that compared two groups on multiple correlated outcomes. Tipton and Pustejovsky (2015) varied the number of independent studies (m), between-study heterogeneity (I^2), and the correlation between outcomes (ρ). Table 2.2 shows the different conditions examined by Tipton and Pustejovsky (2015). The authors included 1 to 10 effect sizes per study and set the sample size in each intervention group between 32 to 130, assuming equal sample size across the two groups. The authors generated a design matrix with five covariates, two study-level and three effect size-level. Some of the covariates exhibited large imbalances or non-normality.

Tipton and Pustejovsky (2015) used fixed effects weighting disregarding the correct working model when estimating RVE to gain computational efficiency. For each combination of conditions, the authors fit different regression specifications including 2, 3, 4, and 5 covariates and calculated $p - 1$ omnibus tests. For the model specification with 5 covariates, the authors ran subset tests for all combination of 2, 3 and 4 of the covariates. The simulation study compared the Naive F-test, two eigen-decomposition based correction methods, and three Hotelling's T^2 based correction methods for tests of multiple-contrast hypotheses. The authors calculated Type I error rates for α values of 0.01, 0.05, and 0.10. The number of simulation replications was 10,000.

The results from Tipton and Pustejovsky (2015) showed that the Naive F-test performed poorly even when the number of studies equaled 100. The eigen-decomposition based methods exhibited high Type I error rate inflation. The three Hotelling's T^2 based methods showed better Type I error rate control.

Of the three methods, the one (HTZ) that performed best was proposed by Zhang

(2012, 2013). This method was originally developed for heteroskedastic one-way ANOVA and MANOVA. The approach matches the total variance in $\mathbf{D}$ to the total variation in the Wishart distribution. The HTZ test corresponds to the CR2 adjustment method with the Satterthwaite degrees of freedom correction for tests of single coefficients (Tipton & Pustejovsky, 2015). The HTZ test can be used for tests of single coefficients as well as of multiple-contrast hypotheses. For the HTZ test, the degrees of freedom are estimated as (Zhang, 2012, 2013):

$$\eta_z = \frac{q(q+1)}{\sum_{s=1}^{q}\sum_{t=1}^{q}\mathrm{Var}\left(d_{st}\right)} \tag{2.30}$$

Here, d_{st} denotes the entry in row s and column t of $\mathbf{D}$. $\mathrm{Var}(d_{st})$ is estimated using the working model (Tipton & Pustejovsky, 2015). The results from Tipton and Pustejovsky (2015) showed that the HTZ test resulted in Type I error rates closest to the nominal rate of 0.05. The Type I error rates of this method ranged from 0.00 to 0.04 across simulation conditions with a median of 0.0254. The other two Hotelling's based methods had Type I error rates near zero. Although the simulation results from Tipton and Pustejovsky (2015) showed that the HTZ test controlled Type I error rates adequately, HTZ had below nominal Type I error rates across conditions, indicating that the method may possibly be conservative. The Type I error rates for the HTZ tests were lower for conditions with lower number of studies and higher number of constraints.

2.4 Cluster Wild Bootstrapping

An alternative method that may account for dependence when the number of studies is small is cluster wild bootstrapping. Cluster wild bootstrapping has been investigated to correct clustered heteroskedastic error terms of regular regression parameter estimates (Cameron et al., 2008). However, as noted by Tipton and Pustejovsky (2015), it has not been studied in a meta-analytic framework. Below is a discussion of the bootstrapping procedure, different ways to bootstrap and the assumptions underlying each of them, and the argument for the use of cluster wild bootstrapping to handle dependence in meta-analyses with small number of studies.

2.4.1 Bootstrapping

Bootstrapping is a technique to estimate unknown quantities like standard errors, confidence intervals, and p-values from statistical models (Boos et al., 2003). The focus of this dissertation will be on using bootstrapping to conduct hypothesis tests. The general idea behind bootstrapping is to emulate the unknown distribution of an estimate or test statistic by creating an empirical distribution based on re-sampling many times from the original dataset (Boos et al., 2003).

The bootstrap data generating process (DGP) refers to how the empirical distribution is created from the original dataset (MacKinnon, 2009). The choice of the DGP can be very important as different processes have different underlying assumptions (MacKinnon, 2009). The DGP can involve re-sampling the data itself, sampling residuals or sampling transformed residuals (Cameron et al., 2008; MacKinnon, 2009). The process of deriving the residuals can involve imposing the null hypothesis or not (MacKinnon, 2009). When bootstrapping is used for conducting hypothesis tests in particular, MacKinnon (2006) recommended imposing the null hypothesis when calculating the residuals as the process of hypothesis testing involves testing where an estimate lies on the distribution of a test statistic when the null hypothesis is true. Furthermore, according to MacKinnon (2006), imposing a null hypothesis makes the bootstrap test more reliable as the parameters, especially nuisance parameters which may be the basis of the distribution of the test statistic, are more precisely and efficiently estimated. Three common bootstrapping DGPs and the assumptions underlying each of them are discussed below.

Pair Bootstrapping

Pair bootstrapping involves re-sampling the pair of outcome and covariates $(\mathbf{y}, \mathbf{X})$ with replacement from the original dataset (Freedman, 1981, 1984). The estimate or test statistic of interest is then calculated on each bootstrap sample. When using clustered data, the clusters are re-sampled (Cameron et al., 2008).

A disadvantage of this procedure is that it involves re-sampling the covariates instead of holding them constant (MacKinnon, 2009). MacKinnon (2009) argued that when each bootstrap sample has a different $\mathbf{X}^*$, inferences made from bootstrap tests can be misleading especially when the sample size is small and the distribution of a test statistic is highly dependent on $\mathbf{X}$. When re-sampling clusters, sample sizes

can vary across replications if clusters sizes are imbalanced (Djogbenou et al., 2019). Further, when the sample size is small, there may be situations where certain covariates may not have any variance after re-sampling, especially when whole clusters of covariates are re-sampled. Therefore, the estimation of the regression coefficient and standard errors may be infeasible (Cameron et al., 2008). For example, if there is a study-level covariate—a binary covariate indicating whether the study is experimental or not—and studies are re-sampled, there may be a situation where the variable does not take on multiple values. Additionally, an assumption underlying the pair bootstrap procedure is that each observation is an independent draw from a multivariate distribution (MacKinnon, 2009). In the case with clusters, each cluster is an independent draw.

Residual Bootstrapping

Residual bootstrapping involves re-sampling residuals with replacement while holding $\mathbf{X}$ constant (MacKinnon, 2009). The residuals are then used to calculate new outcome values for each bootstrap replication (MacKinnon, 2009). When using clustered data, the vector of residuals for each cluster is re-sampled with replacement (Cameron et al., 2008). The null hypothesis can be imposed when calculating the residuals (MacKinnon, 2009).

An assumption underlying this method is that the errors are independently and identically distributed and hence, homoskedastic (Cameron et al., 2008). Another assumption is that $\mathrm{E}[y|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$ (MacKinnon, 2009). In other terms, the functional form of the full model needs to be correct. Additionally, when clusters are involved, all clusters are assumed to have the same sample size (Cameron et al., 2008). An advantage of residual bootstrapping is that $\mathbf{X}$ is held constant (Cameron et al., 2008). Unlike pair bootstrapping, the issue of covariates lacking variance after re-sampling does not occur with residual bootstrapping (Cameron et al., 2008).

Wild and Cluster Wild Bootstrapping

Wild boostrapping, first proposed by Liu et al. (1988) based on work by Wu (1986), involves sampling transformed residuals (Cameron et al., 2008; MacKinnon & Webb, 2018). The general process of conducting a wild bootstrap test is as follows:

1. Fitting a null model and a full model on the original data.

2. Obtaining residuals from the null model.

3. Generating an auxiliary random variable that has mean of 0 and variance of 1 and multiplying the residuals by the random variable (MacKinnon, 2015).

4. Obtaining new outcome scores by adding the transformed residuals to the predicted values from the null model fit on the original data.

5. Re-estimating the full model with the new calculated outcome scores and obtaining the test statistic (Cameron et al., 2008).

Steps 3-5 are repeated over R bootstrap replications. The p-value can be calculated as the proportion of times that the bootstrap test statistic is more extreme than the test statistic from the original data (Cameron et al., 2008; MacKinnon, 2009).

Cameron et al. (2008) extended the wild bootstrap for analyses involving clusters. Within each cluster, the auxiliary weights are set to be constant (Cameron et al., 2008; MacKinnon & Webb, 2017). MacKinnon and Webb (2017) argued that because the weights are constant within each cluster, bootstrap-based inferences preserve the within-cluster variances and covariances of the error terms to the extent that the residuals preserve them. Because the within-cluster variances and covariances of error terms are preserved, the cluster wild bootstrap method is ideal to handle dependence.

Like residual bootstrapping, wild and cluster wild boostrapping involves holding $\mathbf{X}$ constant. This feature is particularly useful in cases where there are few clusters as it prevents lack of variance in $\mathbf{X}$ due to re-sampling (Cameron et al., 2008). Unlike residual bootstrapping, cluster wild bootstrapping does not require the regression error vectors to be identically and independently distributed or the clusters to have the same sample size (Cameron et al., 2008). Therefore, compared to the other DGPs, cluster wild bootstrapping may be a better procedure for conducting hypothesis tests when the number of clusters is small. An assumption underlying wild and cluster wild bootstrapping is that the functional form of the full model is specified correctly.

Davidson and Flachaire (2008) and MacKinnon (2006) noted that the residual terms can be transformed in some way before multiplying them by the auxiliary random variable. For non-clustered data, MacKinnon (2006) suggested dividing the

residuals by $(1 - h_i)^{1/2}$, with h_i denoting the i^{th} diagonal element of the hat matrix. This transformation corresponds to the HC2 type adjustment (Davidson & Flachaire, 2008; MacKinnon, 2006, 2013). MacKinnon (2006) noted that multiplying the residuals by the HC2 correction ensures that the transformed residuals will have the correct variance if the actual errors have constant variance. If the residuals are not transformed with the HC2 correction, the errors can have variance that underestimates the true error variance. MacKinnon (2013) showed through simulations that multiplying the residuals by the HC2 or HC3 corrections when running wild bootstrap tests resulted in better Type I error rates compared to multiplying the residuals by the HC1 correction. In the context with clustered errors, one possible way to transform the residuals is to multiply the residuals by the CR2 adjustment matrices from the null model. The transformed residuals will have the correct error variance under the assumed working model (Bell & McCaffrey, 2002; Pustejovsky & Tipton, 2018). Pustejovsky and Tipton (2018) showed through simulations that using the CR2 adjustment with cluster robust variance estimation (CRVE) helps correct the under-estimation of the error variance even if the working model is incorrect.

MacKinnon (2006) recommended imposing the null hypothesis when generating bootstrap replicates. The null model contains predictors that are not being tested in single coefficient tests or multiple-contrast hypothesis tests. The full model, on the other hand, includes all the predictors of interest. Let $\mathbf{X}_{0j}$ denote a $k_j \times p_0$ matrix of covariates in the null model, $\tilde{\boldsymbol{\beta}}_0$ denote a $p_0 \times 1$ vector of coefficients from the null model fit to the original dataset, and $\tilde{\boldsymbol{e}}_j$ denote a $k_j \times 1$ vector of residuals derived from the null model. In each replication of cluster wild bootstrapping, new outcome scores are calculated with the residuals and the predicted values from the null model fit to the original dataset. The cluster wild bootstrap DGP is as follows: (MacKinnon, 2015):

$$\mathbf{T}_j^* = \mathbf{X}_{0j}\tilde{\boldsymbol{\beta}}_0 + v_j^*\mathbf{B}_j\tilde{\boldsymbol{e}}_j \tag{2.31}$$

Here, $\mathbf{T}_j^*$ denotes a $k_j \times 1$ vector of transformed effect sizes, v_j^* denotes the auxiliary random variable for cluster j, and $\mathbf{B}_j$ denotes a $k_j \times k_j$ adjustment matrix derived from the null model. If the residuals are not to be multiplied by any adjustment matrix prior to being multiplied by the auxiliary weights, $\mathbf{B}_j$ will be an identity matrix.

When calculating the test statistics in each replication of cluster wild bootstrapping, the CR adjustment matrices can be used to estimate the standard errors. MacK-

innon (2013), in the simulation studies, estimated the t-statistic in each bootstrap replication with HC1 to HC4 corrected standard errors and did not find any substantial difference across the tests in terms of Type I error rates. However, the tests differed in power with HC1 showing the highest power and HC2 to HC4 showing decreased power. By analogy, for cluster wild bootstrapping, using the CR0 or CR1 adjustment matrices might be appropriate. For single coefficient test, a t-test statistic can be estimated with the standard error calculated using the CR0 or CR1 adjustment matrices. For single coefficient test and for multiple-contrast hypothesis test, an F statistic can be calculated as $F = Q/q$, the simplest calculation of the statistic, with Q estimated using the CR0 or CR1 adjustment matrices. Using the CR0 correction compared to using the CR1 correction will not affect the calculation of p-values as the CR1 correction involves multiplying the CR0 type adjustment matrix by a constant (Djogbenou et al., 2019). However, the test statistics from the original data and the test statistics from the bootstrap replications should be calculated in the same way (Djogbenou et al., 2019).

Auxiliary Weights

Cluster wild bootstrapping involves sampling transformed residuals (Cameron et al., 2008). The residuals are multiplied by an auxiliary random variable that has a mean of 0 and variance of 1 (MacKinnon, 2015). There are two common types of auxiliary random weight variables (Cameron et al., 2008; MacKinnon, 2009, 2015). The first type of weights is the Mammen weights proposed by Mammen (1993). Mammen weights take on the following values (MacKinnon, 2015; Mammen, 1993):

$$v_m^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}) \end{cases} \tag{2.32}$$

with

$$\mathrm{E}[v_m^*] = 0, \mathrm{E}[v_m^{*2}] = 1, \mathrm{E}[v_m^{*3}] = 1, \mathrm{E}[v_m^{*4}] = 2 \tag{2.33}$$

A bootstrap residual for case i will have the opposite sign as the observed residual with probability of 0.72 (MacKinnon, 2009). Using the Mammen weights can ensure that the third moment of the bootstrap residuals is the same as that of the true errors (MacKinnon, 2015).

The second type of weights is the Rademacher weights proposed by Davidson and Flachaire (2008). Rademacher weights take on the following values (Davidson & Flachaire, 2008; MacKinnon, 2015):

$$v_r^* = \begin{cases} -1 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases} \tag{2.34}$$

with

$$\mathrm{E}[v_r^*] = 0, \mathrm{E}[v_r^{*2}] = 1, \mathrm{E}[v_r^{*3}] = 0, \mathrm{E}[v_r^{*4}] = 1 \tag{2.35}$$

The bootstrap errors are positive and negative with probability of 0.5 each (MacKinnon, 2015). Using the Rademacher weights can ensure that the second and the fourth moments of the bootstrap residuals are the same as those of the true errors (Davidson & Flachaire, 2008). According to Davidson and Flachaire (2008), the Rademacher weights impose symmetry and even if the errors are not actually symmetric, the weights ensure that the fourth moment of the residuals is estimated correctly. Davidson and Flachaire (2008) showed through simulations that wild bootstrap tests that used the Rademacher weights outperformed those that used the Mammen weights in terms of error in rejection probability, even when the distribution of true errors were skewed and not symmetric.

2.4.2 Methodological Studies on Cluster Wild Bootstrapping

Cameron et al. (2008) conducted simulation studies to examine the finite sample properties of CRVE compared to the three bootstrap techniques. The authors examined the Stata modified version of CR1 type CRVE (CR1S), and CR3 type CRVE. The CR1S adjustment matrix is specified as follows:

$$\mathbf{A}_j = \sqrt{[(m)(N-1)]/[(m-1)(N-p)]}\mathbf{I}_j \tag{2.36}$$

The authors generated data from a linear model with a single covariate and with clustered errors. The number of clusters generated ranged from 5 to 30 in increments of 5. The errors were generated to be homoskedastic and heteroskedastic and clusters were generated to be balanced and imbalanced. The authors used ordinary least squares estimation and conducted single coefficient tests using CRVE and the boot-

strap methods. The t-tests based on CRVE were compared to the standard normal distribution. The authors compared using bootstrapping to derive the standard errors and then conducting tests (bootstrap-se) to conducting hypothesis tests directly with bootstrapping (bootstrap-t). For cluster wild bootstrapping, the authors imposed the null hypothesis to calculate the residuals and predicted values but did not multiply the residuals by adjustment matrices. The Rademacher weights were used as the auxiliary random variable. The authors used the CR1S type CRVE to estimate the t-statistic in each bootstrap replication. The number of bootstrap replications was 399 and the number of Monte Carlo iterations was 1,000. Cameron et al. (2008) noted that using a smaller number of bootstrap replications was valid because the bootstrap replication error would have been negated over the Monte Carlo simulation iterations. Simulation results showed that bootstrap-t methods performed better in terms of controlling Type I error rates compared to bootstrap-se methods. Even with a small number of clusters and unbalanced groups, the cluster wild bootstrap-t method performed well in terms of controlling Type I error rates compared to the CRVE methods and other bootstrapping methods for single coefficient t-tests.

MacKinnon and Webb (2017) examined cases where cluster sizes were extremely unequal. The authors generated data from a difference-in-difference regression model. The authors ran four simulation studies, two with 50 clusters and two with 100 clusters. The cluster sizes were either equal or were dependent on the population of the 50 US states. The authors compared the CR1S type CRVE to cluster wild boostrapping. The t-tests based on CRVE were compared to a t-distribution with $m - 1$ degrees of freedom. For cluster wild bootstrapping, the authors imposed the null hypothesis to calculate the residuals and predicted values but did not multiply the residuals by adjustment matrices. The Rademacher weights were used as the auxiliary random variable. The authors used the CR1S correction to estimate the t-statistic in each bootstrap replication. The number of bootstrap replications was 399 and the number of simulation iterations was 400,000. The simulation results showed that the cluster wild bootstrap test performed best in terms of controlling Type I error rates. MacKinnon and Webb (2017) also generated data with varying number of treated clusters. The treatment variable, a cluster-level predictor variable, was generated with varying levels of imbalance. When the number of treated clusters was small, the study showed that CRVE over-rejected. Simulation results showed

that the cluster wild bootstrap test performed better compared to CRVE in terms of controlling Type I error rates. However, cluster wild bootstrapping under-rejected in conditions where the number of treated clusters was very small (less than 5 out of 50).

Djogbenou et al. (2019) conducted simulations to examine inferences based on the cluster wild bootstrap test. The authors generated data from a regression model with clustered errors. The cluster sizes were generated to be unequal. The number of clusters ranged from 10 to 200. Djogbenou et al. (2019) compared restricted and unrestricted cluster wild bootstrapping. The null hypothesis was imposed when calculating the residuals for restricted cluster wild bootstrapping whereas the null hypothesis was not imposed when calculating the residuals for unrestricted bootstrapping. The authors also compared the use of the Rademacher weights to the use of the Mammen weights. Djogbenou et al. (2019) did not multiply the residuals by adjustment matrices. The authors used the CR1S correction to estimate the t-statistic in each bootstrap replication. The authors compared the bootstrap tests to CR1S type CRVE. The t-tests based on CRVE were compared to a t-distribution with $m - 1$ degrees of freedom. The number of bootstrap replications in Djogbenou et al. (2019) was 399 and the number of simulation iterations was 400,000. The results showed that the restricted cluster wild bootstrap test with the Rademacher weights performed the best compared to the other tests in terms of Type I error rate and power even when the error terms were skewed.

MacKinnon (2015) examined CRVE and cluster wild bootstrapping in terms of confidence interval coverage. The authors generated data from a difference-in-difference regression model. The number of clusters generated ranged from 9 to 30 in increments of 3. MacKinnon (2015) compared CR1S type CRVE Wald statistic based confidence intervals to bootstrap based confidence intervals. For cluster wild bootstrapping, MacKinnon (2015) did not multiply the residuals by adjustment matrices. MacKinnon (2015) also compared several auxiliary weights for cluster wild boostrapping including the Rademacher weights, the Mammen weights, a six-point distribution proposed by Webb (2013), and several continuous weight distributions. The number of bootstrap replications in MacKinnon (2015) was 999 and the number of simulation iterations was 100,000. Simulation results from MacKinnon (2015) showed that bootstrap based confidence intervals provided more accurate coverage than the intervals

based on CRVE. Moreover, intervals based on cluster wild bootstrapping using the Rademacher weights had the most accurate coverage compared to intervals based on bootstrapping using any of the other weights even in the condition where the number of clusters was 9.

The studies reviewed here showed that the cluster wild bootstrap test maintained adequate Type I error rates even when the number of clusters was small, cluster sizes were imbalanced, and when a categorical covariate was highly imbalanced. Furthermore, the studies provided evidence in favor of using the Rademacher two-point weights even with very small number of clusters. The type of models evaluated in the simulation studies were linear regression models with clustering and difference-in-difference models that are relevant to the econometrics literature. Moreover, most of the studies examined conditions with very small number of clusters.

Pustejovsky and Tipton (2018) found that using the CR2 adjustment with the Satterthwaite degrees of freedom for fixed effects estimation with panel data models resulted in better Type I error rate control compared to using the CR1 or CR3 estimators without any further corrections. Thus, the CR2 estimator with the Satterthwaite degrees of freedom may have provided a better comparison than the CR1S or CR3 type CRVE examined in Cameron et al. (2008), MacKinnon and Webb (2017), Djogbenou et al. (2019), and MacKinnon (2015).

Based on the results of the studies reviewed in this section, cluster wild bootstrapping may be an attractive alternative to the HTZ test for meta-analysis. However, the performance of cluster wild bootstrapping in the context of meta-analysis has not been examined. Furthermore, no study has compared the performance of cluster wild bootstrapping to that of the CR2 correction method with the Satterthwaite degrees of freedom—the HTZ test. No study has examined whether multiplying residuals by adjustments matrices for cluster wild bootstrapping results in any difference in Type I error rates and power. Moreover, the performance of cluster wild bootstrapping for multiple-contrast hypothesis tests has not been examined.

2.4.3 Cluster Wild Bootstrapping in Meta-Analysis

Cluster wild bootstrapping has not been examined methodologically in a meta-analytic framework. However, it has been used in a handful of applied meta-analytic studies with dependent effect sizes and small number of studies. Examples include

McEwan (2015), in Review of Educational Research, examining school-based interventions on learning in developing countries; Gallet and Doucouliagos (2014), in the Annals of Tourism Research, examining income elasticity of air travel; Oczkowski and Doucouliagos (2015), in the American Journal of Agricultural Economy, examining the relationship between price of wine and its quality; and Ola and Menapace (2020), in the World Development journal, examining determinants of entry into high-value markets. All of the articles used bootstrapping to correct for single coefficient t-tests. None used it for multiple-contrast hypothesis tests.

2.5 Application

To demonstrate the differences between several methods for small sample correction to handle dependence in meta-analysis, I present analyses conducted using the data collected by Tanner-Smith and Lipsey (2015). I analyzed a random sample of 20 studies from the original data to evaluate the methods when the number of studies is small. I only included individual randomized control trials and only included effect sizes related to the alcohol consumption outcome. I selected studies that did not have any missing data. The number of effect sizes per study ranged from 1 to 42. I chose average age and the dependent variable measure as the moderators of interest. The dependent variable measures used in the primary studies were: blood alcohol concentration, combined measures, frequency of heavy use, frequency of use, peak consumption, and quantity of use. I compared the Naive F-test, the HTZ test, and the cluster wild bootstrap test with and without multiplying the residuals by the CR2 adjustment matrices. For single coefficient test, I focused on testing the coefficient for the age variable. For multiple-contrast hypothesis test, I focused on testing the difference in the effects of the interventions across the different dependent variable measures. The full model included both age and the dependent variable measure as predictors.

To run the analyses, I used R Version 4.0.0 (R Core Team, 2020). I used the `robu()` function from the `robumeta` package to fit the meta-regression models (Fisher et al., 2017). I used the `Wald_test()` function from the `clubSandwich` package to run small sample corrections for tests of both single coefficient and multiple-contrast hypothesis (Pustejovsky, 2020a).

For the Naive F-test, the variance-covariance matrix was set to "CR1" and the test to "Naive-F" in the `Wald_test()` function. For the HTZ test, the variance-covariance matrix was set to "CR2" and the test to "HTZ". I used the `constrain_zero()` function from the `clubSandwich` package to specify the appropriate contrast matrix.

To run cluster wild bootstrapping, the null and the full model were fit on the original dataset using `robu()`. For the single coefficient tests, the null model only included the dependent variable measure as a predictor and did not include age. For the multiple-contrast hypothesis tests, the null model only included age and did not include the dependent variable measure. The residuals and predicted values for the two types of tests were calculated from the respective null models. The CR2 adjustment matrices—one matrix per study—used to estimate the respective null models were extracted using an internal `clubSandwich` function. Transformed residuals were created by multiplying the adjustment matrices with the residuals from the respective null models. I ran separate bootstrap replications for the single coefficient test and for the multiple-contrast hypothesis test.

Following is a description of how one set of bootstrap replications was estimated. For each bootstrap replication, the Rademacher weights were sampled, set to be constant within a study. Using the Rademacher weights, two new outcomes were created: (1) one that added the residuals from the null model multiplied by the Rademacher weights to the predicted values from the null model (CWB), and (2) another that added the transformed residuals—those multiplied by the adjustment matrices—multiplied by the Rademacher weights to the predicted values from the null model (CWB Adjusted). The full model was then re-estimated with the two new outcome scores and the F-test statistics extracted. The covariance matrix was set as "CR1" and test as "Naive-F" in the `Wald_test()` function to estimate the F-test statistics for the relevant tests from the re-estimated models for each replication. The "Naive-F" option for the test argument calculates the F statistic as $F = Q/q$. The number of bootstrap replications was 999. To derive the bootstrap p-value, the proportion of times the bootstrap test statistic was greater than the F-test statistic from the original full model was calculated. The F-test statistic from the original full model was calculated using the CR1 matrices and the Naive F-test.

Table 2.3 below shows the results for the test of average sample age. The p-value from the HTZ test was greater than the p-value from the Naive F-test. The p-values

Table 2.3

Tanner-Smith and Lipsey (2015) Analysis: Tests for Age

Method	F	delta	df num	df denom	p
Naive-F	0.320	1.000	1.000	19.000	0.578
HTZ	0.204	1.000	1.000	2.983	0.682
CWB					0.685
CWB Adjusted					0.689

from the CWB and CWB Adjusted tests were slightly greater than the p-value from the HTZ test and greater than the p-value from the Naive F-test. The p-value from the CWB Adjusted test was slightly greater than the p-value from the CWB test.

For multiple-contrast hypothesis test, I tested whether the difference between each dependent variable measure level and the reference level (blood alcohol concentration) is equal to zero—whether the effect of the interventions vary by the dependent variable measure used. Table 2.4 shows the results for the multiple-contrast hypothesis tests. The p-value from the HTZ test was greater than the p-value from the Naive F-test. The p-values from the CWB and CWB Adjusted tests were smaller than the p-value from the HTZ test but greater than the p-value from the Naive F-test. The p-value from the CWB Adjusted test was slightly greater than the p-value from the CWB test.

Table 2.4

Tanner-Smith and Lipsey (2015) Analysis: Multiple-Contrast Hypothesis Tests for Dependent Variable Measure

Method	F	delta	df num	df denom	p
Naive-F	0.911	1.000	5.000	19.000	0.495
HTZ	0.473	0.615	5.000	6.384	0.786
CWB					0.701
CWB Adjusted					0.707

The p-values observed from different methods are different suggesting the methods may differ in Type I error rates and power. For multiple-contrast hypothesis tests, the CWB and CWB Adjusted tests had p-values that were smaller than the p-value from the HTZ test but greater than the p-value from the Naive F-test. The p-value from the CWB Adjusted test was slightly greater than the p-value from the CWB

test. For single coefficient tests, the p-values from the HTZ test and the CWB and CWB Adjusted tests were similar. Cluster wild boostrapping may maintain Type I error rates adequately and have more power than the HTZ test, especially for multiple-contrast hypothesis tests. The results shown in Tables 2.3 and 2.4 are based on only one study so I cannot draw any conclusions about the performance of the methods. Therefore, I conducted simulation studies to examine the different methods for handling dependence when the number of studies is small.

2.6 Purpose of Study

Although cluster wild bootstrapping offers a promising alternative to small sample corrections proposed by Tipton (2015) and Tipton and Pustejovsky (2015), its performance under a meta-analytic framework has not been evaluated in any methodological study. Thus, the goal of my dissertation was to examine whether using cluster wild bootstrapping improved upon the performance of the HTZ test, which corresponds to the CR2 adjustment method with the Satterthwaite degrees of freedom. I also examined the Naive F-test, which is the same as the CR1 adjustment method with $m - p$ degrees of freedom, as a baseline comparison method.

Tipton and Pustejovsky (2015) showed that the HTZ test provided adequate control of Type I error rate inflation. However, the results from Tipton and Pustejovsky (2015) showed that the HTZ test had below nominal Type I error rates indicating the possibility that the test may be conservative, especially for hypothesis tests with many constraints. Tipton and Pustejovsky (2015) did not directly compare power. In this dissertation, I examined whether cluster wild bootstrapping maintained Type I error rate control and provided improved power compared to the HTZ test.

Furthermore, cluster wild bootstrapping has been used in a few applied meta-analytic studies for tests of single coefficients. Therefore, in this dissertation, I examined single coefficient tests as well as multiple-contrast hypothesis tests. The results of the study can guide applied meta-analytic researchers to choose the most appropriate correction method for dependence when the number of studies in their meta-analysis is small.

I also examined if multiplying the residuals by the CR2 adjustment matrices when running cluster wild bootstrapping impacted Type I error rates and power. Davidson

and Flachaire (2008) and MacKinnon (2006) suggested multiplying the residuals by the HC2 correction when running wild bootstrapping. An extension of that suggestion to cluster wild bootstrapping has not been examined, even in the econometrics literature.

2.6.1 Research Question

The guiding research question for this dissertation was: To what extent do the CWB and the CWB Adjusted tests improve upon the HTZ test and the Naive F-test in terms of Type I error rates and power for tests of single coefficients and multiple-contrast hypothesis in the context of meta-analysis?

Chapter 3

Methods

I ran two simulation studies that used different design matrices. I evaluated the Naive F-test, the HTZ test, the CWB test, and the CWB Adjusted test in terms of Type I error rates and power. I ran the simulations in R Version 3.5.1 (R Core Team, 2020). I used the Stampede2 supercomputer provided by the Texas Advanced Computing Center (TACC) to run the simulations using supercomputing resources as bootstrapping can be computationally expensive. I used the `tidyverse` set of packages for data-munging (Wickham et al., 2019), the `mvtnorm` package for generating data (Genz et al., 2020), the `simhelpers` package for organizing the code for the simulation studies (Joshi & Pustejovsky, 2020), and the `Pusto` package for running the simulations using parallel processing (Pustejovsky, 2020b). Below, I outline the data generation process, estimation methods, experimental design, and performance criteria of the two simulation studies.

3.1 Study 1

3.1.1 Data Generation

The data generation process of this study followed that of Tipton and Pustejovsky (2015). I generated SMDs like Tipton and Pustejovsky (2015), as the SMD is the most common type of effect size measure for intervention research (Tipton, 2015). Each simulated data was comprised of m studies. A given study j contained k_j effect sizes.

Standardized Mean Differences

The SMDs were generated based on the distribution of summary statistics from primary studies. Study j consisted of two groups, treatment and control, and k_j correlated outcome variables. Let N_j denote the total sample size of study j assuming equal sample size per group, $\boldsymbol{\delta}_j$ denote the vector of true effect sizes for study j, and $\boldsymbol{\Sigma}_j$ denote the outcome variance-covariance matrix for study j. The formulation below requires that the outcomes have unit variance. Thus, $\boldsymbol{\Sigma}_j$ contained 1's along

the diagonal and the correlation between the outcomes on the off-diagonals. Let $\bar{\mathbf{y}}_{Tj}$ and $\bar{\mathbf{y}}_{Cj}$ denote the $k_j \times 1$ vectors of sample means for the treatment and control groups respectively and $\mathbf{S}_j$ denote the $k_j \times k_j$ sample variance-covariance matrix of the outcomes, pooled across the treatment and control groups. Assuming multivariate normality:

$$\bar{\mathbf{y}}_{Cj} \sim N\left(\mathbf{0}, \frac{2}{N_j}\boldsymbol{\Sigma}_j\right), \qquad \bar{\mathbf{y}}_{Tj} \sim N\left(\boldsymbol{\delta}_j, \frac{2}{N_j}\boldsymbol{\Sigma}_j\right) \tag{3.1}$$

and

$$(\bar{\mathbf{y}}_{Tj} - \bar{\mathbf{y}}_{Cj}) \sim N\left(\boldsymbol{\delta}_j, \frac{4}{N_j}\boldsymbol{\Sigma}_j\right) \tag{3.2}$$

The pooled sample covariance matrix follows a multiple of a Wishart distribution with $N_j - 2$ degrees of freedom and scale matrix equal to $\boldsymbol{\Sigma}_j$ (Anderson & Girshick, 1944):

$$(N_j - 2)\mathbf{S}_j \sim Wishart\,(N_j - 2, \boldsymbol{\Sigma}_j) \tag{3.3}$$

The denominators of the SMD estimates from study j—the pooled standard deviations of the outcome variables—were generated by simulating a single Wishart matrix, extracting the diagonal elements, dividing them by $N_j - 2$, and then taking the square root. Below, let s_{ij} indicate the pooled standard deviation corresponding to effect size i in study j. SMD i in study j was calculated as follows:

$$d_{ij} = \frac{\bar{y}_{Tij} - \bar{y}_{Cij}}{s_{ij}} \tag{3.4}$$

Hedges's g bias correction was applied to the SMD as follows (Borenstein & Hedges, 2019; Hedges, 1981):

$$g_{ij} = J(N_j - 2) \times d_{ij} \tag{3.5}$$

where

$$J(df) = 1 - \frac{3}{4 \times df - 1} \tag{3.6}$$

The variance of g_{ij} was calculated as follows (Borenstein & Hedges, 2019; Hedges, 1981):

$$\mathrm{Var}(g_{ij}) = [J(N_j - 2)]^2 \left(\frac{4}{N_j} + \frac{d_{ij}^2}{2(N_j - 2)}\right) \tag{3.7}$$

Covariates

I used the design matrix generated by Tipton and Pustejovsky (2015) in my simulation. Tipton and Pustejovsky (2015) generated five covariates, two binary and three continuous. The first binary covariate, X_1, is a study-level covariate with large imbalance, equaling 1 in 15% of the studies. The second binary covariate, X_2, is an effect size-level covariate, equaling 1 in 10% of the effect size estimates overall and in 0 to 20% of the effect size estimates within a study. X_3 is a normally distributed study-level covariate, X_4 is a normally distributed continuous effect size-level covariate, and X_5 is a continuous, highly skewed effect size-level covariate. Tipton and Pustejovsky (2015) noted that these types of variables are common in applied meta-analyses, with the covariates that have large imbalances or high skewness representing the worst cases.

The generated data has 200 rows, containing the design matrix for 200 effect sizes, with 10 rows per study, totaling 20 studies. Following the procedures of Tipton and Pustejovsky (2015), in cases where there were more than 20 studies, the rows of the design matrix were repeated. For studies with less than 10 effect sizes, the first k_j rows from the design matrix were selected.

To examine the Type I error rates, I simulated SMDs that were unrelated to the covariates. To examine power, the experimental design contained conditions with varying magnitude of relationships between each covariate and the effect sizes. The conditions are detailed in the Experimental Design section below.

Meta-Analytic Data

To generate the SMDs based on the distribution of summary statistics, the number of effect sizes, the sample size, the outcome covariance matrix, and the true effect size parameters were generated for each study. For this simulation, the sample sizes of the treatment and control groups were assumed to be equal.

The number of effect size per study was generated as follows:

$$k_j \sim min(1 + Poisson(4), 10) \tag{3.8}$$

The number varied from 1 to 10 effect sizes per study. The range of the number of effect sizes followed that from Tipton and Pustejovsky (2015) and captured the range

seen in real meta-analytic data. Tipton et al. (2019), in a review of 64 meta-analyses published in education and psychology journals, found an average of 4.5 (SD = 5.6) effect sizes per study. All conditions in my simulation study included varied number of effect sizes to reflect real meta-analytic data.

The sample size per study was generated as follows:

$$N_j \sim min(20 + 2 \times Poisson(30), 200) \tag{3.9}$$

The total sample size per study was set to range from 20 to 200. I simulated sample sizes based on Equation 3.9 for varying number of studies and found a median sample size of 80 with the 25^{th} and 75^{th} percentiles equal to 72 and 88 respectively. The per group sample size per study generated by Tipton and Pustejovsky (2015) ranged from 32 to 130. Rodgers and Pustejovsky (2019) examined the sample size per study in a meta-analysis conducted by Lehtonen et al. (2018) on the effect of bilingualism that included 152 studies. Rodgers and Pustejovsky (2019) found a median sample size of 48 and range from 12 to 343. Furthermore, Park and Beretvas (2016) reviewed meta-analyses published in the Journal of Review of Educational Research from 2010 to 2015 and found a median study sample size of 171 and the 25^{th} and 75^{th} percentiles of 66 and 702. Many of the primary studies reviewed by Park and Beretvas (2016) likely included designs that examined the effects of cluster-level interventions—for example, effects of interventions administered to students nested within schools that were randomized (Rodgers & Pustejovsky, 2019). Therefore, the sample sizes of interest may be smaller than those reported by Park and Beretvas (2016). The range of total sample sizes per study that I generated roughly followed that from Tipton and Pustejovsky (2015) and reflected the range of study sample sizes found in published meta-analyses.

The correlation between the outcomes per study, r_j, assuming that the correlation between pairs of outcomes is constant within studies, was generated as follows:

$$r_j \sim Beta\left(\rho\nu, (1-\rho)\nu\right) \tag{3.10}$$

where ν controlled the variability of r_j across studies as $\text{Var}(r_j) = \rho(1-\rho)/(1+\nu)$. Thus, smaller ν values correspond to more variable correlations. The covariances in $\mathbf{\Sigma}_j$ were populated with r_j. The variances were populated with 1's. The value for

ν was set to 50 to generate moderate amount of variance in the correlation values across studies. The value for ρ differed across conditions.

True effect size parameters were generated based on the relationships between the covariates and the effect sizes. Below, let δ_{ij} denote the true SMD i in study j, $\mathbf{x}_{ij}$ denote a $1 \times p$ vector of covariates for SMD i in study j, $\boldsymbol{\beta}$ denote a $p \times 1$ vector of meta-regression coefficients, and v_j denote the between-study sampling error term with mean of 0 and variance of τ^2. Here, $\mathbf{x}_{ij}$ was generated as described in the Covariates section above. The true effect size parameters were generated as follows:

$$\delta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + v_j, \quad \text{where} \quad v_j \sim N(0, \tau^2) \tag{3.11}$$

With the generated number of effect sizes, sample size, outcome covariance matrix, and the vector of true effect sizes for each study, I generated correlated effects meta-analytic data. Each simulated data contained SMDs and associated variances, and the corresponding design matrix for k_j effect sizes for each of the specified number of studies.

3.1.2 Estimation Methods

For computational efficiency, I used a version of the `robu()` function from the `robumeta` package to estimate correlated effects weighted least squares meta-regression models. The version of the function that I used only runs weighted least squares estimation for the correlated effects working model and omits checks that are necessary when the function is used with real data but not when the function is used with simulated data. I calculated the Naive F and HTZ p-values using the `Wald_test()` function from the `clubSandwich` package (Pustejovsky, 2020a). I used the `Wald_test()` function for both single coefficient tests and multiple-contrast hypotheses tests.

I fit a model including the main effects of all of the covariates on each simulated dataset. This model was the full model for all of the tests. Tests for the single coefficients were estimated from this model for each of the five covariates. For multiple-contrast hypotheses tests, I examined all combinations of 2 to 5 regression coefficients for the covariates, totaling 26 different sets of coefficients. In the conditions designed to test power, I only examined sets that included the covariates with coefficients not set to be equal to zero when generating the true effect sizes. The null hypothesis

was whether each set of the coefficients equaled zero. Tipton and Pustejovsky (2015) found that the small sample correction methods performed differently depending on the number of contrasts used. Therefore, I examined all possible contrasts.

Naive F-Test

To run the Naive F-test, I used the `Wald_test()` function. The variance-covariance matrix was specified as "CR1" and the test specified as "Naive-F".

HTZ Test

To run the HTZ test, I also used the `Wald_test()` function. The variance-covariance matrix was specified as "CR2" and the test specified as "HTZ".

Cluster Wild Bootstrapping

With the original dataset from each simulation iteration, I estimated the null model and the full model with random effects weighted least squares estimation using the computationally less expensive version of `robu()`. The full model included all the five covariates. The null model included all covariates except for the ones being tested. Residuals and predicted values were calculated from the null model. I aimed to examine whether multiplying the residuals by the CR2 adjustment matrices from the null model impacted Type I error rates and power. Therefore, I ran the cluster wild bootstrap with and without the multiplication of the residuals by the CR2 adjustment matrices when transforming the residuals—the CWB Adjusted test and the CWB test respectively.

For each bootstrap replication, a random auxiliary weighting variable was sampled, set to be constant within a study. The original residuals and the CR2 transformed residuals were multiplied by the random auxiliary weights. For this simulation, I used the Rademacher weights as Djogbenou et al. (2019) and MacKinnon (2015) showed that using the Rademacher weights for cluster wild bootstrapping outperformed using the other types of weights even when the number of clusters was small. Two new outcome scores were then calculated by adding the weighted residuals—the non-adjusted weighted residuals and the CR2 adjusted weighted residuals—to the predicted outcome scores from the null model. Using the new outcome scores, the

full model was re-estimated. To estimate the F statistics for the relevant tests from the re-estimated full models, I calculated the CR0 adjustment matrices, calculated the Q statistic following Equation 2.24, and converted the test statistic as $F = Q/q$.

The number of bootstrap replications was set to 399 following Cameron et al. (2008), MacKinnon and Webb (2017), and Djogbenou et al. (2019). After running all the replications, the p-value was calculated as the proportion of times the bootstrap test statistic was greater than the test statistic from the original full model for the CWB and the CWB Adjusted tests (Cameron et al., 2008; MacKinnon, 2009):

$$p = \frac{1}{R} \sum_{r=1}^{R} I\left(F^{(r)} > F\right) \tag{3.12}$$

3.1.3 Experimental Design

The experimental design for this study generally followed that of Tipton and Pustejovsky (2015), but with fewer conditions, as the bootstrapping procedure is computationally intensive. The parameters examined in this study included the number of studies (m), between-study heterogeneity in the effect sizes (τ), within-study correlation between outcomes (ρ), and the regression coefficients used to generate effect sizes ($\boldsymbol{\beta}$).

Table 3.1 shows the different conditions that I examined. The factors in this study included 4 values for the number of studies $\times$ 2 τ values $\times$ 2 ρ values $\times$ 11 sets of meta-regression coefficients, totaling 176 conditions. The conditions are discussed below. The number of simulation iterations was 2,400.

Number of Studies

The number of independent studies (m) was set to 10, 20, 40, and 80 to cover realistically small to moderate sample sizes. These values are a subset of the number of studies evaluated by Tipton and Pustejovsky (2015). Tipton and Pustejovsky (2015) did not find much difference in the performance of the Naive F-test and the HTZ test when the number of studies was set to 80 versus 100. Therefore, I only included conditions with up to 80 studies. In a review of meta-analyses across education and psychology journals, Tipton et al. (2019) found an average of 65.5 (SD = 65.8) studies. The conditions used in my dissertation covered the range of the number of studies

Table 3.1

Data Generating Conditions: Study 1

Conditions	Values
Number of studies (m)	10, 20, 40, 80
Between-study heterogeneity (τ)	0.1, 0.3
Correlation between outcomes (ρ)	0.5, 0.8
Regression coefficients($\boldsymbol{\beta}$)	A. $\beta_0 = 0.3$, $\beta_1, ..., \beta_5 = 0$
	B. $\beta_0 = 0.3$, $\beta_1 = 0.1$ or 0.5, $\beta_2, ..., \beta_5 = 0$
	C. $\beta_0 = 0.3$, $\beta_2 = 0.1$ or 0.5, β_1 and $\beta_3, .., \beta_5 = 0$
	D. $\beta_0 = 0.3$, $\beta_3 = 0.1$ or 0.5, $\beta_1, \beta_2, \beta_4, \beta_5 = 0$
	E. $\beta_0 = 0.3$, $\beta_4 = 0.1$ or 0.5, $\beta_1, ..., \beta_3$ and $\beta_5 = 0$
	F. $\beta_0 = 0.3$, $\beta_5 = 0.1$ or 0.5, $\beta_1, ..., \beta_4 = 0$

found in applied meta-analyses, especially on the small sample side.

Between-Study Heterogeneity

The values for τ were set to 0.1 and 0.3 representing small to large heterogeneity (Pigott, 2012). Unlike Tipton and Pustejovsky (2015), I used τ instead of I^2 values to measure between-study heterogeneity due to the drawbacks of interpreting I^2 values as discussed in the Characterizing Variability section above (Borenstein et al., 2017). Pigott (2012) suggested $0.33\sigma^2$, σ^2, and $1.33\sigma^2$ for small, moderate, and large τ^2 values. I examined possible ranges of σ^2 values that would be generated given the range of sample size per study in my simulation and settled on an average σ^2 value of 0.05 to calculate τ. The results from Tipton and Pustejovsky (2015) showed that the maximum Type I error rates increased only slightly with increase in the I^2 values even when an incorrect fixed effects model was used. Therefore, I only examined a small set of τ values in this simulation study.

Correlation between Outcomes

The values for ρ were set to 0.5 and 0.8. The `robumeta` package uses 0.8 as the default value for the within-study correlation between effect sizes when estimating meta-regression models. I included ρ value of 0.5 as one of the conditions to examine the performance of the methods when the assumed value for the correlation between effect sizes in `robumeta` is incorrect. The correlation values between outcomes were

drawn from a beta distribution as specified in Equation 3.10 with mean ρ value and a parameter ν that dictated the variability of the correlations across studies. The value for ν was set to 50 to introduce moderate amount of variation in the correlation values across studies.

Hedges et al. (2010) showed that the estimation of τ^2 and the estimation of the standard errors for regression coefficients using RVE were robust to differences in the values of the within-study correlation between effect sizes. Results from Tipton (2015) showed that the differences in the Type I error rates across different values of ρ and τ^2 were more pronounced for the CR1 estimator than the CR2 and CR3 estimators and were more pronounced in Study 1, which included one covariate at a time in the regression models, than in Study 2, which included sets of covariates in the models. The differences were even less pronounced for estimators when the Satterthwaite degrees of freedom was used. Furthermore, Tipton and Pustejovsky (2015) did not find any relationship between their results and the ρ values. Therefore, I only examined a small set of values for ρ.

Meta-Regression Coefficients

Let β_0 indicate the intercept, and β_1 to β_5 indicate the coefficients related to each of the five moderator variables. To examine the Type I error rates, one of the conditions specified β_1 to β_5 all equal to zero. To examine power, I generated the data with the coefficients for each of the five covariates taking on each of the following values: 0.1 and 0.5. All the other coefficients, except the intercept were set to 0. β_0 was set to 0.3 in all of the conditions. Table 3.1 shows the different conditions.

The coefficients for the moderators can be thought of as effect sizes—change or difference in the outcome in terms of standard deviation—as the outcome variable in the meta-regression models consists of SMDs. In a review of studies on educational interventions, Kraft (2020) examined 747 studies that included 1,942 effect sizes. Kraft (2020) reported that the median effect size estimate across all the studies was 0.1 and the 90^{th} percentile was 0.5. These two values were used as regression coefficients in my simulation study to generate the true effect sizes.

3.1.4 Performance Criteria

To evaluate the performance of the CWB and the CWB Adjusted tests compared to the HTZ test and the Naive F-test, the performance criteria of interest were Type I error rate and power. Type I error rate and power both capture the proportion of times that the p-values derived from simulation iterations is below a specified α level—the proportion of times the null hypothesis is rejected (Morris et al., 2019). Let K denote the number of simulation iterations, and p_k denote the p-value from simulation replication k, for $k = 1, ..., K$. The rejection rate for a specified α level is defined as:

$$\rho_\alpha = \Pr(p_k) < \alpha \tag{3.13}$$

The rejection rate is calculated as:

$$r_\alpha = \frac{1}{K} \sum_{k=1}^{K} I(p_k < \alpha) \tag{3.14}$$

The Monte Carlo standard error for the estimate of the rejection rate, which captures the level of uncertainty in the estimation of the rejection rate, is calculated as (Morris et al., 2019):

$$r_\alpha MCSE = \sqrt{r_\alpha(1 - r_\alpha)/K} \tag{3.15}$$

Following Tipton and Pustejovsky (2015), I examined α values of 0.01, 0.05, and 0.10 with 0.05 being the most conventional.

3.2 Study 2

This study was designed to examine whether the results from Study 1 would generalize to analyses with a different design matrix. The design matrix in this study included one nominal covariate with multiple categories. Applied meta-analysts are likely to conduct a multiple-contrast hypothesis test to examine whether the effect of an intervention is similar across different categories of a moderator variable. Thus, I designed this study to examine the methods in contexts that would be useful to applied meta-analysts. The data-generation procedure, estimation methods, and performance criteria calculations for Study 2 were mostly similar to those for Study 1. The differences are discussed below.

3.2.1 Data Generation

Instead of using the design matrix created by Tipton and Pustejovsky (2015), I generated a single covariate that had 3, 4, or 5 categories. The covariate was generated to vary either at the study level or at the effect size level. Applied meta-analysts are likely to encounter both study-level and effect size-level variables—for example, a study-level covariate could be the type of experimental design and an effect size-level covariate could be the type of outcome measure. Further, results from Tipton (2015) showed that small sample correction methods performed differently for different types of covariates. Therefore, I generated both types. I constrained each category to be present in at least 2 studies for the study-level covariate type and in at least 2 effect size estimates for the effect size-level covariate type. Each category had equal probability of being sampled.

The number of effects per study, the sample size per primary study, and the correlation between outcomes were generated exactly like in Study 1. The number of effects per study was set to vary between 1 to 10. The total sample size per primary study was set to range from 20 to 200. The correlation between outcomes per study was generated based on a beta distribution with mean ρ value and a ν parameter that controlled the variability of the correlation values between studies.

3.2.2 Estimation Methods

For Study 2, I only tested multiple-contrast hypothesis examining whether the effect of the intervention was equal across all of the categories. The null hypothesis of the multiple-contrast hypothesis test was that the effect of the intervention did not vary across different categories.

3.2.3 Experimental Design

The experimental design of this study mostly followed that from Study 1. The major difference was the specification of the regression coefficients used to generate the true effect sizes. For Type I error, all the β values, except for the intercept, were set to 0. For power analysis, all β values, expect for the intercept and β_1, were set to 0. The value for β_1, which represents the difference between the effect of the second category and that of the first category, was set to 0.1, 0.3, or 0.5 to study power

curves. The intercept was set to 0.3 in all conditions as in Study 1. Furthermore, I added another condition, covariate type, which indicated whether to generate a study-level covariate or an effect size-level covariate.

Table 3.2 shows the different conditions that I examined in Study 2. The factors in this study included 4 values for the number of studies $\times$ 2 τ values $\times$ 2 ρ values $\times$ 3 different number categories in the covariate $\times$ 2 covariate types $\times$ 4 sets of meta-regression coefficients, totaling 384 conditions. The number of simulation iterations was 2,400.

Table 3.2

Data Generating Conditions: Study 2

Conditions	Values
Number of studies (m)	10, 20, 40, 80
Between-study heterogeneity (τ)	0.1, 0.3
Correlation between outcomes (ρ)	0.5, 0.8
Number of categories	3, 4, 5
Covariate type	study-level, effect size-level
Regression coefficient(β_1)	0.0, 0.1, 0.3, 0.5

3.3 Number of Iterations

For both studies, I set the number of iterations to 2,400. The number was chosen as a compromise between computing time and desired level of precision. The time required to run both studies on TACC with 2400 iterations each was approximately 300 hours, with Study 1 requiring around 290 hours and Study 2 requiring around 10 hours. For Type I error rate, I calculated the MCSE for 2400 iterations across the different nominal α levels. For a nominal α level of 0.05, the MCSE is 0.004. For a nominal α level of 0.01, the MCSE is 0.002. For a nominal α level of 0.10, the MCSE is 0.006. For a rejection rate of 0.5 for power, the MCSE is 0.01. The MCSEs calculated are relatively much smaller than the nominal rejection rates.

Chapter 4

Results

In this chapter, I present results from the two simulation studies. I examined the Type I error rates for all the tests. For power, I only examined tests that maintained adequate Type I error rates. I also discuss sensitivity of the results to differing values of τ and ρ. I used R version 4.0.3 to analyze the results of the simulation studies (R Core Team, 2020). I used the `tidyverse` set of packages to clean, analyze, and visualize the results (Wickham et al., 2019), and the `patchwork` package to combine plots (Pedersen, 2020).

4.1 Study 1

4.1.1 Type I Error Rates

The box-plots for Type I error rates show the range of Type I error rates by the number of studies (m), the number of contrasts (q), and the nominal α levels. The rejection rates range over the rates from specific tests and variables, and the rates for different τ and ρ values. The solid lines indicate the nominal α levels and the dashed lines indicate the upper bounds for simulation error. I calculated the bound as:

$$U_B = \alpha + 1.96 \times MCSE_\alpha \tag{4.1}$$

Here, α denotes the nominal α level and $MCSE_\alpha$ denotes the MCSE at the nominal α level. Tests that have Type I error rates that fall below the simulation error bounds are considered to maintain Type I error rates adequately.

Naive F-test

In Study 1, I examined the Naive F-test as a baseline comparison method. The results from the simulation study replicated findings from Tipton and Pustejovsky (2015). Figure 4.1 shows the Type I error rates of the Naive F-test. The Type I error rates were higher than the nominal α level, especially for conditions with lower number of studies, i.e., 10 or 20 studies. The rates were also higher than the nominal

rates for tests of higher number of contrasts. Even in conditions with the number of studies equal to 80, the median Type I error rates were higher than the nominal rates.

Cluster Wild Bootstrapping versus HTZ

Figures 4.2, 4.3, and 4.4 show the range of Type I error rates of the CWB, the CWB Adjusted, and the HTZ tests for the nominal α levels of 0.01, 0.05, and 0.10 respectively. I did not include the results of the Naive F-tests in these graphs as the Naive F-test exhibited much higher Type I error rates than the tests examined in these plots.

The results for the HTZ test replicated findings from Tipton and Pustejovsky (2015). Type I error rates of the HTZ test tended to be below the nominal level. For conditions with smaller number of studies and for tests of higher number of contrasts, the Type I error rates of the HTZ test were far below the nominal level. For example, in conditions with 10 studies and for tests of 5 contrasts, the Type I error rates of the HTZ test were near 0. Even in conditions with 80 studies, the Type I error rates of the HTZ test were slightly below the nominal level. In contrast, the CWB and the CWB Adjusted tests had Type I error rates near the nominal rate across all conditions. The Type I error rates of the two CWB tests were very similar across all conditions. The pattern of results were similar across all three nominal α levels.

4.1.2 Power

I examined results for power in several different ways. First, I examined absolute power levels for tests that maintained Type I error rates adequately. Second, I examined power ratios as a way to compare CWB to the current standard test, the HTZ test. Third, I disaggregated the absolute power and power ratio results by sets of covariates tested as different types of covariates had different power levels.

The first set of box-plots for absolute power show the range of power of the tests by the number of studies, the number of contrasts (q), the regression coefficient used to generate the true effect sizes (β), and the nominal α levels. The rejection rates range over the rates from specific tests and variables, and the rates for different τ and ρ values. The first set of box-plots for power ratio show the range of the power ratio

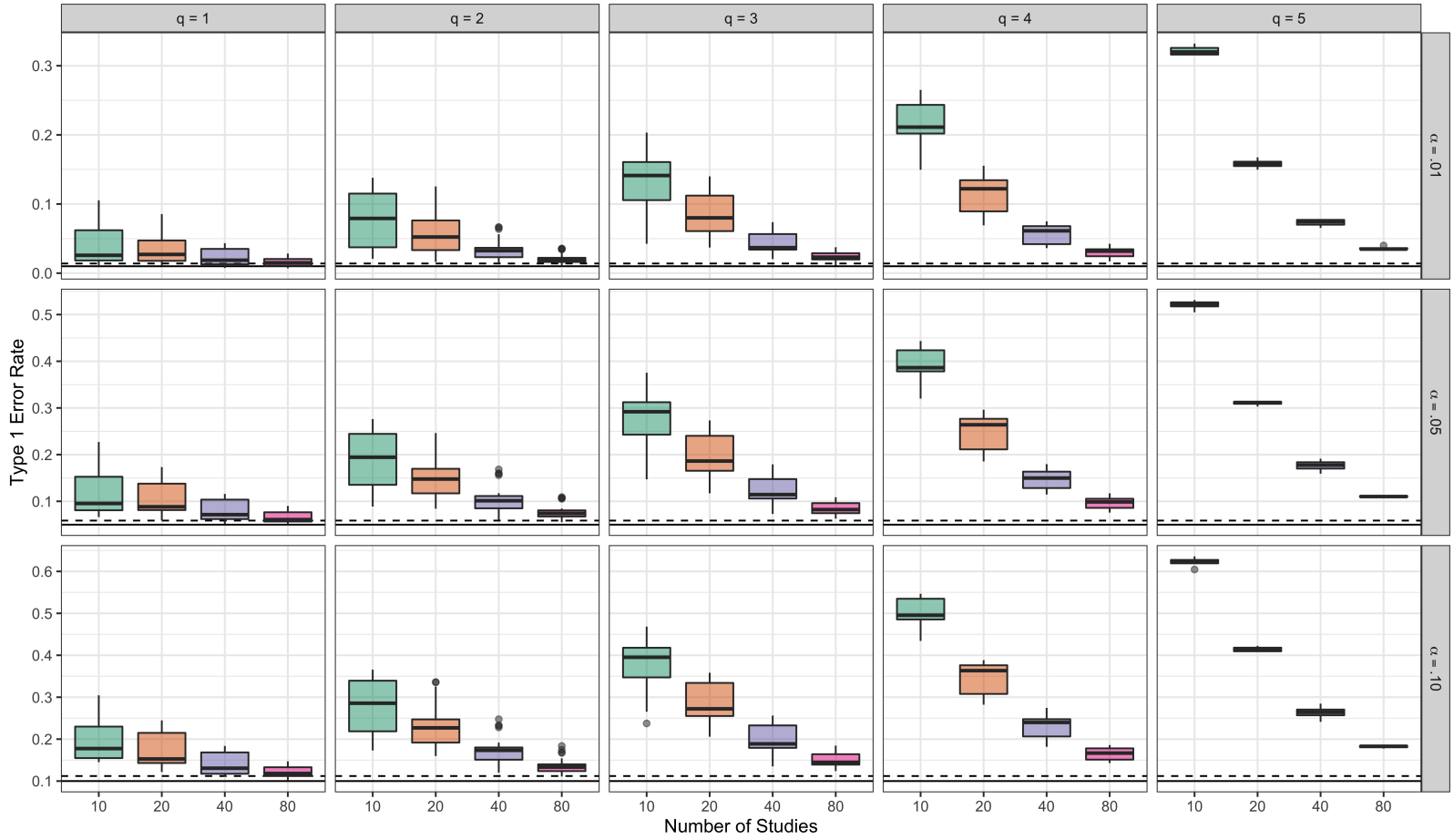

Figure 4.1. Study 1: Type I error rates of the Naive F-test by the number of studies, the number of contrasts (q), and the nominal α level. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error. The maximum MCSE across all conditions and for each of the nominal α levels was 0.01.

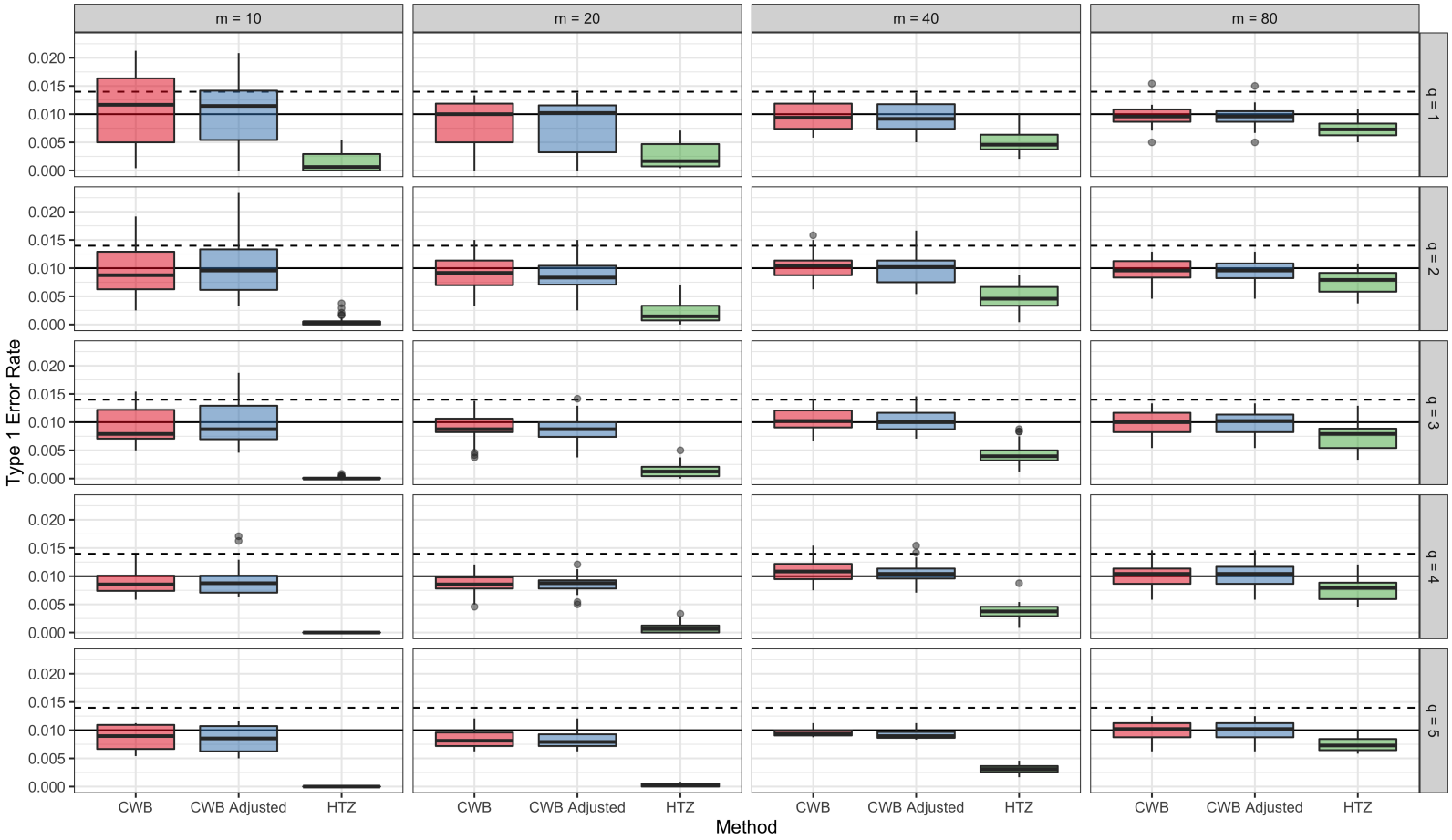

Figure 4.2. Study 1: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m) and the number of contrasts (q) for nominal α level of 0.01. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error. The maximum MCSE for the CWB test and the CWB Adjusted test across all conditions was 0.003, and the maximum for the HTZ test was 0.002.

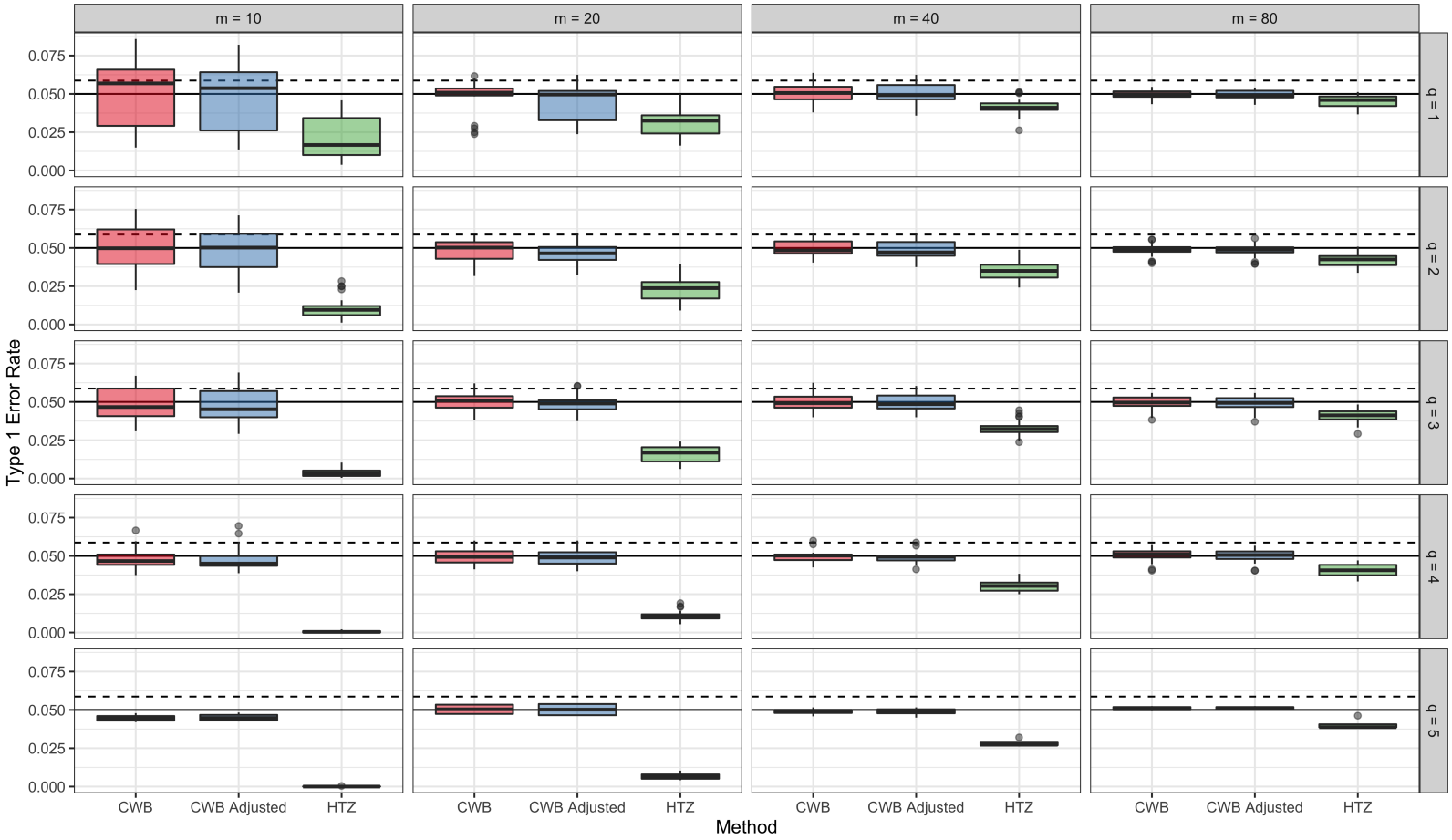

Figure 4.3. Study 1: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m) and the number of contrasts (q) for nominal α level of 0.05. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error. The maximum MCSE for the CWB test and the CWB Adjusted test across all conditions was 0.006, and the maximum for the HTZ test was 0.005.

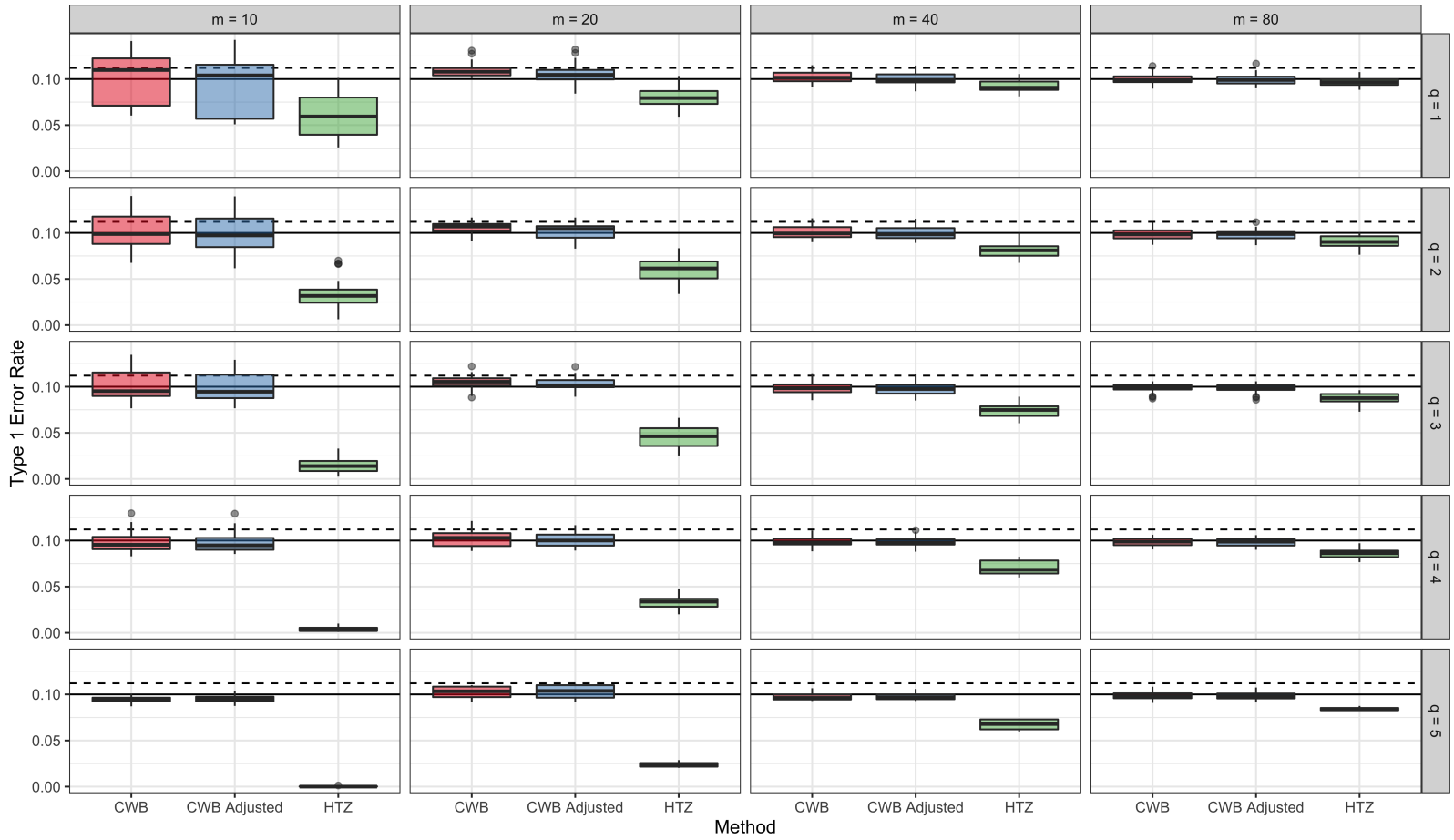

Figure 4.4. Study 1: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m) and the number of contrasts (q) for nominal α level of 0.10. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error. The maximum MCSE for the CWB test and the CWB Adjusted test across all conditions was 0.007, and the maximum for the HTZ test was 0.006.

instead of absolute power. The second set of box-plots for absolute power and power ratio show disaggregated results by sets of covariates that were tested.

Absolute Power

I did not consider the Naive F-test in the evaluation of power as the Naive F-test had extremely high Type I error rates. Figures 4.5, 4.7, and 4.9 show absolute power of the CWB, the CWB Adjusted, and the HTZ tests for the nominal α levels of 0.01, 0.05, and 0.10 respectively.

The median power of the HTZ test was generally lower than those of the CWB and the CWB Adjusted tests. The power of the HTZ test was particularly low for conditions with lower number of studies and for tests of higher number of contrasts. The power of the CWB and the CWB Adjusted tests were similar across all conditions.

Power Ratio

Figures 4.6, 4.8, and 4.10 show the ratio of power of the HTZ test over the power of the CWB test for the nominal α levels of 0.01, 0.05, and 0.10 respectively. I only examined the CWB test as the performance of the CWB and the CWB Adjusted tests were nearly identical in terms of power. In the plots, ratios below the solid lines at 1 indicate the loss of power from using the HTZ test rather than the CWB test.

The results show that the CWB test had higher power than the HTZ test across most conditions. In particular, for conditions with smaller number of studies (i.e., 10 or 20 studies) and for tests of higher number of contrasts, the CWB test had much higher power than the HTZ test. For example, the CWB test had 100% more power than the HTZ in conditions with 10 studies and for tests of 5 contrasts. One exception is the condition with 10 studies for tests of 1 or 2 contrasts. In these particular cases, the box-plots are slightly split across the solid line at 1 indicating that the HTZ test had higher power compared to the CWB test in some conditions or for some particular tests. Additionally, power losses for single coefficient tests were smaller compared to those for multiple-contrast hypotheses tests. For tests of 4 or 5 contrasts in conditions with smaller number of studies, power losses were high. Overall, the CWB test had higher power compared to the HTZ test.

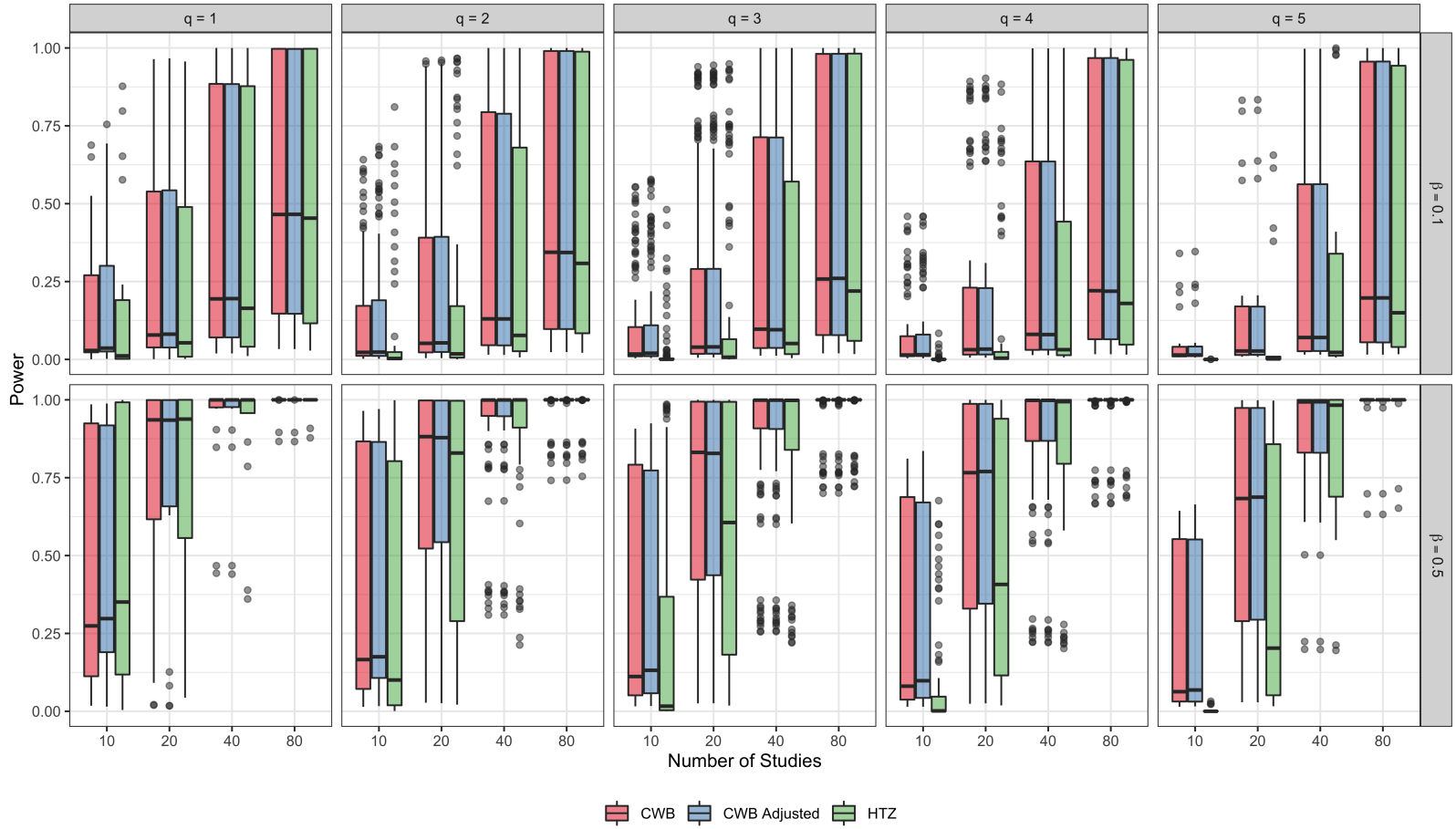

Figure 4.5. Study 1: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the number of contrasts (q), and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.01. The maximum MCSE for each of the tests across all conditions was 0.01.

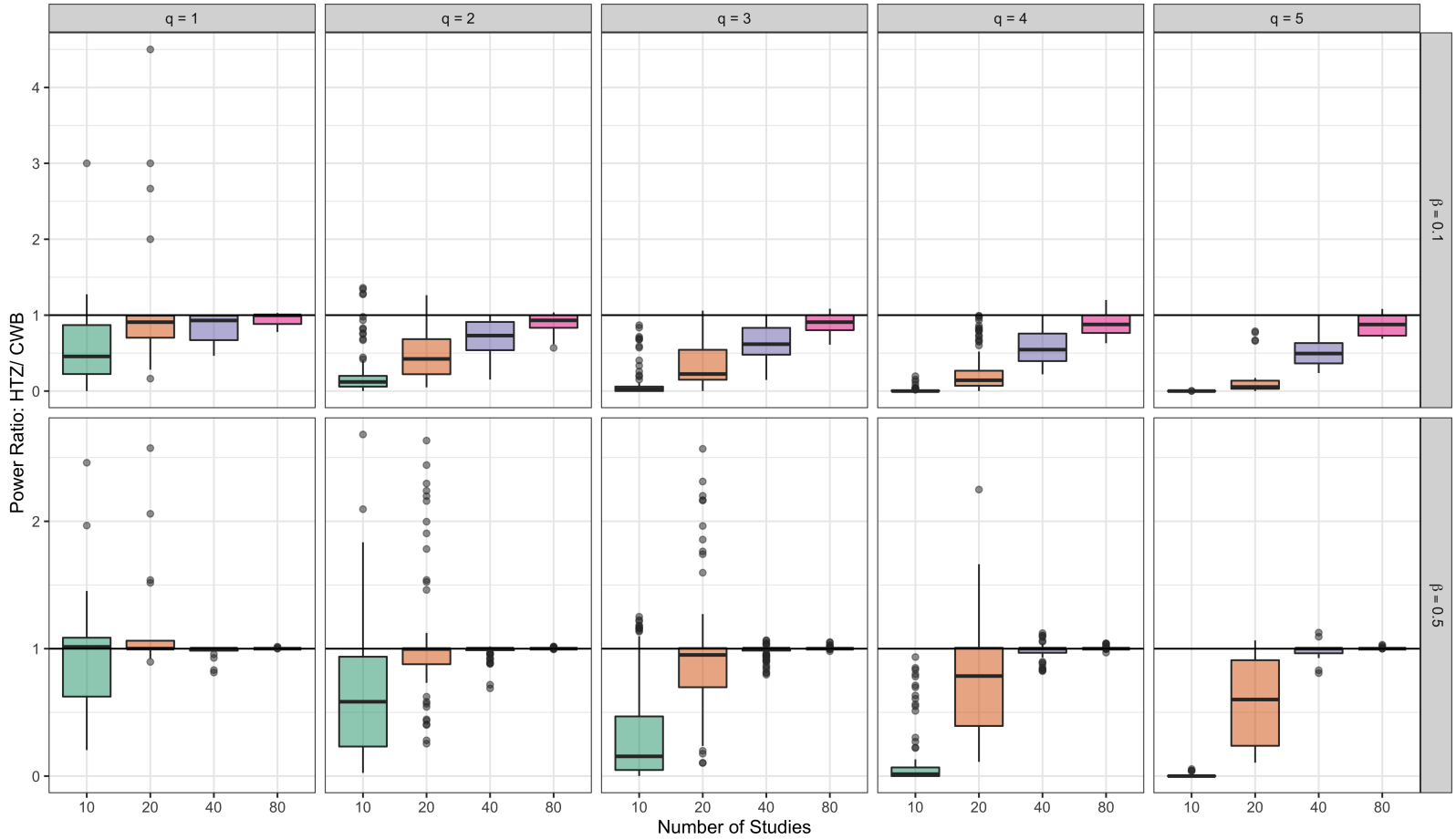

Figure 4.6. Study 1: Ratio of power of the HTZ test and the CWB test by the number of studies, the number of contrasts (q), and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.01. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

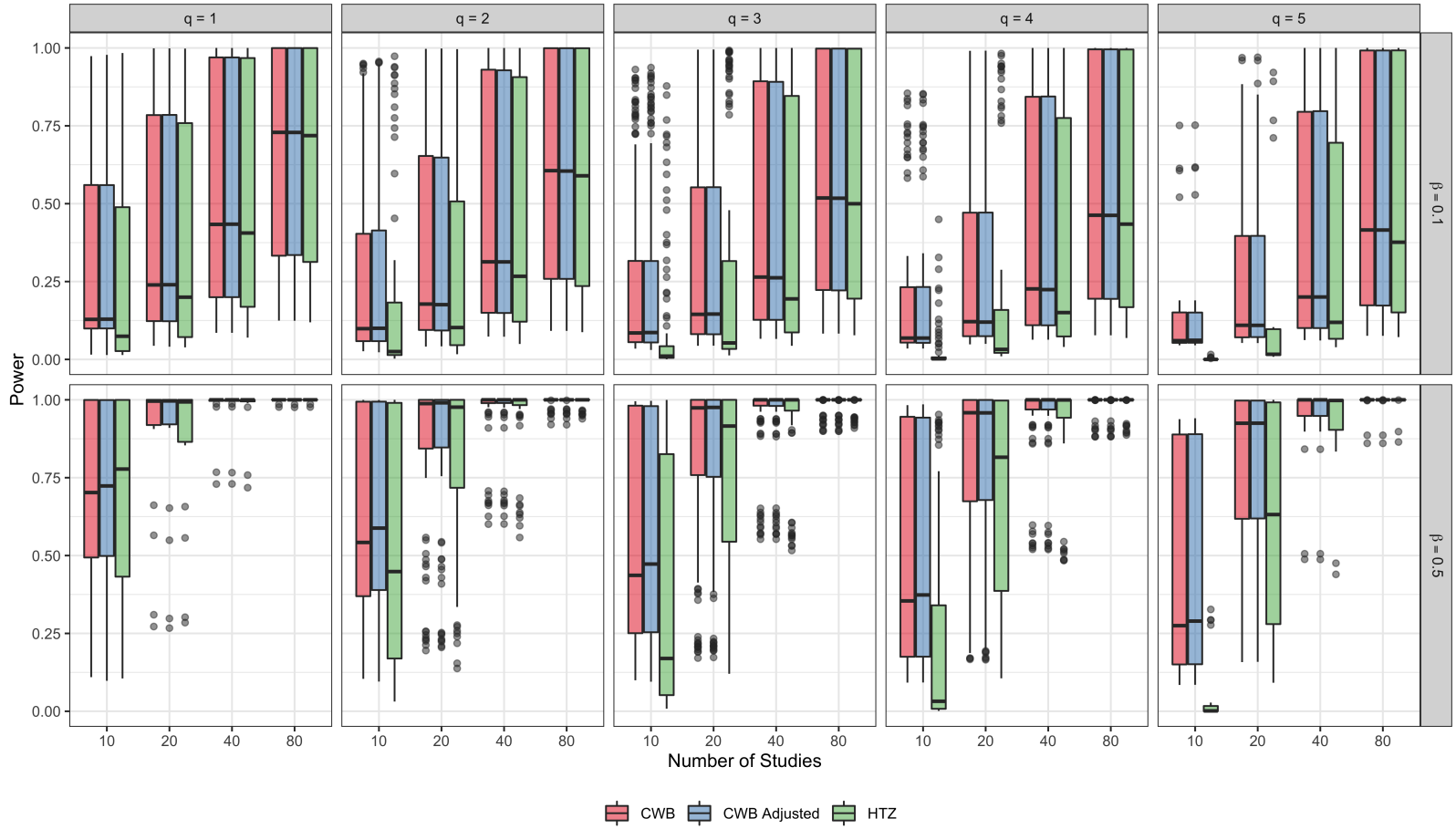

Figure 4.7. Study 1: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the number of contrasts (q), and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The maximum MCSE for each of the tests across all conditions was 0.01.

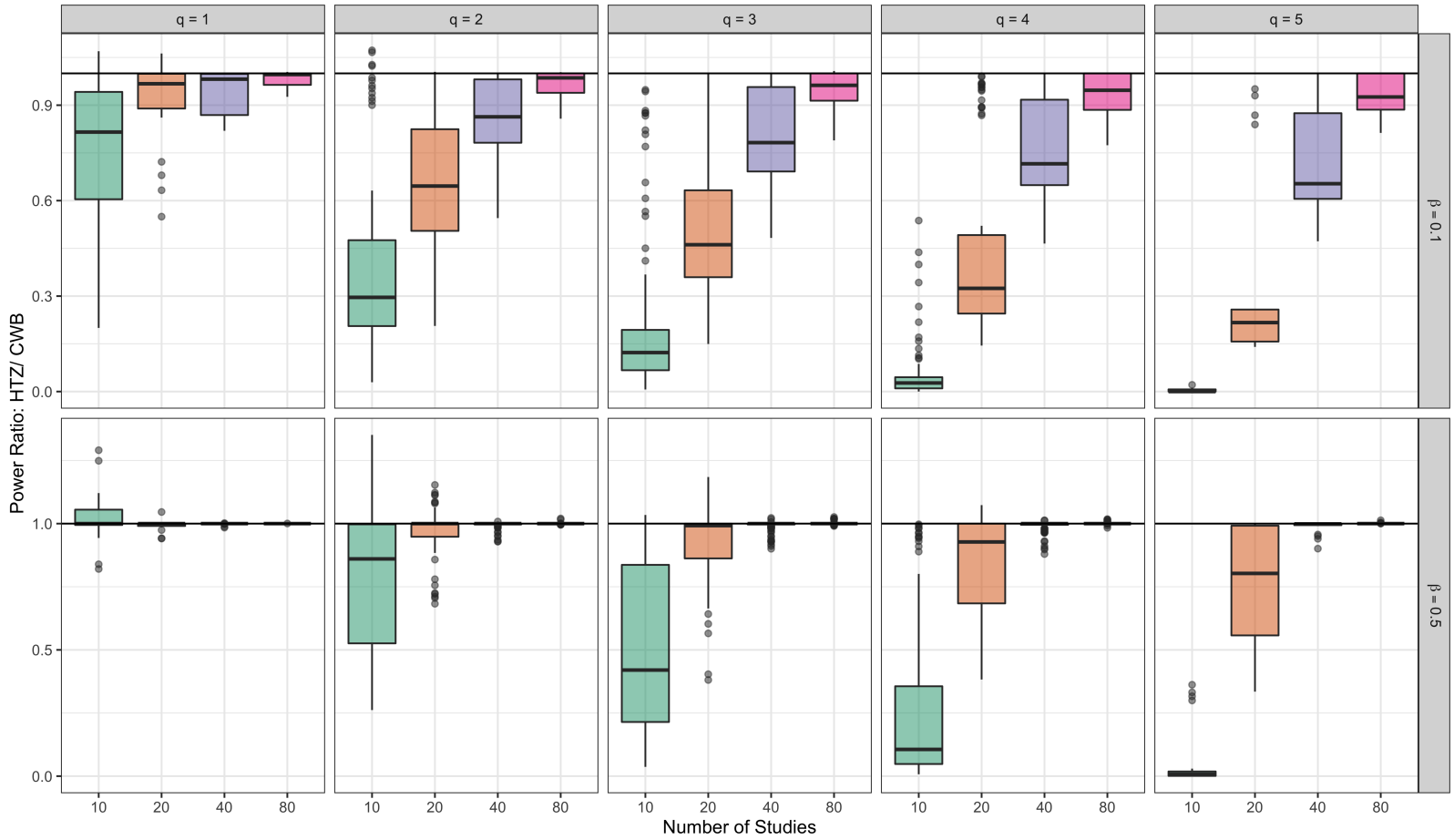

Figure 4.8. Study 1: Ratio of power of the HTZ test and the CWB test by the number of studies, the number of contrasts (q), and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

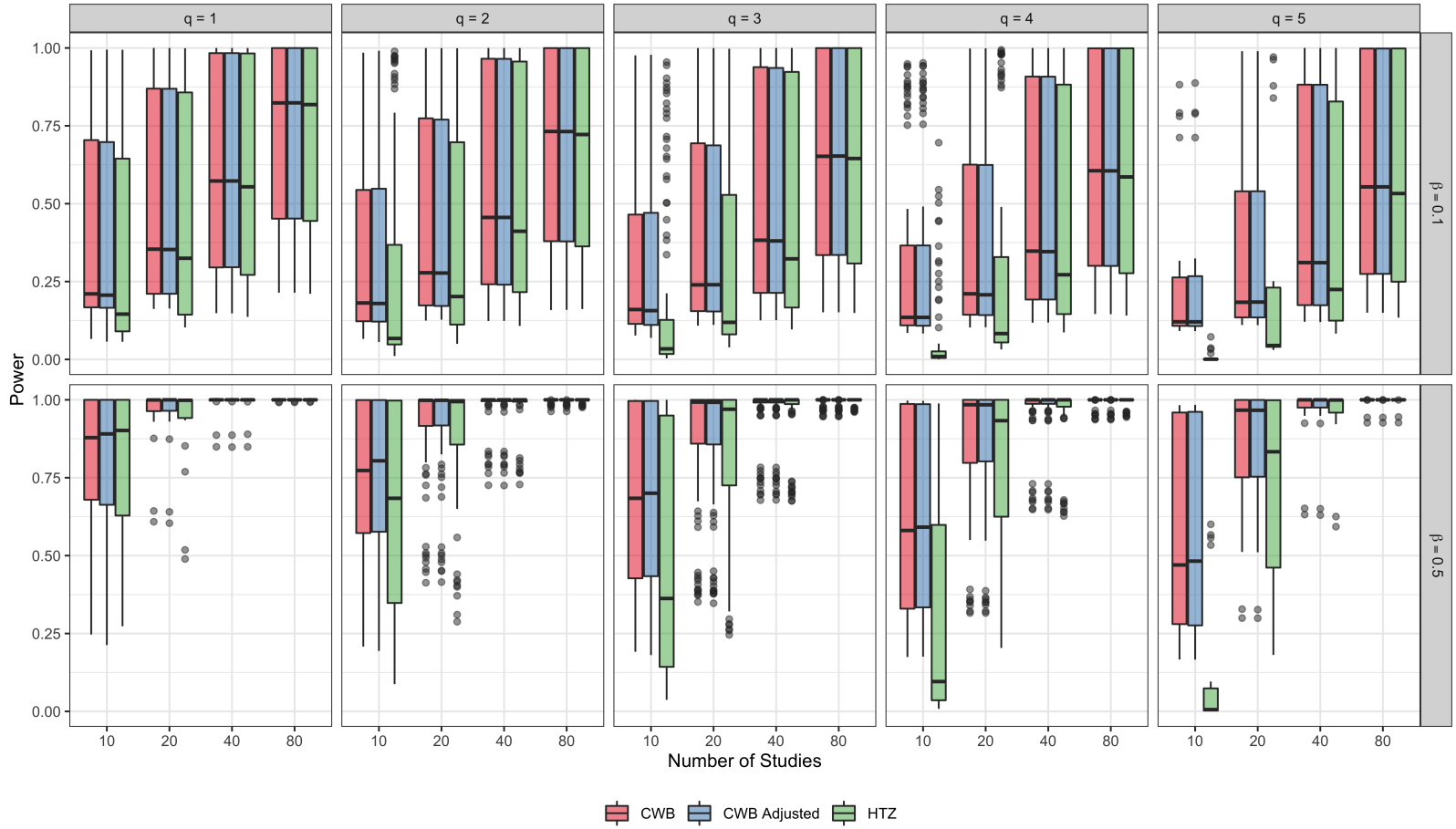

Figure 4.9. Study 1: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the number of contrasts (q), and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.10. The maximum MCSE for each of the tests across all conditions was 0.01.

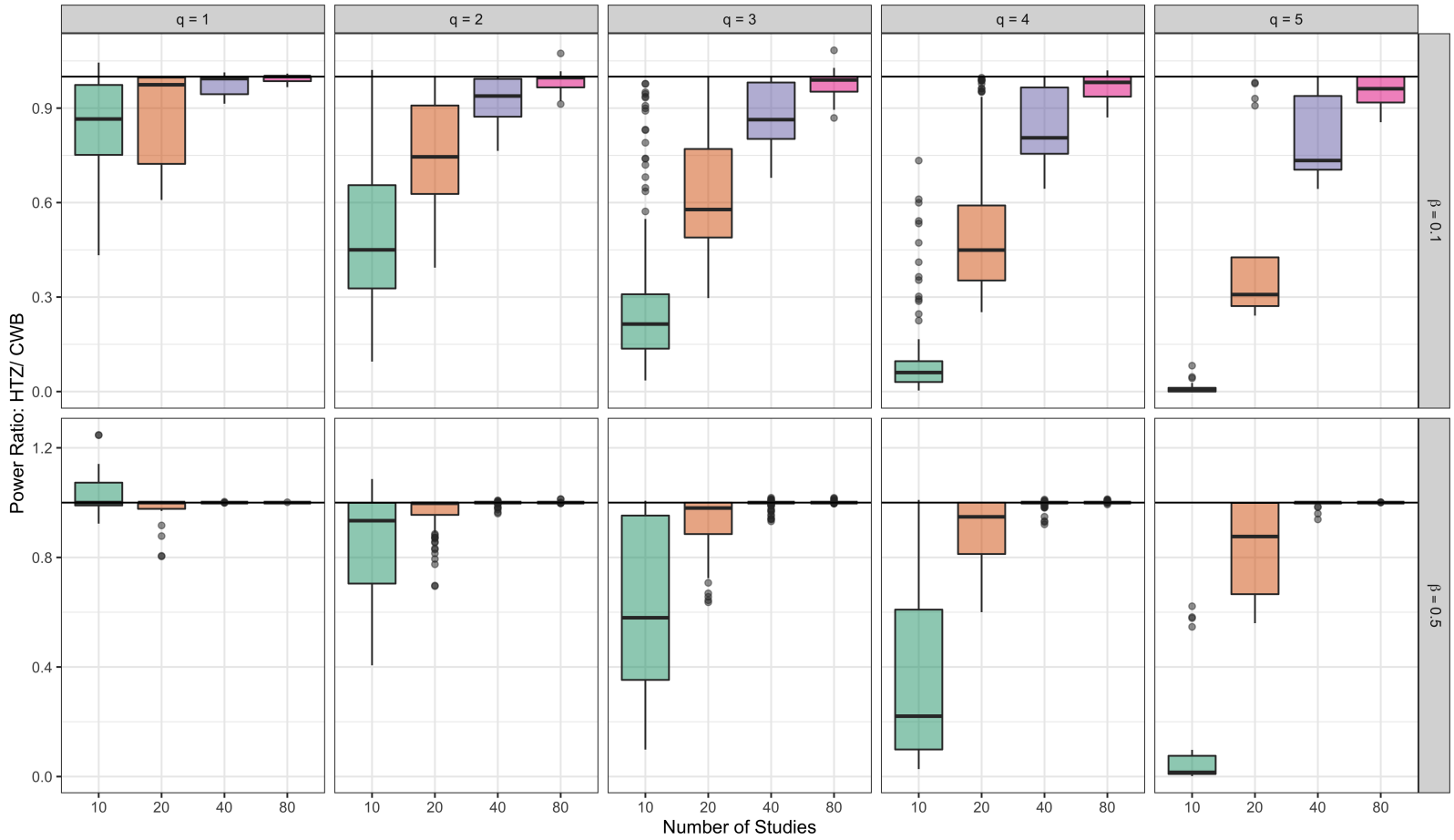

Figure 4.10. Study 1: Ratio of power of the HTZ test and the CWB test by the number of studies, the number of contrasts (q), and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.10. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

Power and Power Ratio by Sets of Covariates Tested

Figures 4.11 to 4.18 show the results of absolute power and power ratio disaggregated by sets of covariates that were tested. I present results for 1 to 4 contrasts. For tests of 5 contrasts, there is only one possible set of covariates and thus, the graphs would be similar to the ones presented in the aggregated results above. For brevity, I only present results for nominal α level of 0.05.

For single coefficient tests, X1, a study-level binary variable with large imbalance, had lowest absolute power compared to other variables. X2, an effect size-level binary variable with large imbalance, also had low absolute power. X3, a normally distributed study-level continuous covariate, had higher power than the binary moderators but lower power compared to the other continuous covariates. X4, an effect size-level normally distributed covariate, had higher power than X3. X5, an effect size-level skewed continuous covariate, had the highest power compared to all other variables. For conditions with higher number of studies and for conditions where the regression coefficient was set to 0.5, the power levels of all tests were around 1. For multiple-contrast hypotheses tests, the sets that included X1 generally had lower power compared to the other sets and the ones that included X5 generally had higher power compared to the other sets.

For single coefficient tests, in conditions with 10 studies and in conditions with the regression coefficient equaling 0.5, the box-plots for X1 and X2 are above the solid line at 1 indicating that the HTZ test had higher power than the CWB test. In most other conditions, the box-plots are below the solid line. For multiple-contrast hypothesis tests, the box-plots are almost always below the solid line at 1 indicating that the CWB test had higher power than the HTZ test. Different sets of covariates tended to have slightly different ranges of power ratios. However, the ratios varied more strongly with the number of studies and the regression coefficient used to generate the effect sizes.

4.1.3 Sensitivity to τ and ρ Values

In this section, I present analyses of the sensitivity of the results to differing values of τ and ρ used in the experimental design of Study 1. I examined varying values of ρ to examine the sensitivity of the results to mis-specification of the within-study

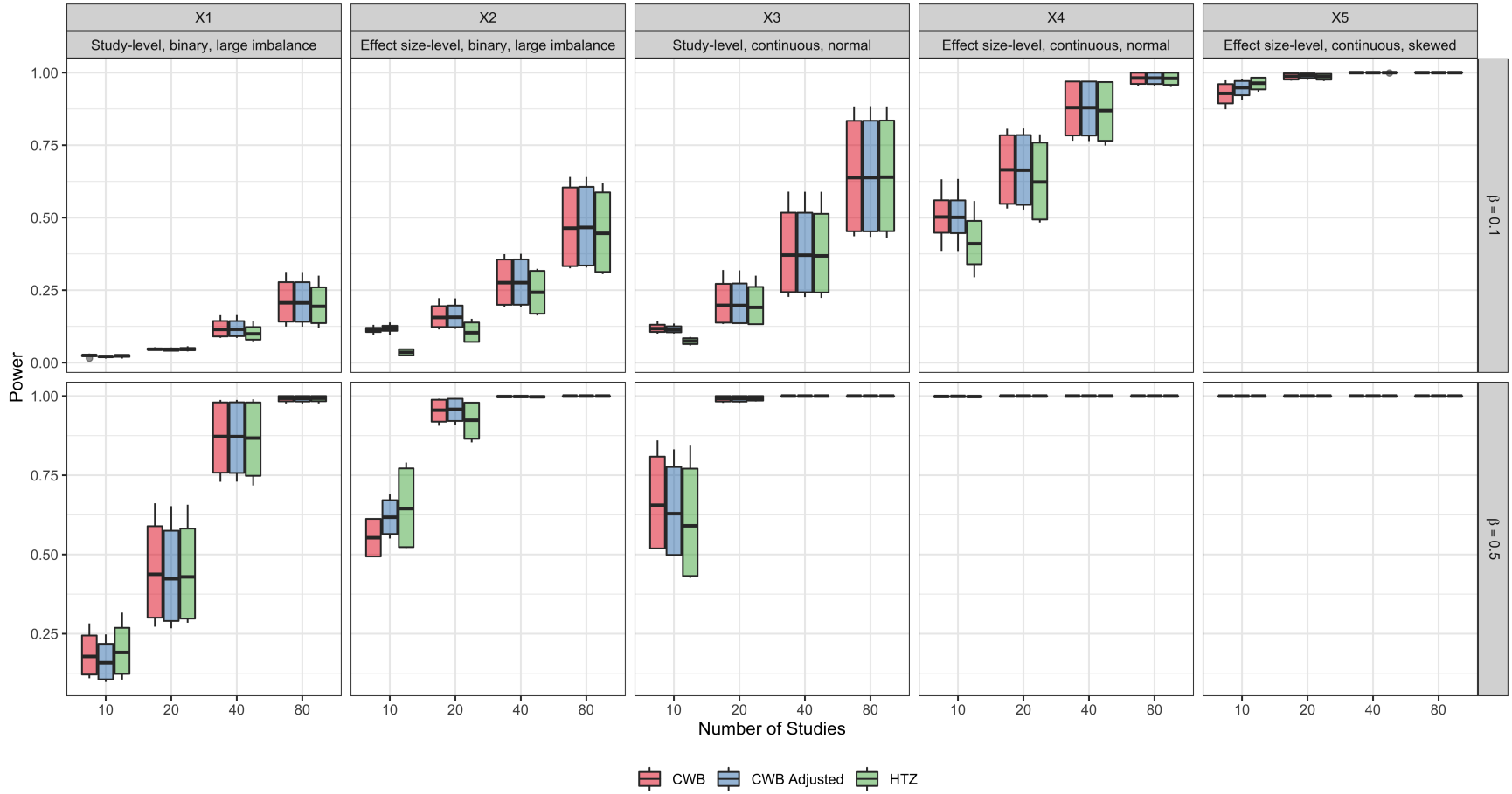

Figure 4.11. Study 1: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the covariate tested, and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The maximum MCSE for each of the tests across all conditions was 0.01.

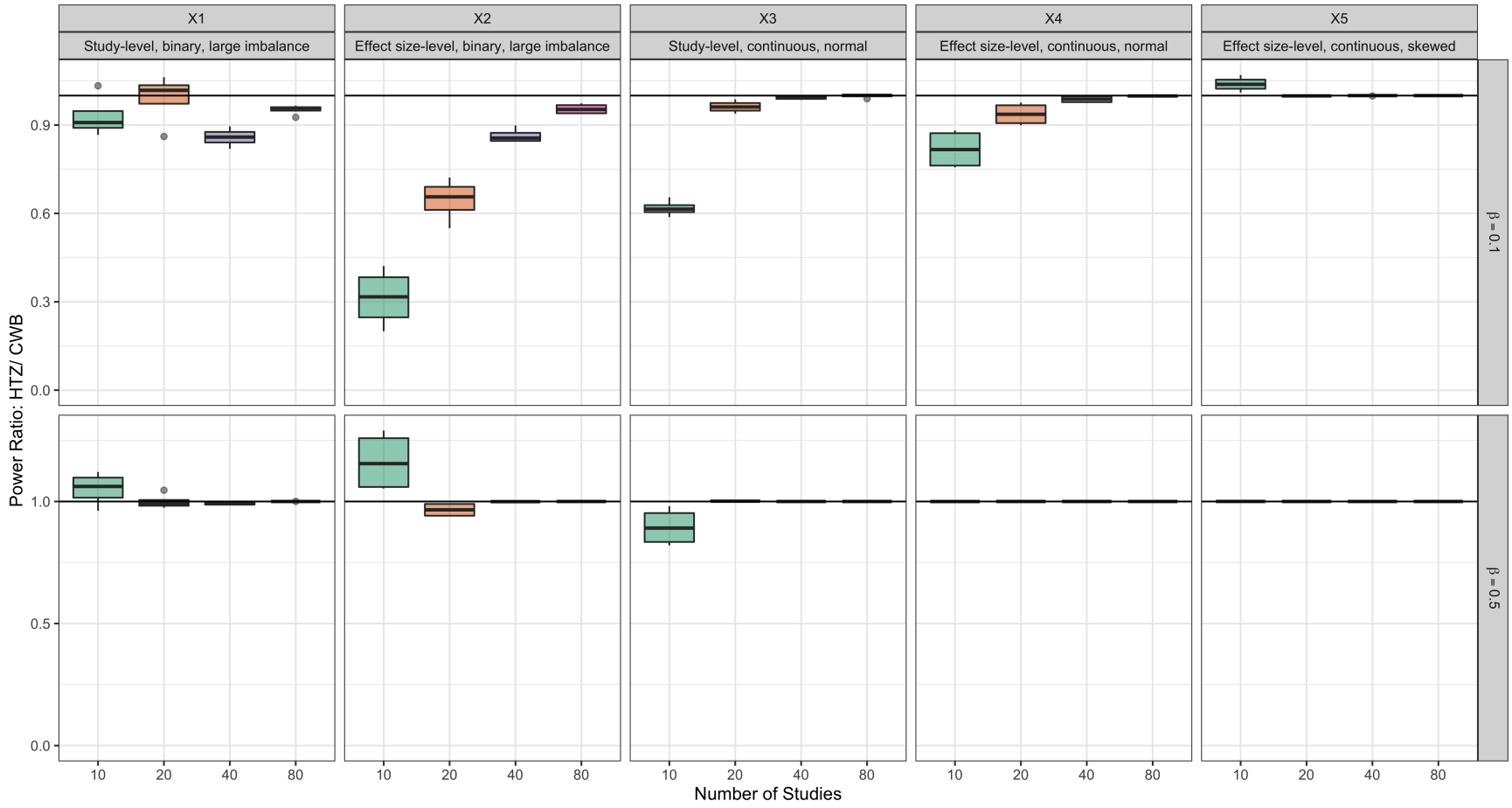

Figure 4.12. Study 1: Ratio of power of the HTZ test and the CWB test by the number of studies, the covariate tested, and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

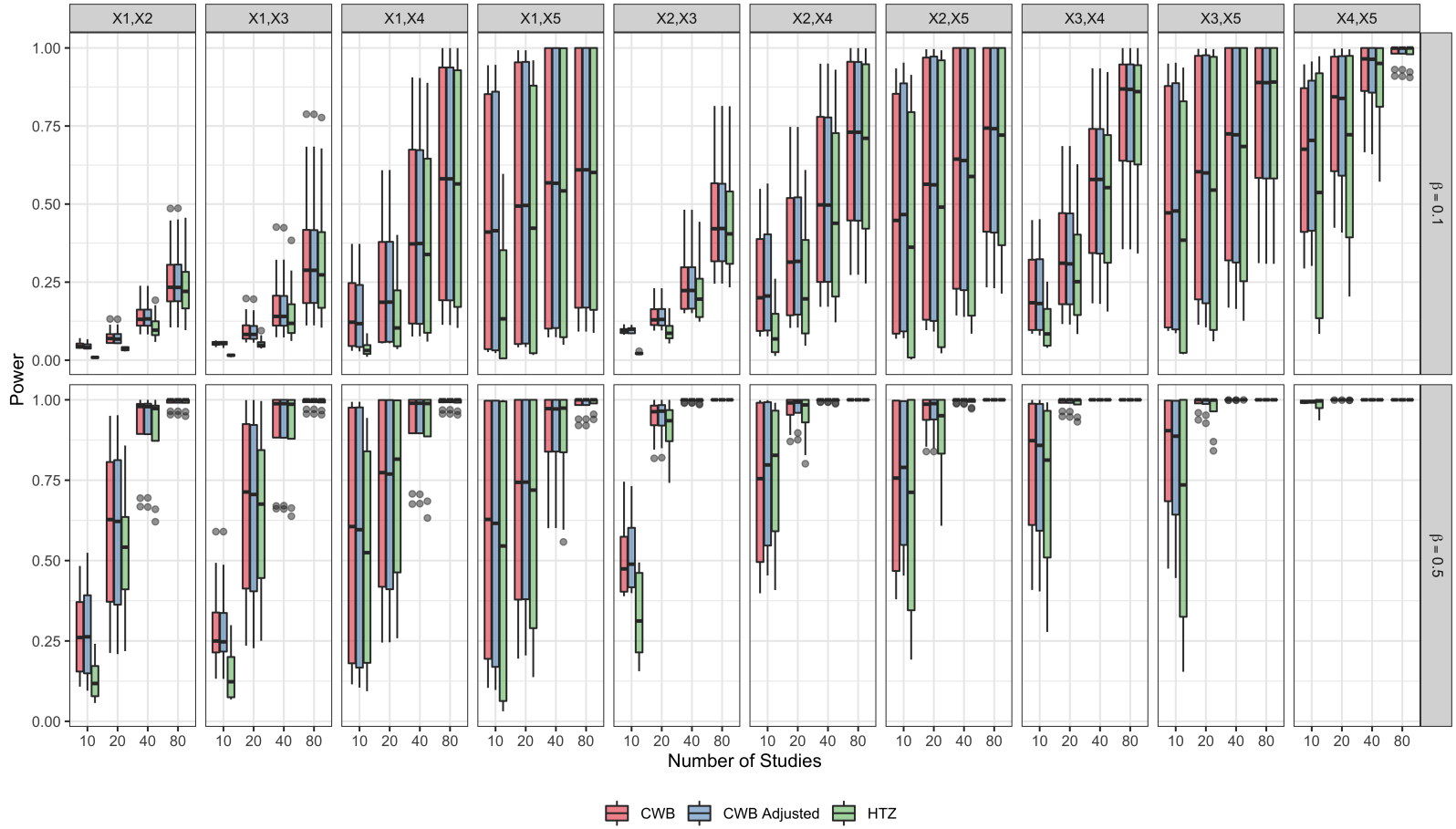

Figure 4.13. Study 1: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the covariate tested, and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The maximum MCSE for each of the tests across all conditions was 0.01.

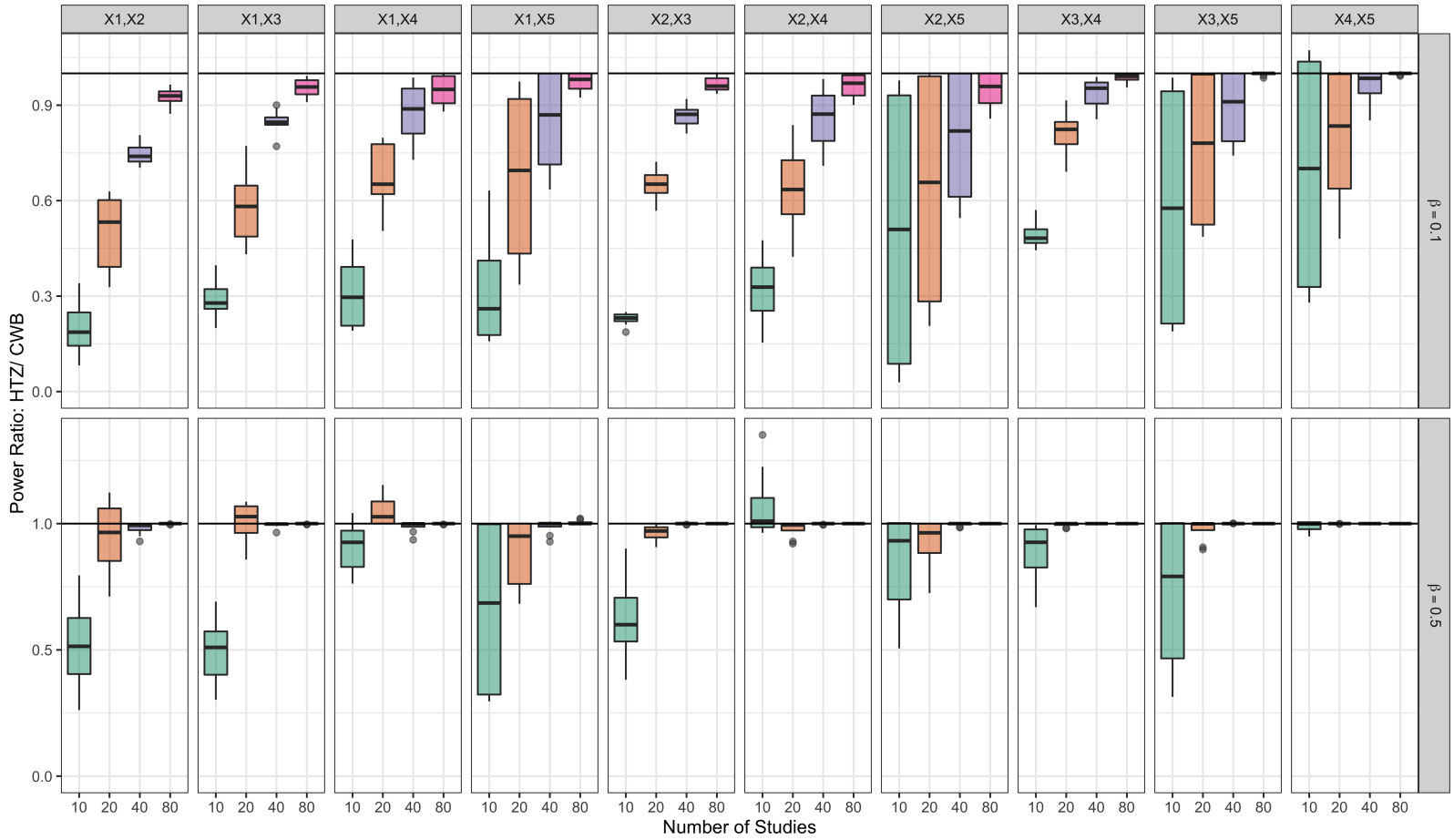

Figure 4.14. Study 1: Ratio of power of the HTZ test and the CWB test by the number of studies, the covariate tested, and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

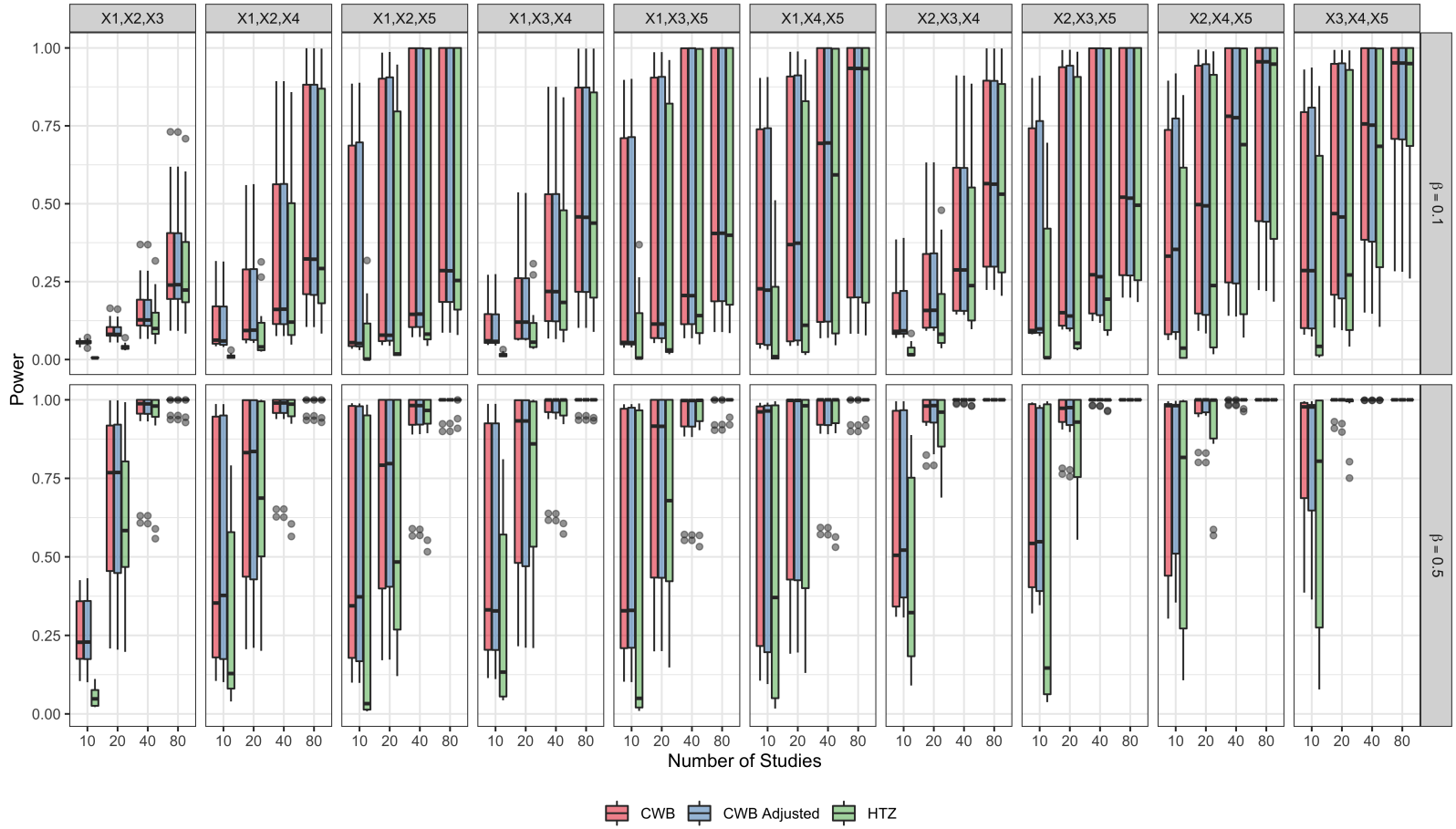

Figure 4.15. Study 1: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the covariate tested, and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The maximum MCSE for each of the tests across all conditions was 0.01.

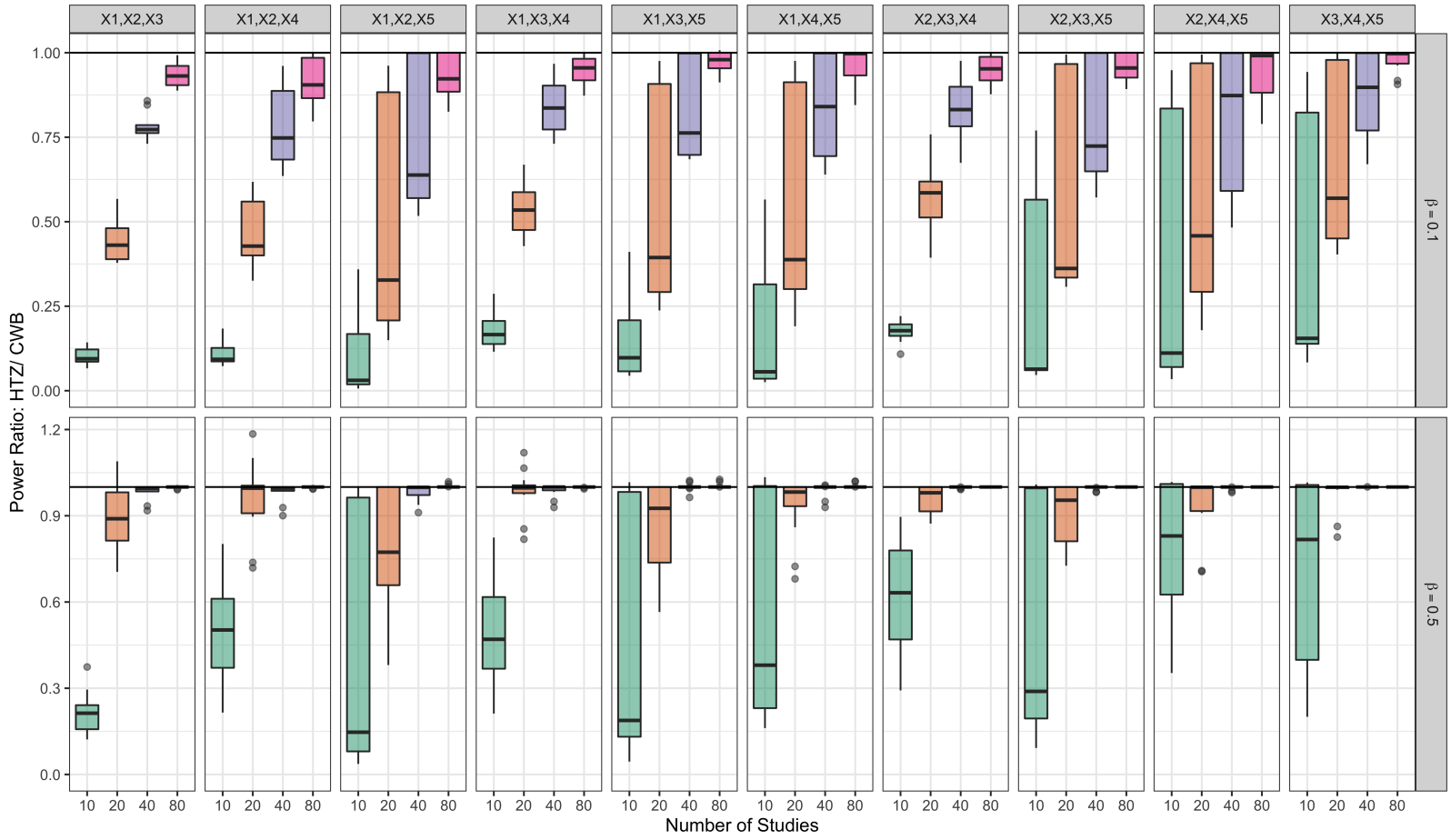

Figure 4.16. Study 1: Ratio of power of the HTZ test and the CWB test by the number of studies, the covariate tested, and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

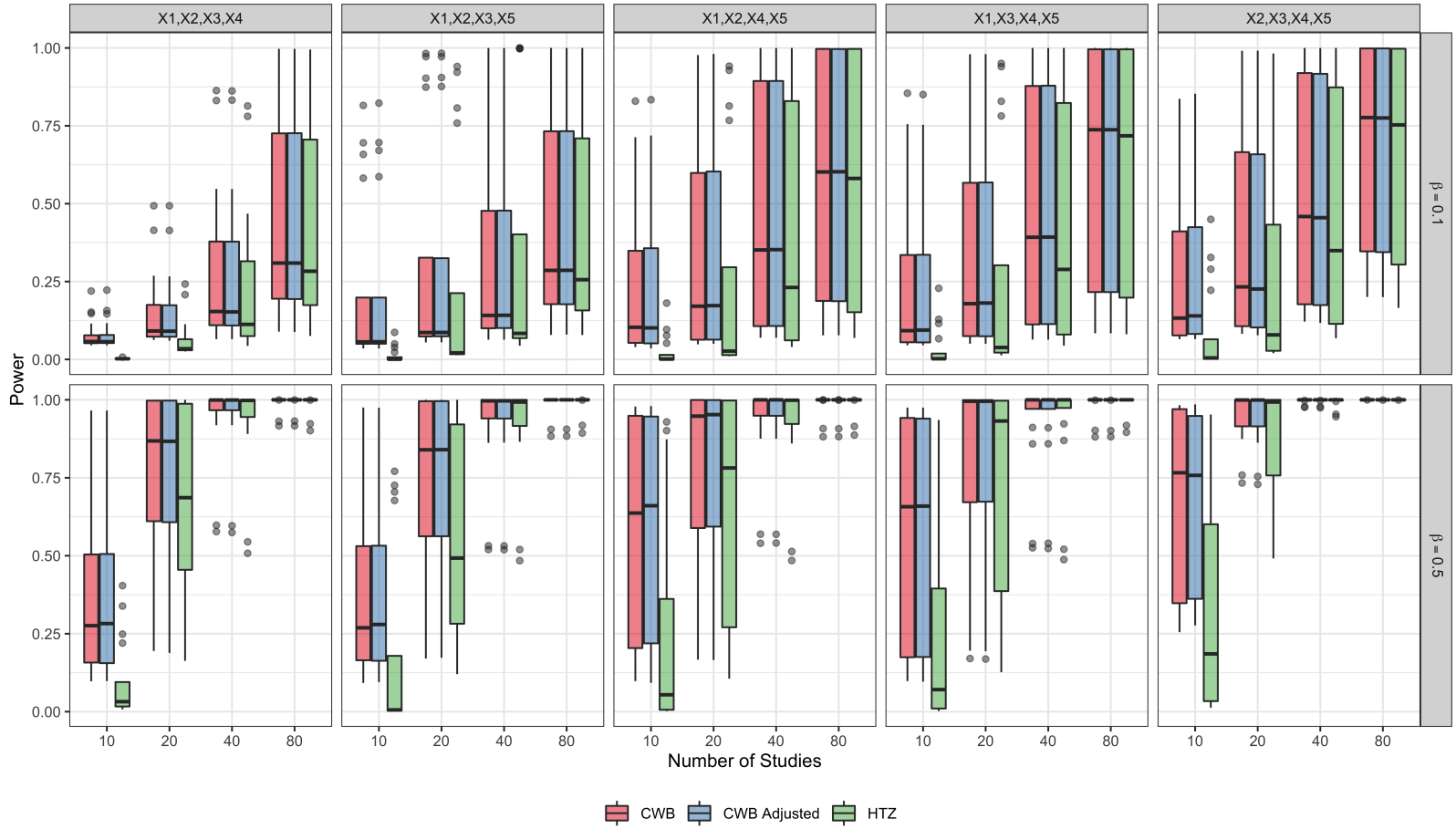

Figure 4.17. Study 1: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the covariate tested, and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The maximum MCSE for each of the tests across all conditions was 0.01.

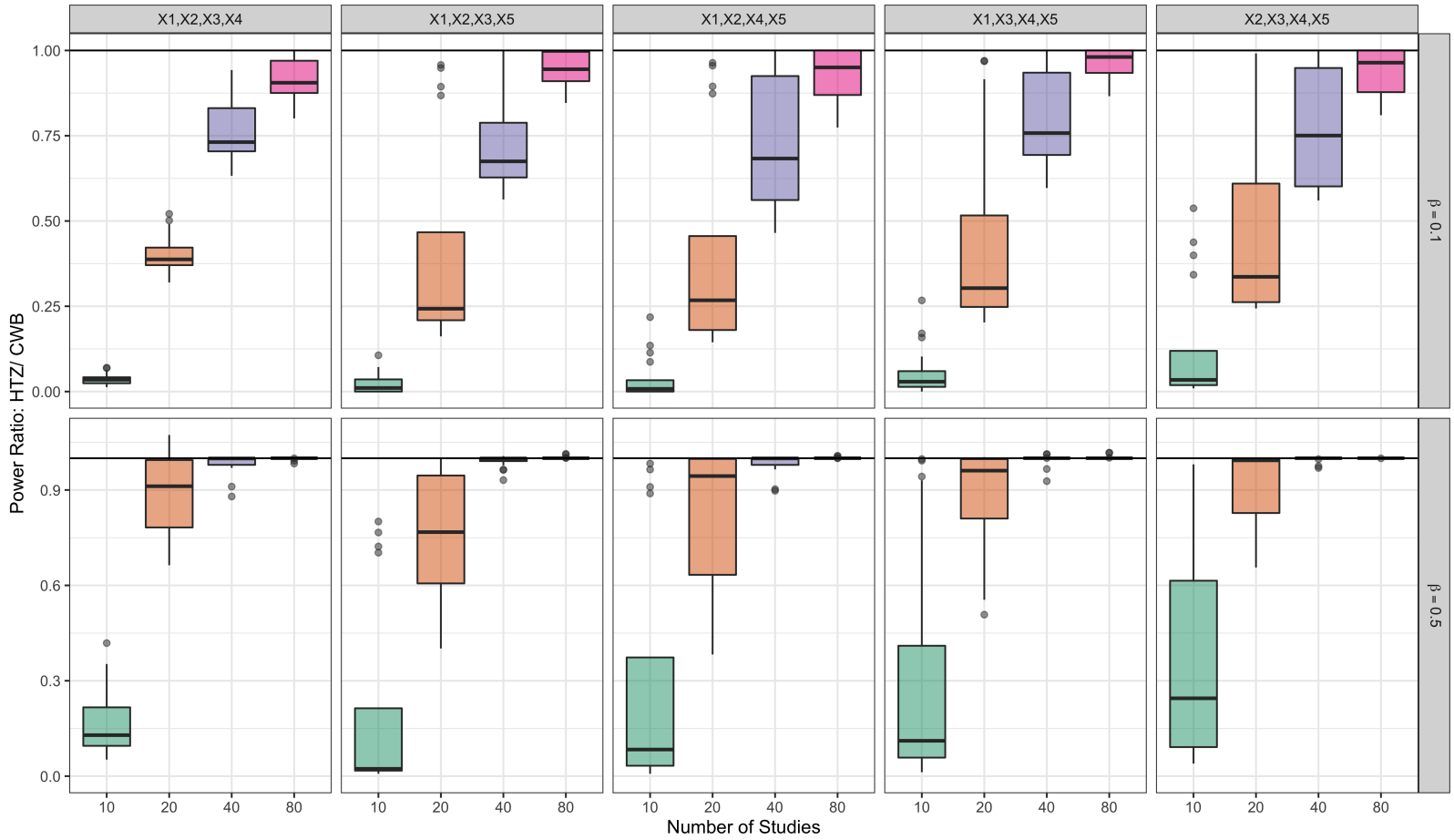

Figure 4.18. Study 1: Ratio of power of the HTZ test and the CWB test by the number of studies, the covariate tested, and the regression coefficient used to generate the true effect sizes (β) for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

correlation value in the working model. In conditions where the value of ρ is 0.5, the working model was mis-specified. I only present the results for nominal α level of 0.05. However, the pattern of results hold for nominal α levels of 0.01 and 0.10 as well.

Figures 4.19 and 4.20 show Type I error rates by the different values of τ and ρ respectively that were used to generate the data in Study 1. Across the different values of τ and ρ, Type I error rates were very similar. Figures 4.21 and 4.22 show ratio of power of the HTZ test and the CWB test by the different values of τ and ρ. Overall, the ranges of the power ratios were similar across the different values of τ and ρ. Figure 4.21 shows some discrepancies in the median power ratios between the two values of τ, especially for multiple-contrast hypotheses tests of higher number of contrasts. Generally, higher value of τ resulted in lower median power ratio, i.e., higher power advantage for the CWB test. Overall, the pattern of the results were not sensitive to the different values of τ and ρ.

4.2 Study 2

Study 2 was designed to examine whether the results from Study 1 would generalize to a study using a different design matrix. The design matrix of Study 2 contained only one categorical moderator with varying number of categories. The following are the results from Study 2.

4.2.1 Type I Error Rates

The box-plots for Type I error rates show the range of Type I error rates by the number of studies (m), the number of contrasts (q), the covariate type, and the nominal α levels. The rejection rates range over the rates for different τ and ρ values. The solid lines indicate the nominal α level and the dashed lines indicate the bounds for simulation error.

Naive F-test

Figure 4.23 shows the Type I error rates of the Naive F-test. The Naive-F test resulted in high Type I error rates across most conditions. The Type I error rates for the effect size-level covariate type were lower than those for the study-level covariate

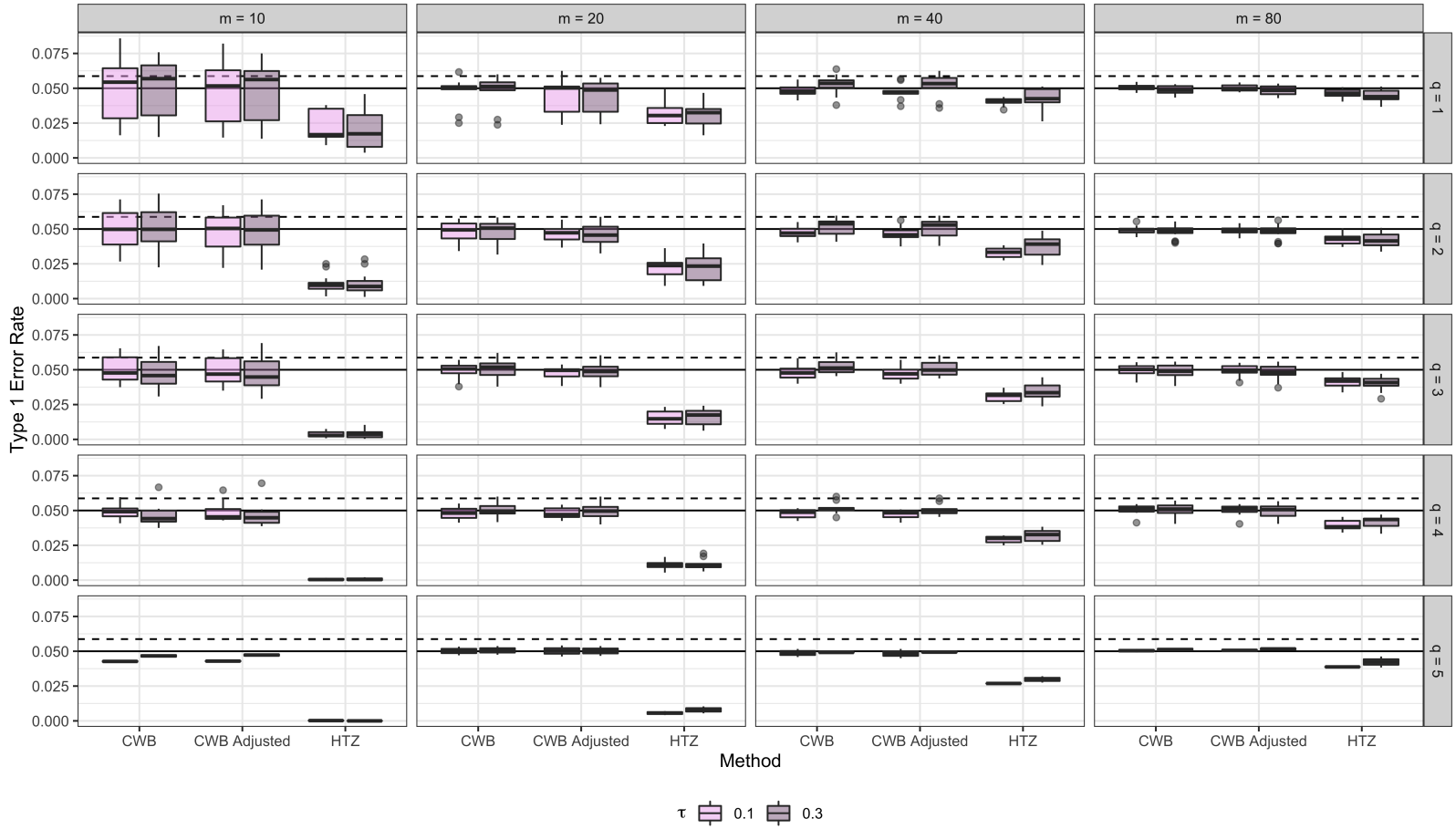

Figure 4.19. Study 1: Sensitivity of Type I error rates results to τ values: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m), the number of contrasts (q), and τ values for nominal α level of 0.05. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error.

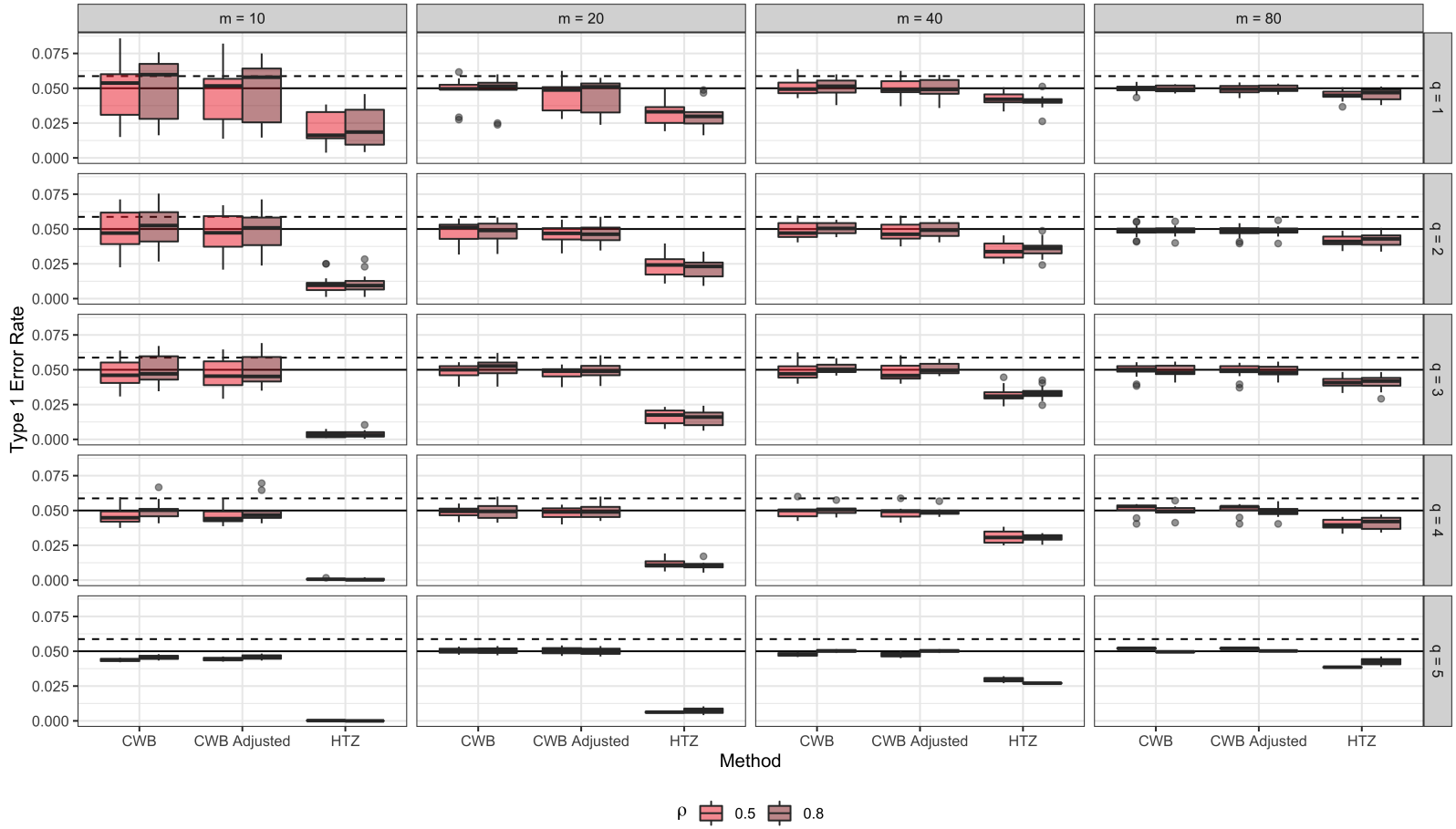

Figure 4.20. Study 1: Sensitivity of Type I error rates results to ρ values: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m), the number of contrasts (q), and ρ values for nominal α level of 0.05. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error.

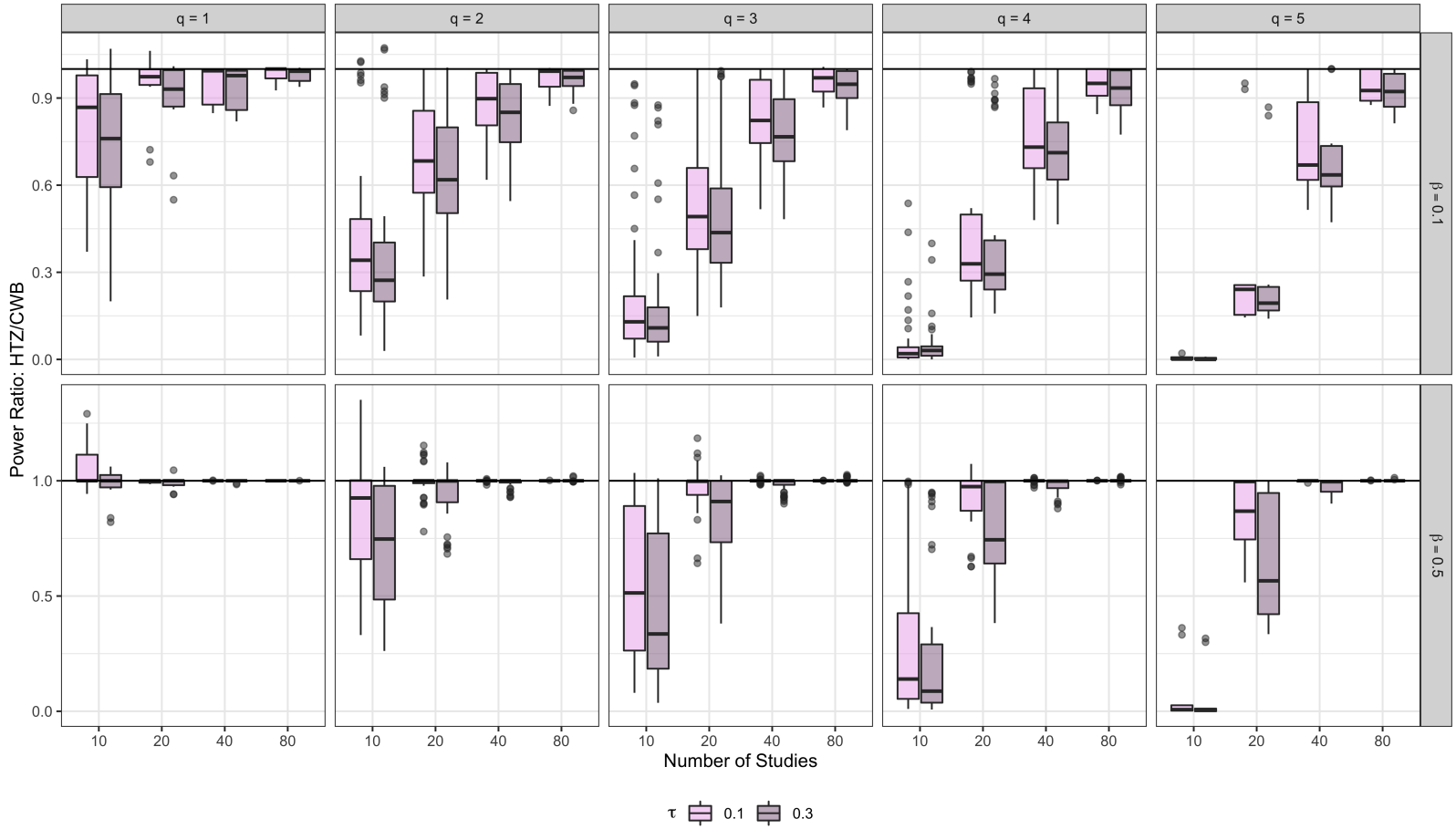

Figure 4.21. Study 1: Sensitivity of power results to τ values: Ratio of power of the HTZ test and the CWB test by the number of studies, the number of contrasts (q), the regression coefficient used to generate the true effect sizes (β), and τ values for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

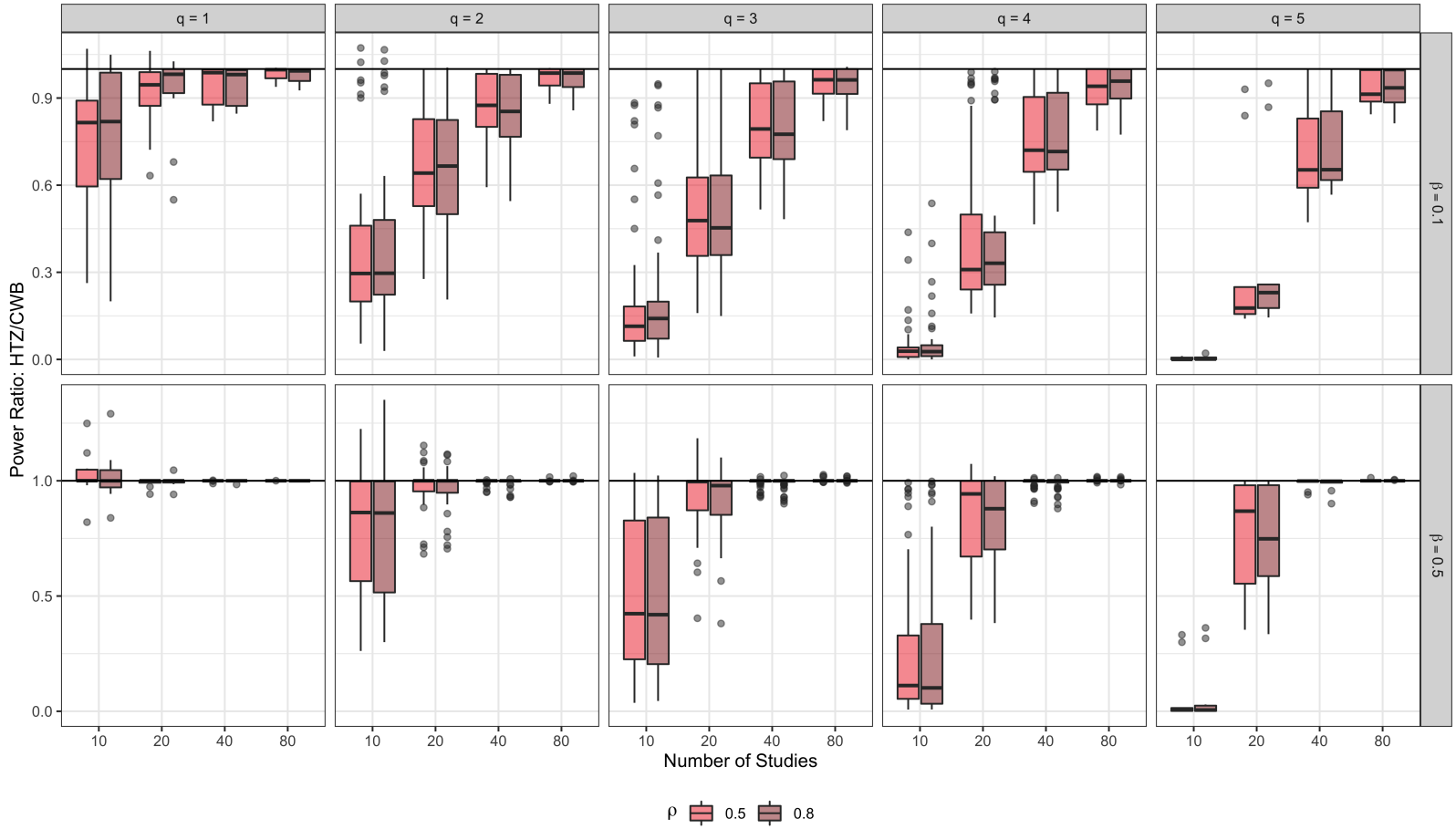

Figure 4.22. Study 1: Sensitivity of power results to ρ values: Ratio of power of the HTZ test and the CWB test by the number of studies, the number of contrasts (q), the regression coefficient used to generate the true effect sizes (β), and ρ values for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

type. For conditions with 2 to 3 contrasts, the Type I error rates were near the nominal α level, especially for conditions with higher number of studies. In the plots, note that the range of the Type I error rates in the y-axes are different for the different covariate types and the nominal α levels.

Cluster Wild Bootstrapping versus HTZ

Figures 4.24, 4.25, and 4.26 show the range of Type I error rates of the CWB, the CWB Adjusted, and the HTZ test for the nominal α levels of 0.01, 0.05, and 0.10 respectively. I did not include the results of the Naive F-test in this set of graphs as the Naive F-test exhibited higher than nominal Type I error rates across most conditions.

Overall, the pattern of results was similar to that of Study 1. The HTZ test had Type I error rates below the nominal rate especially for conditions with lower numbers of studies and for tests of higher number of contrasts. The CWB and the CWB Adjusted tests had Type I error rates closer to the nominal α level. The rates exceeded the nominal level slightly in some conditions but were still within the Monte Carlo simulation error bound across most conditions. For conditions with 10 studies, the Type I error rates of the three tests for the effect size-level covariate type were higher compared to those for the study-level covariate type. However, in conditions with 40 and 80 studies, the Type I error rates of the three tests for the effect size-level covariate type were lower compared to those for the study-level covariate type. The differences between the Type I error rates of the two covariate types in conditions with higher number of studies were much more pronounced for the HTZ test than for the CWB and the CWB Adjusted tests.

4.2.2 Power

The box-plots for absolute power show the range of power of the tests by the number of studies, the number of contrasts (q), the regression coefficient used to generate the true effect sizes (β), the covariate type, and the nominal α levels. The rejection rates range over the rates for different τ and ρ values. The box-plots for power ratio show the range of the power ratio instead of absolute power.

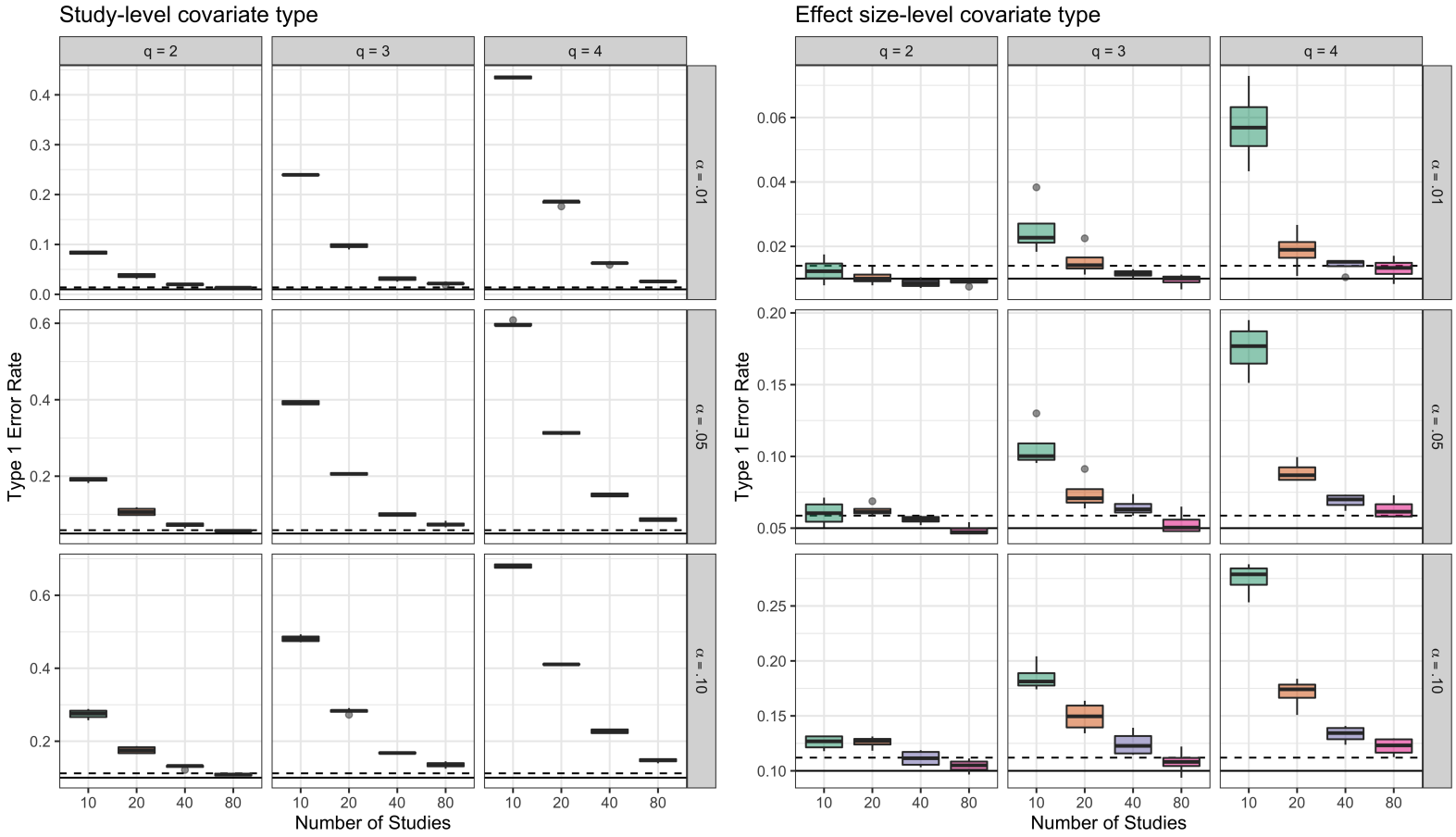

Figure 4.23. Study 2: Type I error rates of the Naive F-test by the number of studies, the number of contrasts (q), the nominal α level, and the covariate type. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error. The maximum MCSE across all conditions and for each of the nominal α levels was 0.01.

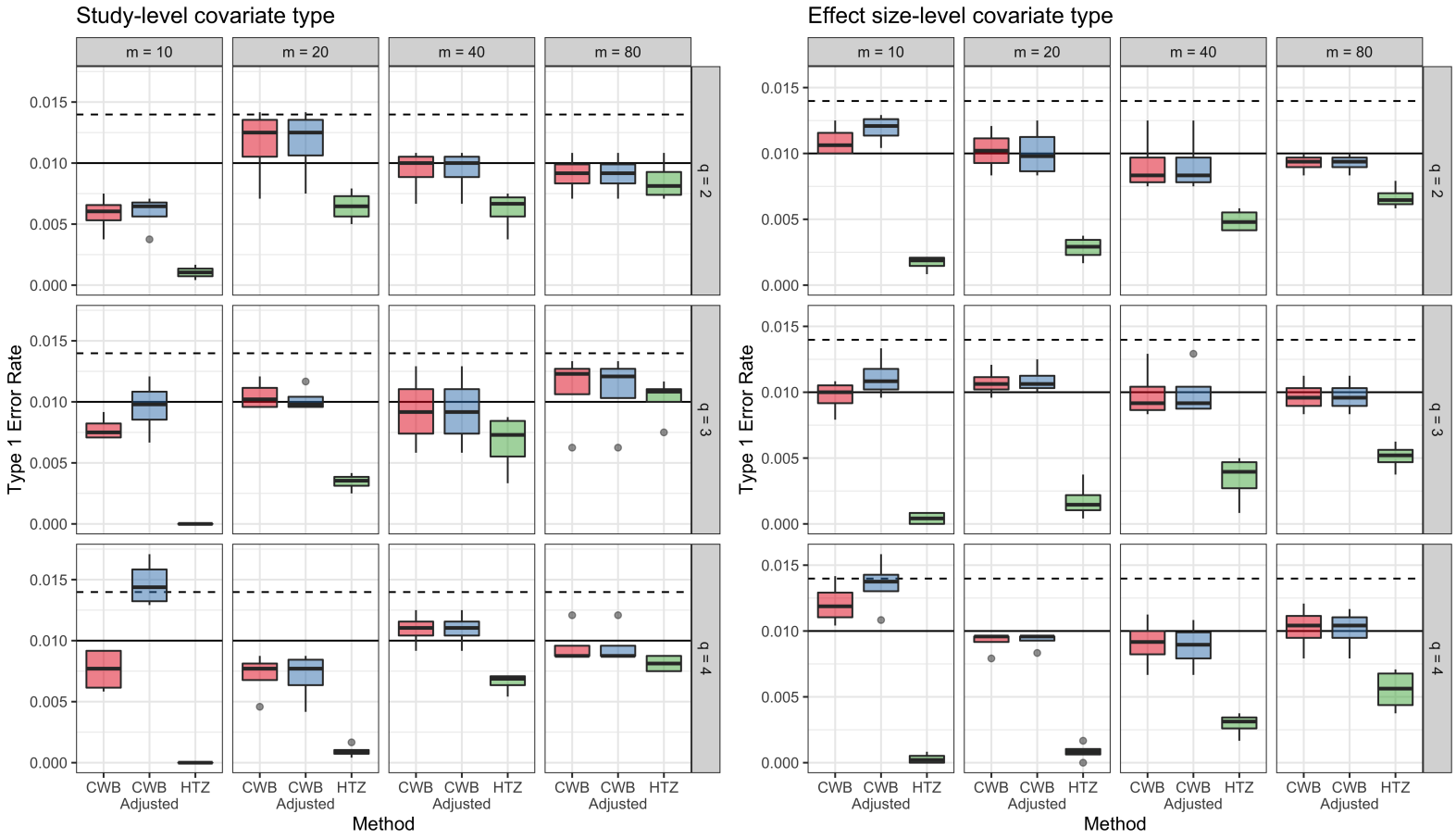

Figure 4.24. Study 2: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m), the number of contrasts (q), and the covariate type for nominal α level of 0.01. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error. The maximum MCSE for the CWB test and the HTZ test across all conditions was 0.002, and the maximum for the CWB Adjusted test was 0.003.

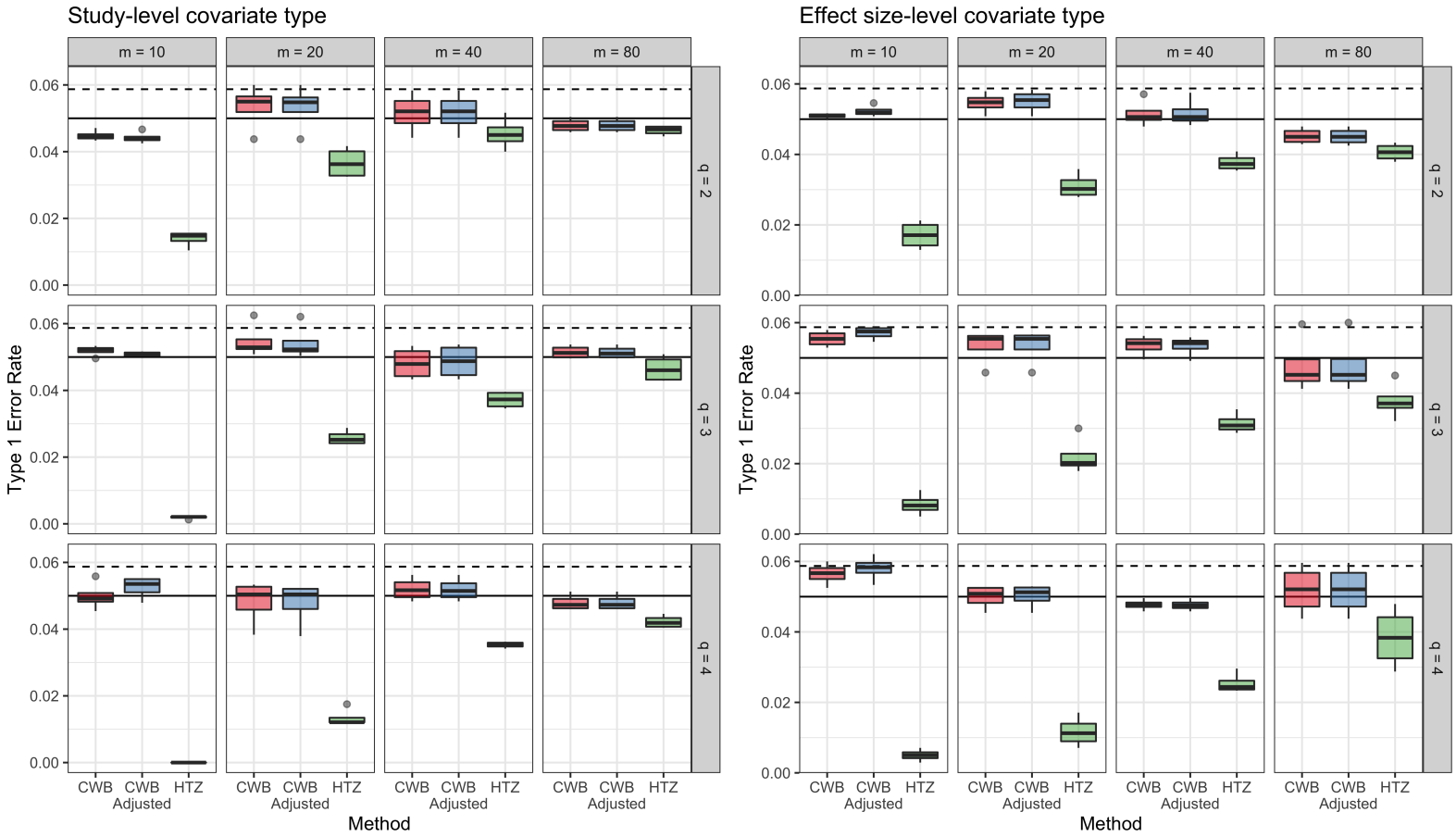

Figure 4.25. Study 2: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m), the number of contrasts (q), and the covariate type for nominal α level of 0.05. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error. The maximum MCSE for the CWB test, the CWB Adjusted test, and the HTZ test across all conditions was 0.005.

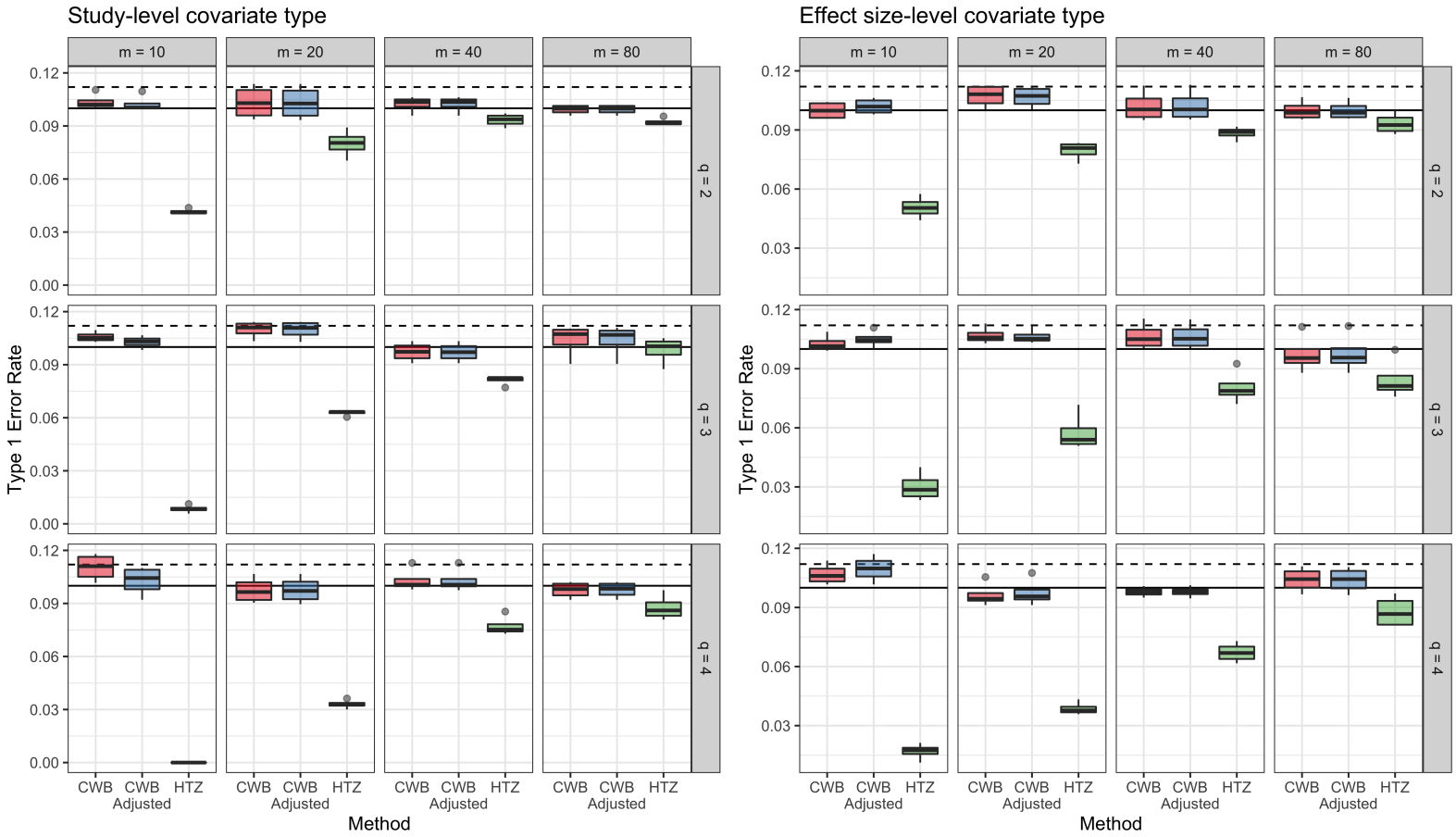

Figure 4.26. Study 2: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m), the number of contrasts (q), and the covariate type for nominal α level of 0.10. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error. The maximum MCSE for the CWB test and the CWB Adjusted test across all conditions was 0.007, and the maximum for the HTZ test was 0.006.

Absolute Power

As in the results for Study 1, I did not consider the Naive F-test for power analyses as the Type I error rates of the Naive F-test were generally too high. Figures 4.27, 4.29, and 4.31 show absolute power of the CWB, the CWB Adjusted, and the HTZ tests for the nominal α levels of 0.01, 0.05, and 0.10 respectively.

The power of the HTZ test was lower than those of the CWB and the CWB Adjusted tests. The power of the HTZ test was particularly low for conditions with lower number of studies and for tests of higher number of contrasts. There was no difference in power of the CWB and the CWB Adjusted tests across all conditions.

Power Ratio

Figures 4.28, 4.30, and 4.32 show the ratio of power of the HTZ test over the power of the CWB test for the nominal α levels of 0.01, 0.05, and 0.10 respectively. As in the results of Study 1, I only examined the CWB test as the performance of the CWB and the CWB Adjusted tests were nearly identical in terms of power. In the plots, ratios below the solid lines at 1 indicate the loss of power from using the HTZ test rather than the CWB test.

Overall, the results from Study 2 replicated the results from Study 1. The CWB test had power advantage in nearly all conditions. The power loss was greater for the study-level covariate type compared to the effect size-level covariate type across most conditions.

4.2.3 Sensitivity to τ and ρ Values

In this section, I present analyses of the sensitivity of the results to differing values of τ and ρ. As in the results for Study 1, I only present the results for nominal α level of 0.05. However, the pattern of results hold for nominal α levels of 0.01 and 0.10.

Figures 4.33 and 4.34 show Type I error rates by the different values of τ and ρ. Type I error rates differed slightly for different values of τ and ρ. Figures 4.35 and 4.36 show the power ratios between the HTZ and the CWB tests by the different values of τ and ρ. The power ratios differed slightly for different values for τ. However, the ratios were similar across different values of ρ for both covariate types. The overall pattern of results were not sensitive to the different values of τ and ρ.

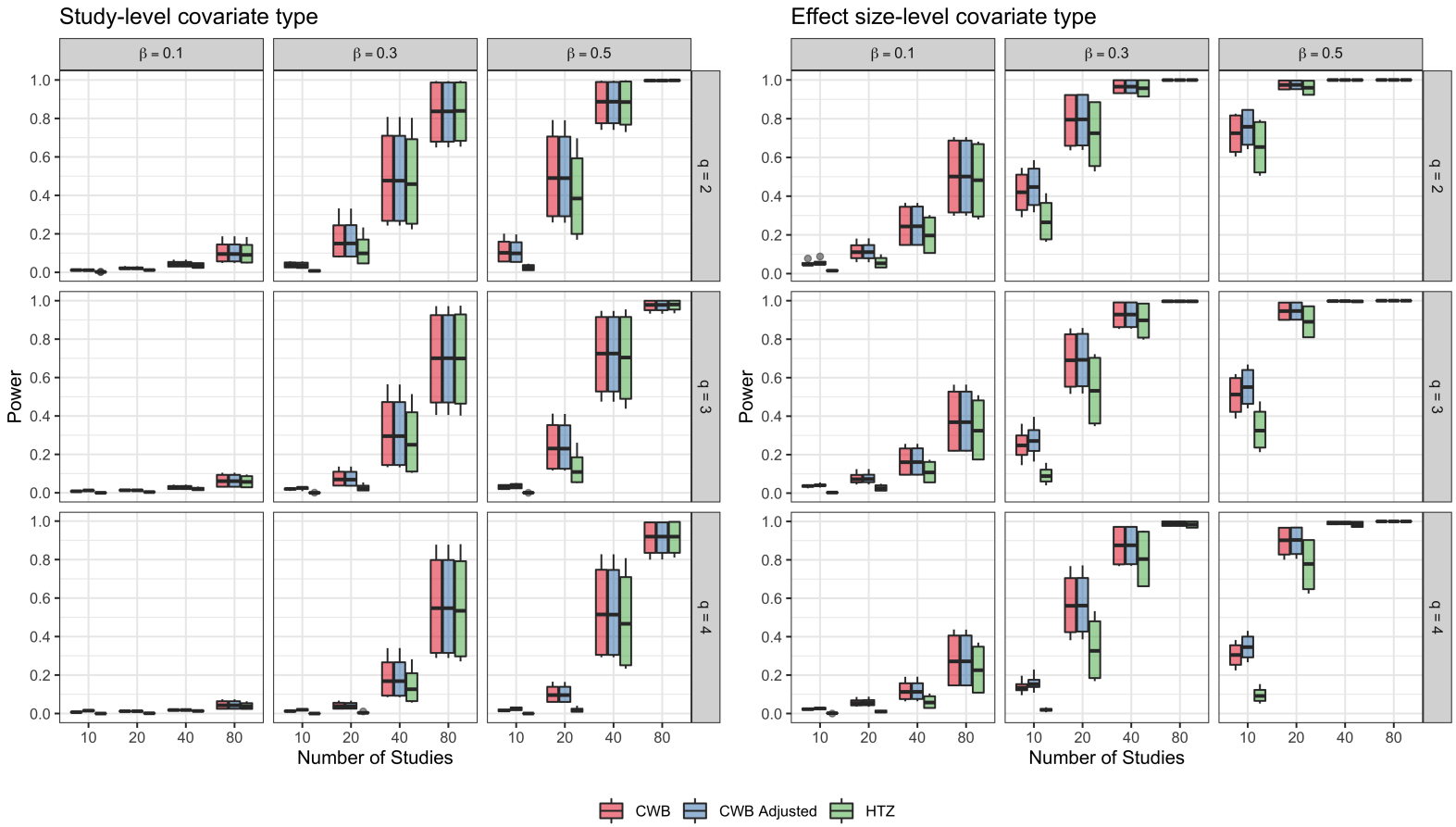

Figure 4.27. Study 2: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the regression coefficient used to generate the true effect sizes (β), the number of contrasts (q), and the covariate type for nominal α level of 0.01. The maximum MCSE for each of the tests across all conditions was 0.01.

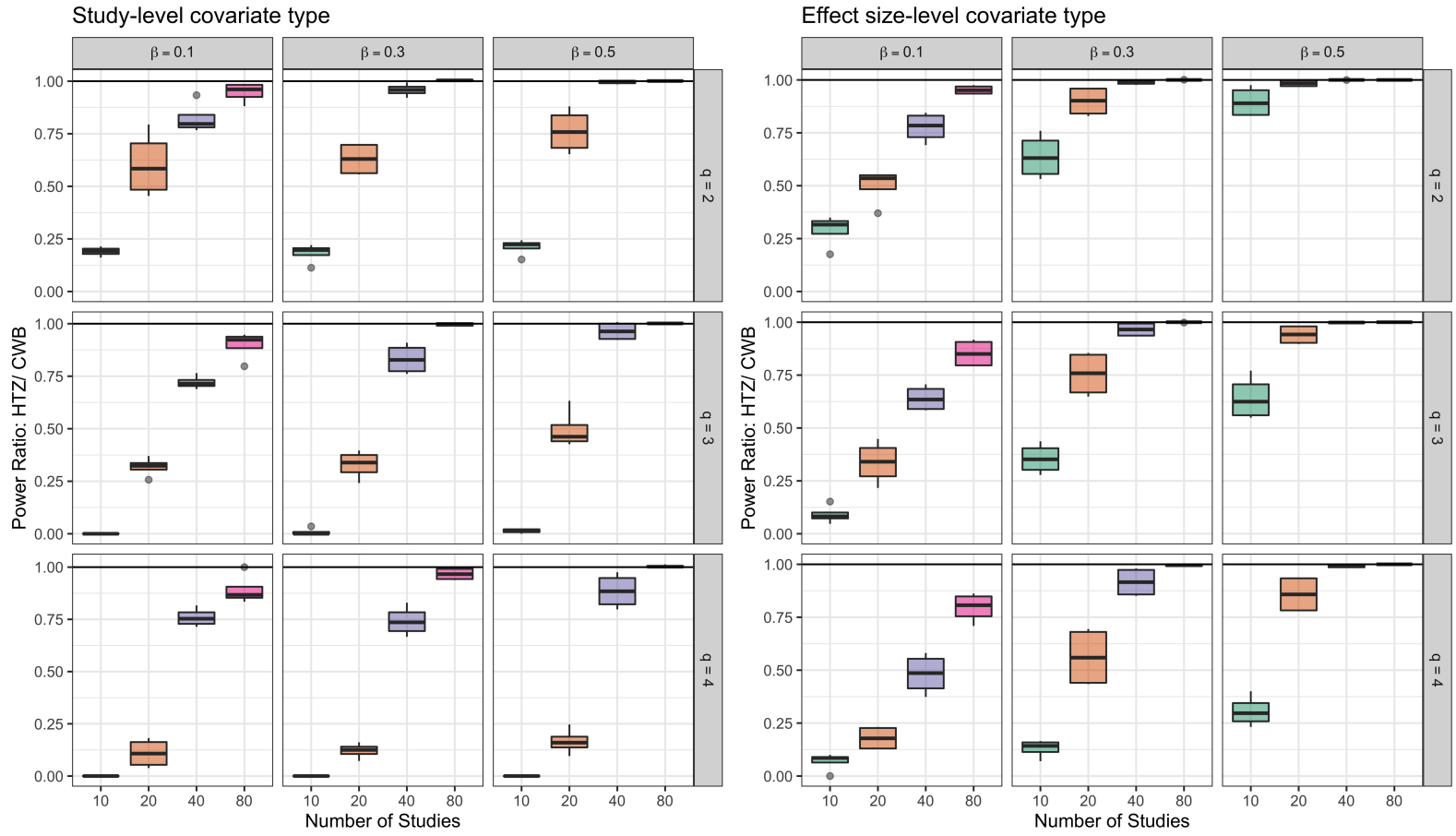

Figure 4.28. Study 2: Ratio of power of the HTZ test and the CWB test by the number of studies, the regression coefficient used to generate the true effect sizes (β), the number of contrasts (q), and the covariate type for nominal α level of 0.01. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

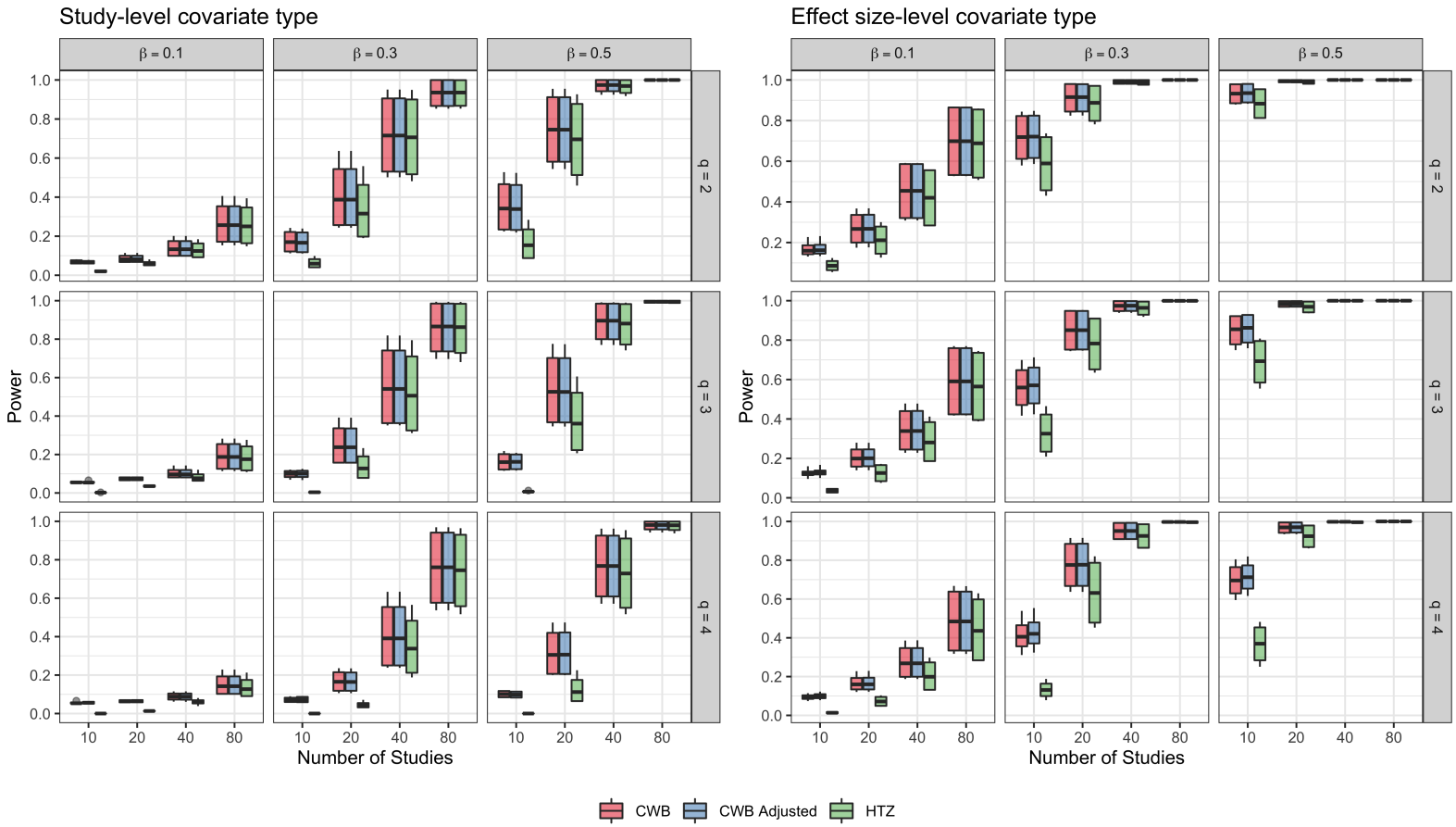

Figure 4.29. Study 2: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the regression coefficient used to generate the true effect sizes (β), the number of contrasts (q), and the covariate type for nominal α level of 0.05. The maximum MCSE for each of the tests across all conditions was 0.01.

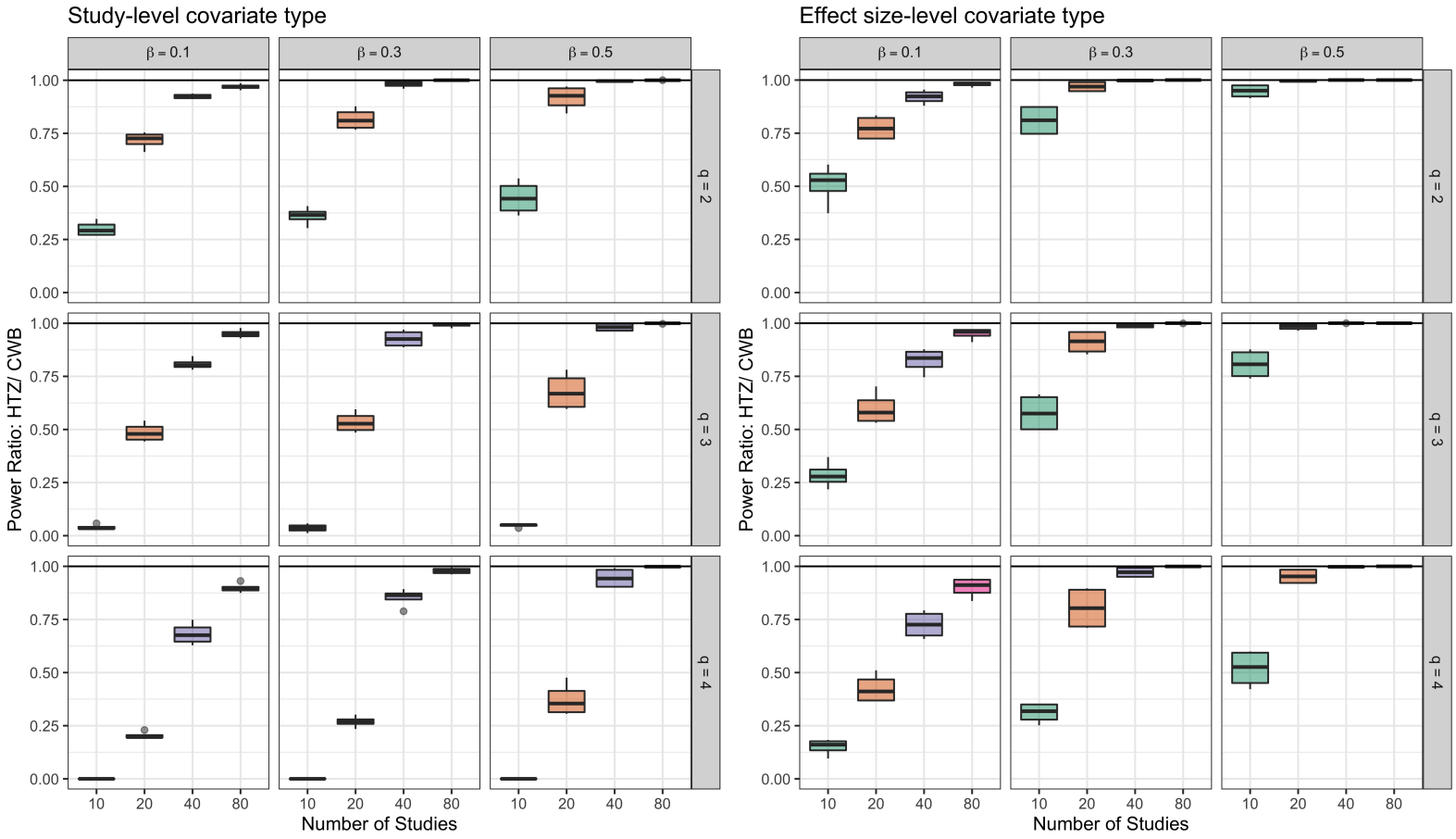

Figure 4.30. Study 2: Ratio of power of the HTZ test and the CWB test by the number of studies, the regression coefficient used to generate the true effect sizes (β), the number of contrasts (q), and the covariate type for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

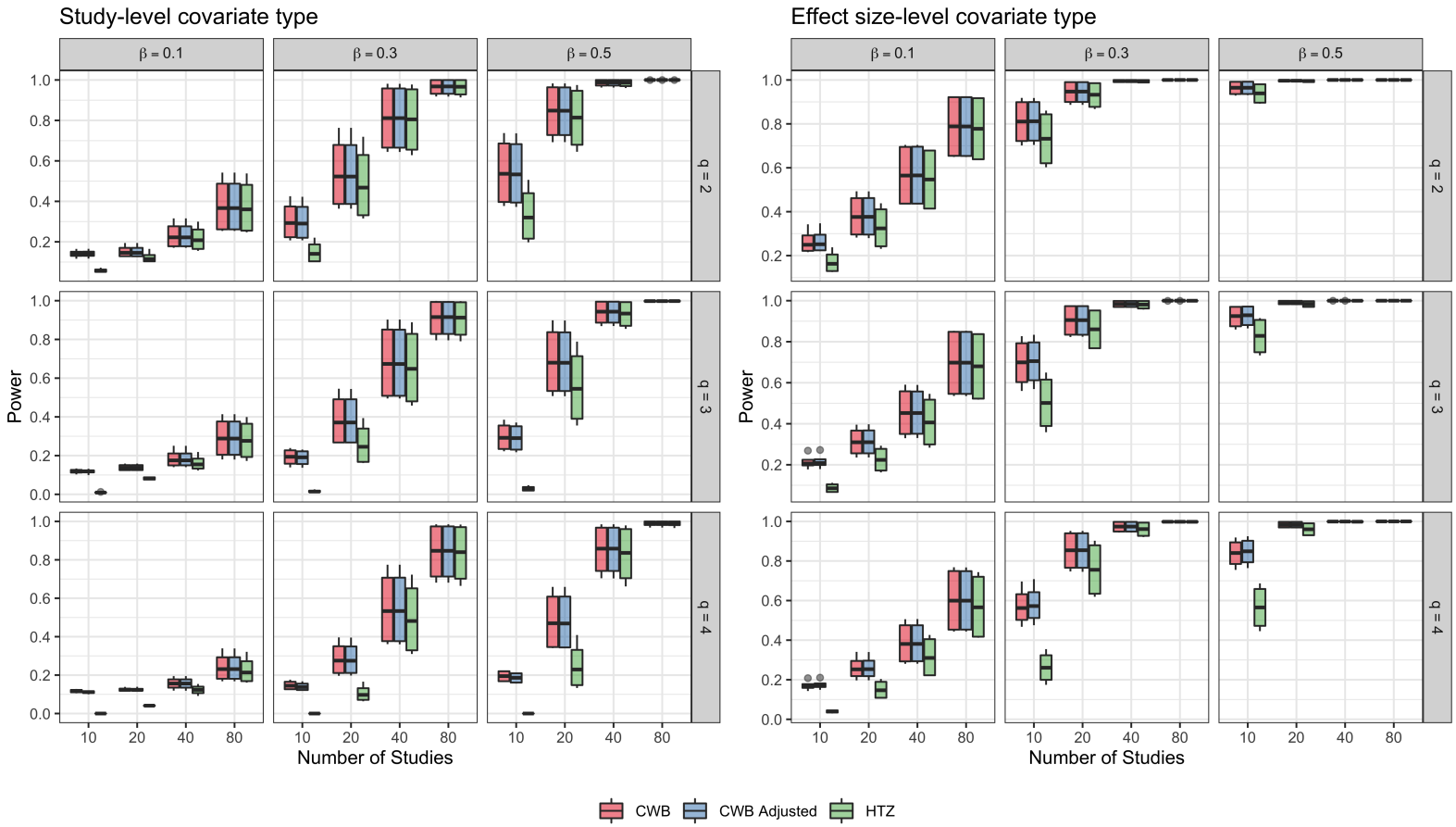

Figure 4.31. Study 2: Power of the CWB, CWB Adjusted, and HTZ tests by the number of studies, the regression coefficient used to generate the true effect sizes (β), the number of contrasts (q), and the covariate type for nominal α level of 0.10. The maximum MCSE for each of the tests across all conditions was 0.01.

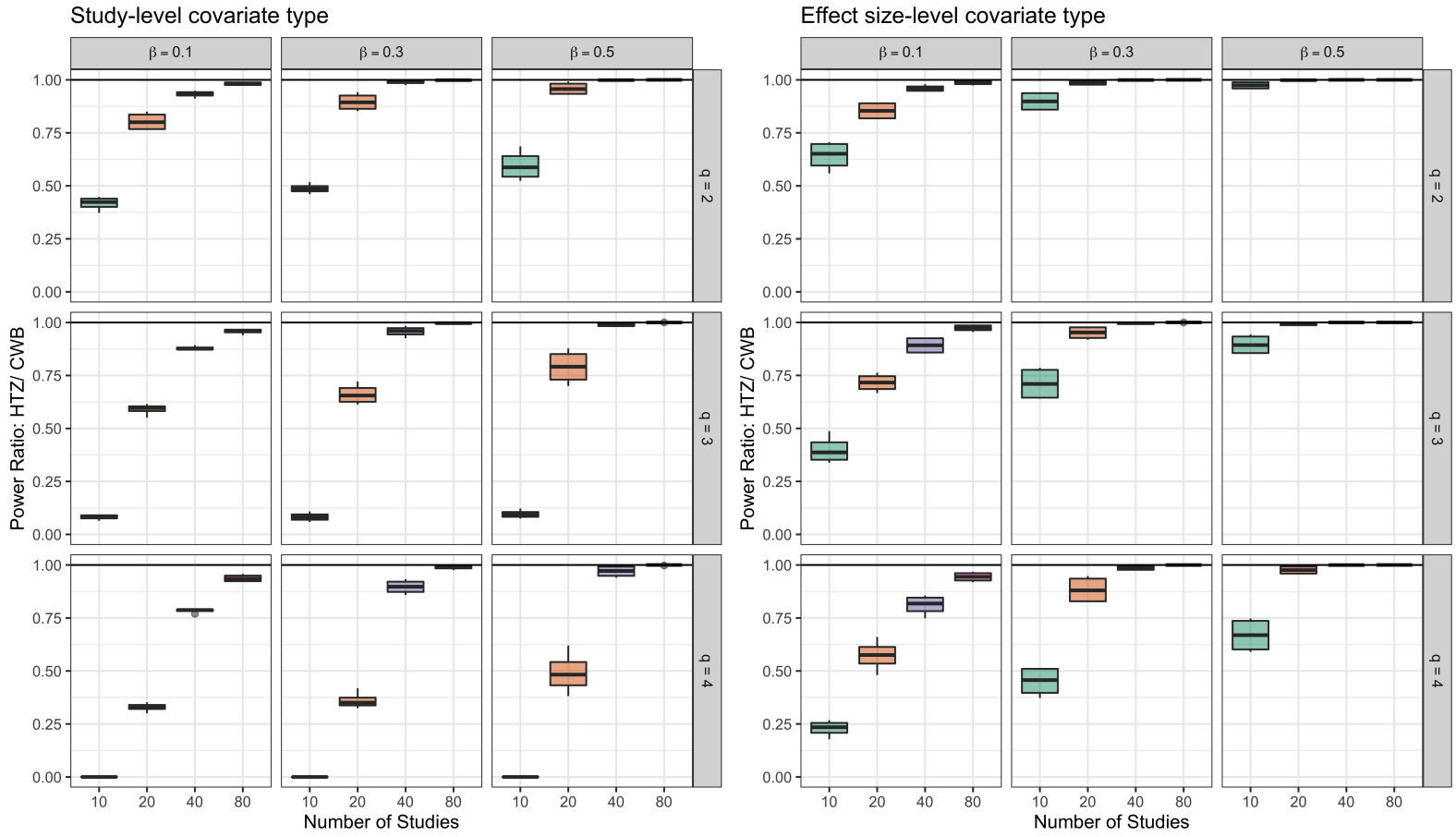

Figure 4.32. Study 2: Ratio of power of the HTZ test and the CWB test by the number of studies, the regression coefficient used to generate the true effect sizes (β), the number of contrasts (q), and the covariate type for nominal α level of 0.10. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

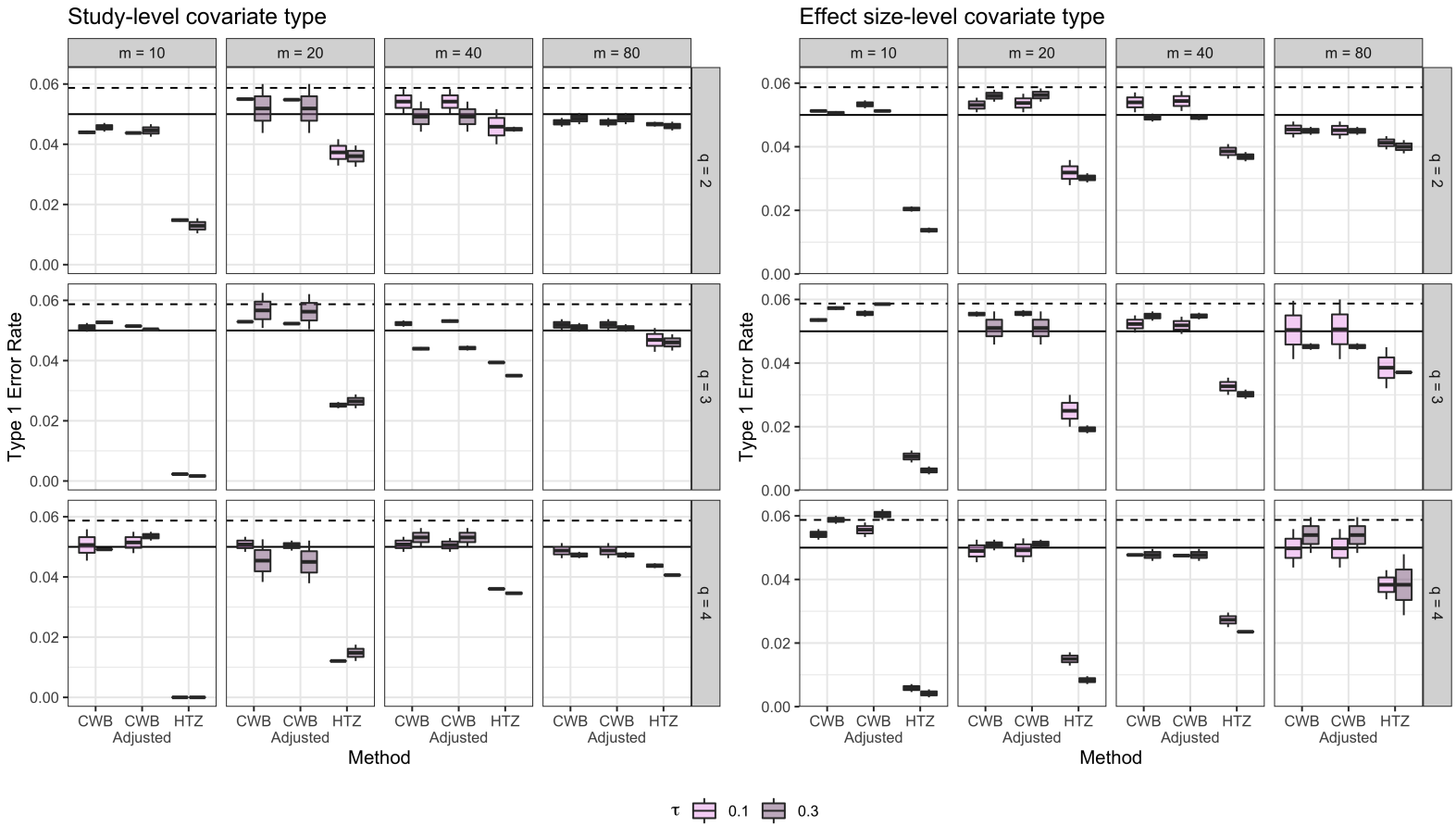

Figure 4.33. Study 2: Sensitivity of Type I error rates results to τ values: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m), the number of contrasts (q), τ values, and the covariate type for nominal α level of 0.05. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error.

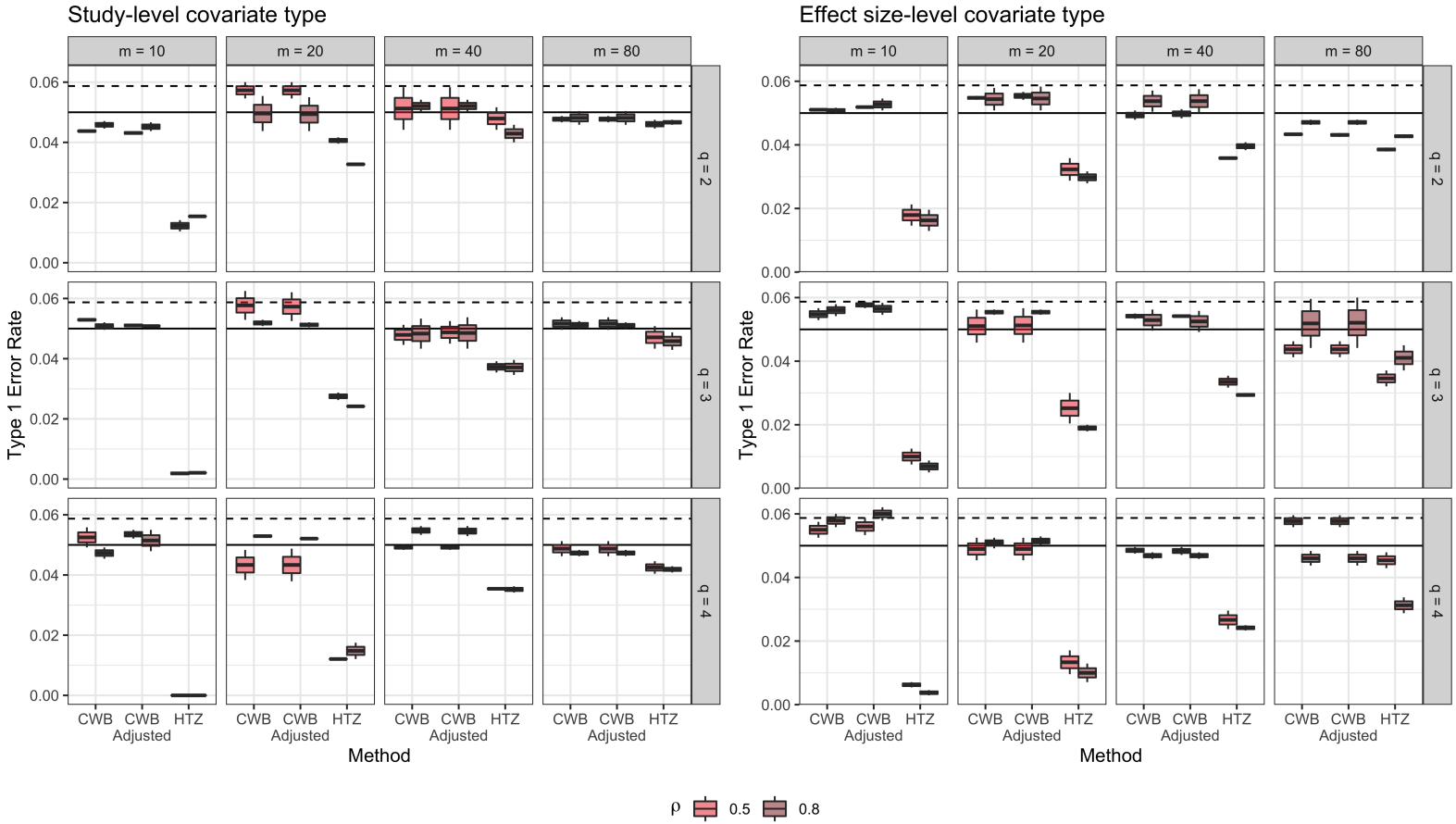

Figure 4.34. Study 2: Sensitivity of Type I error rates results to ρ values: Type I error rates of the CWB, CWB Adjusted, and HTZ tests by the number of studies (m), the number of contrasts (q), ρ values, and the covariate type for nominal α level of 0.05. The solid lines indicate the nominal α level. The dashed lines indicate bounds for simulation error.

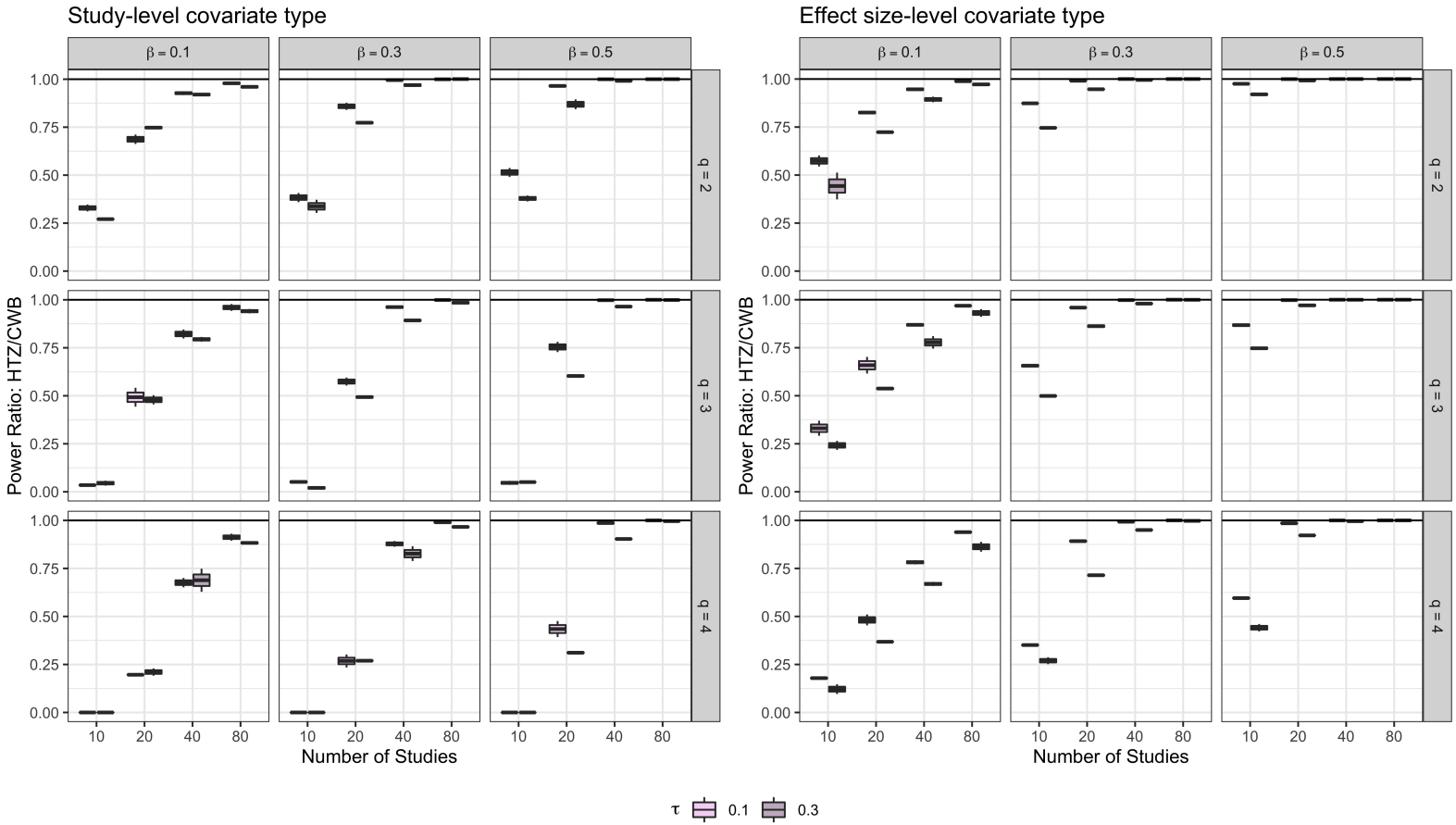

Figure 4.35. Study 2: Sensitivity of power results to τ values: Ratio of power of the HTZ test and the CWB test by the number of studies, the regression coefficient used to generate the true effect sizes (β), the number of contrasts (q), τ values, and the covariate type for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

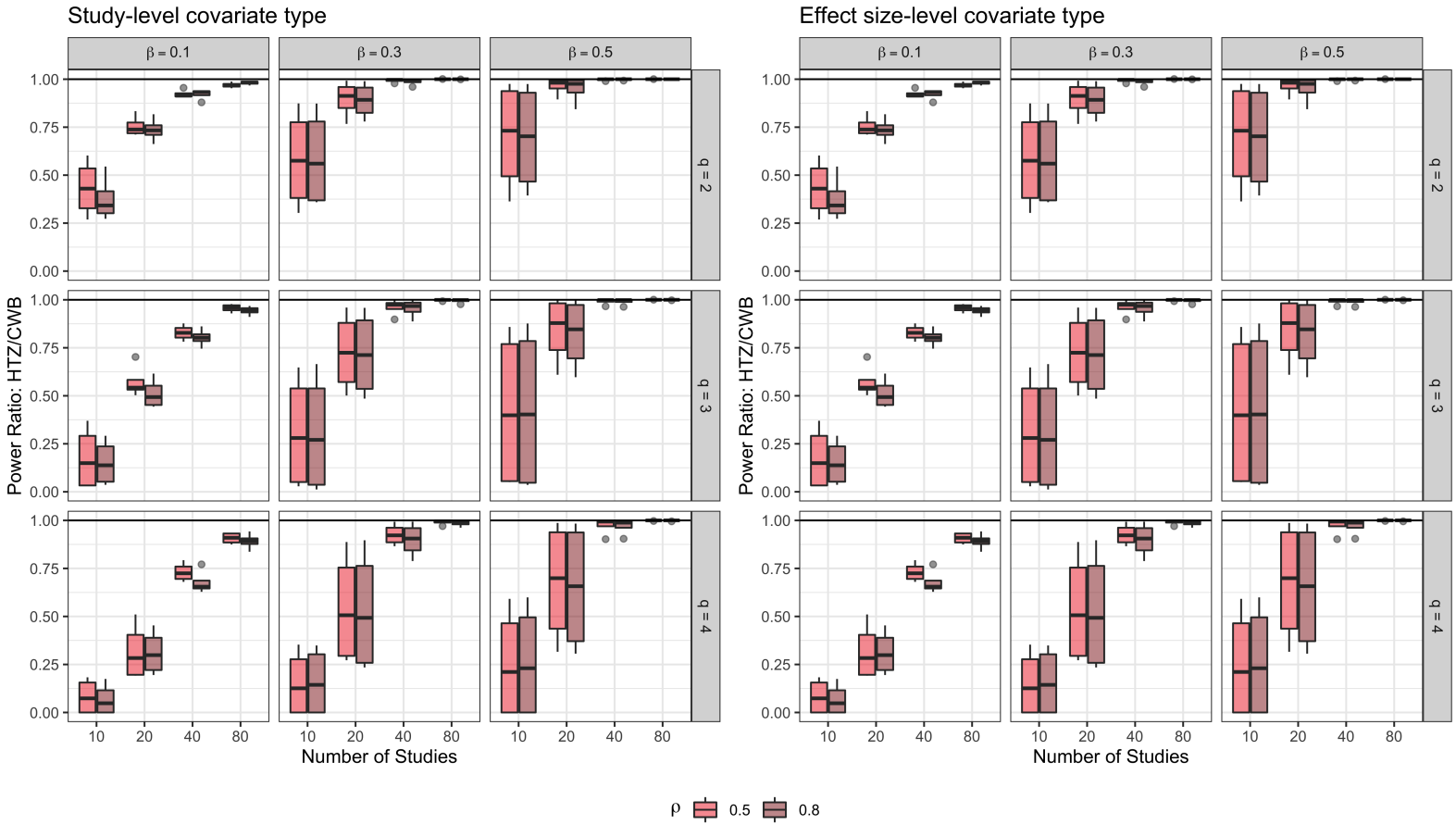

Figure 4.36. Study 2: Sensitivity of power results to ρ values: Ratio of power of the HTZ test and the CWB test by the number of studies, the regression coefficient used to generate the true effect sizes (β), the number of contrasts (q), ρ values, and the covariate type for nominal α level of 0.05. The solid lines at 1 indicate no advantage in power of the HTZ test compared to that of the CWB test. Ratios below the solid lines indicate that CWB has higher power.

Chapter 5

Software Implementation

To ease the implementation of the CWB test for applied researchers, I created an R package called `wildmeta` available on `GitHub` (Joshi et al., 2020). The major function in the package is called `cwb()`, which runs the CWB and the CWB Adjusted tests for meta-analytic models. The function currently works with models fit using the `robumeta` package (Fisher et al., 2017).

5.1 Downloading the Package

Currently, the package is available for installation through `GitHub`. The following code can be used to install the package. Note that the `devtools` package must be installed first (Wickham et al., 2020).

```
# install.packages("devtools")
devtools::install_github("meghapsimatrix/wildmeta")
```

5.2 Documentation

The arguments required to run the function, the output, and a short example are detailed at: https://meghapsimatrix.github.io/wildmeta/reference/cwb.html. The package also contains a vignette that explains the algorithm underlying the CWB test and provides a thorough example on how to implement the CWB test. The vignette is available at: https://meghapsimatrix.github.io/wildmeta/articles/cwbmeta.html.

5.3 Future Development

Future additions to the package will include the extension of the `cwb()` function to incorporate models fit using the `metafor` package (Viechtbauer, 2010). I will also extend the function to work with the working models detailed in Pustejovsky and Tipton (2020). Moreover, I plan to add another function that will calculate bootstrap-based confidence intervals for tests of single coefficients.

Chapter 6

Discussion

6.1 Summary and Implications

Primary studies often report multiple dependent estimates of effect sizes (Hedges et al., 2010). Meta-analytic techniques that ignore dependence can lead to incorrectly estimated standard errors and thus, incorrect inferences from hypothesis tests. Hedges et al. (2010) proposed the use of RVE to handle dependence. However, RVE, as originally proposed by Hedges et al. (2010), can result in inflated Type I error rates when the number of the studies is small (Hedges et al., 2010; Tipton, 2015). Tipton (2015) and Tipton and Pustejovsky (2015) have proposed and evaluated several small sample correction methods for tests of single coefficients and tests of multiple-contrast hypotheses. Although the proposed correction method, the HTZ test, has been shown to control Type I error rates adequately, the results from Tipton and Pustejovsky (2015) suggested that the test may not have adequate power, especially for multiple-contrast hypothesis tests.

In this dissertation, I examined an alternative method, cluster wild bootstrapping, that has been examined in the econometrics literature, but not in the context of meta-analysis. I compared the performance of CWB to those of the Naive-F test and the HTZ test. I replicated parts of the results from Tipton and Pustejovsky (2015). The Naive F-test resulted in high Type I error rate inflation across all conditions across both simulation studies. If meta-analysts use the Naive F-test in practice for single coefficient tests or multiple-contrast hypothesis tests, they may reject the null hypothesis when it is actually true—i.e., provide evidence for existence of the effect of a moderator when the moderator does not have any effect (Type I error). Unlike the Naive-F test, the HTZ test maintained Type I error rates across all conditions. However, replicating the findings of Tipton and Pustejovsky (2015), my results showed that the HTZ test resulted in Type I error rates that were below the nominal level. For tests of high number of contrasts with data containing small number of studies, the HTZ test had Type I error rates near 0. The HTZ test controlled Type I error rates better compared to the Naive F-test. However, the results showed that the test was too conservative especially for multiple-contrast hypothesis tests. If applied

meta-analysts use the HTZ test, they may fail to reject the null when it is actually false—i.e., fail to discover an effect when it is actually present (Type II error).

On the other hand, the results from my two simulation studies showed that the CWB and the CWB Adjusted test had Type I error rates near the nominal α level for α values of 0.01, 0.05, and 0.10 across all conditions that I examined across both simulation studies. Furthermore, the CWB test had higher power compared to the HTZ test in almost all of the conditions that I examined. The one exception was in the condition with 10 studies for single coefficient tests in Study 1. The advantage in power when using the CWB test rather than using the HTZ test was especially high for tests of higher number of contrasts. The results also showed that the CWB and the CWB Adjusted tests did not differ in Type I error rates or power. The CWB and the CWB Adjusted tests controlled Type I error rates adequately meanwhile providing better power compared to the HTZ test, especially for multiple-contrast hypothesis tests. Therefore, the CWB tests balance between the Type I and Type II error rates much better than the Naive F-test and the HTZ test. For multiple-contrast hypothesis tests, I recommend using the CWB test instead of the HTZ test as it provides more power and maintains Type I error rates adequately.

6.2 Explanations

The Naive F-test performed poorly in terms of controlling Type I error rates, especially in Study 1. Tipton and Pustejovsky (2015) explained that the poor performance of the Naive F-test can be attributed to the fact that the test only adjusts for degrees of freedom, $m - p$, and does not account for any features of the design matrix, like outliers or imbalanced categories. The design matrix in Study 1 contained worst-case scenario covariates with high leverage. Therefore, the Naive F-test performed poorly. The Naive F-test performed better in some conditions in Study 2, exhibiting Type I error rates that were close to the nominal α level. However, in Study 2, the covariates were generated to be balanced. The design matrices in Study 2 mimic ideal distributions of covariates and even then, across most of the conditions, the Naive F-test performed poorly.

In contrast to the Naive F-test, the HTZ test adjusts for leverage (Tipton & Pustejovsky, 2015). The HTZ test has two layers of correction for leverage: the first

in the CR2 adjustment of the RVE, and the second in the Satterthwaite correction for the degrees of freedom. Therefore, the HTZ test performed better than the Naive F-test in terms of controlling for the Type I error rates. However the results from Tipton and Pustejovsky (2015) and from my studies suggest that the HTZ test might be conservative. The test may over-correct for the features of the design matrix.

The results of my simulation studies replicated those of Tipton and Pustejovsky (2015), despite few differences in the data generation process and estimation methods. The design matrix that I used in Study 2 was not examined in Tipton and Pustejovsky (2015). Further, unlike Tipton and Pustejovsky (2015) who used a fixed effects working model for all their estimations, I used a correlated effects working model, including estimation of the between-study variance, to conduct meta-regression analyses. My results showed that the results of Tipton and Pustejovsky (2015) are robust across different design matrices and different model assumptions.

The CWB and the CWB Adjusted tests showed improved performance compared to the HTZ test in both Study 1 and Study 2. The CWB tests had Type I error rates around the nominal α level and also had better power than the HTZ test across almost all conditions. Cluster wild bootstrapping is not separate from RVE (MacKinnon, 2013). Rather, it works under the RVE framework as the test statistics are calculated based on the CR0 sandwich estimator (MacKinnon, 2013). Unlike the HTZ test, the CWB test does not involve explicitly correcting for leverage. However, the computational techniques of re-sampling residuals and using the bootstrap distribution to calculate p-values resulted in better performance compared to the Naive F-test and the HTZ test in terms of both controlling Type I error rates adequately and providing better power. Theoretical explanations of why the CWB test performs better than the HTZ test in terms of power are scant. Liu et al. (1988), Mammen (1993), Djogbenou et al. (2019), and MacKinnon et al. (2019) provide theorems proving the asymptotic validity and asymptotic refinement of the CWB test. However, the reason why the CWB test outperforms the HTZ test when the number of clusters is small is unclear.

The HTZ test had higher power than the CWB test for single coefficient tests for certain conditions in Study 1. Across all conditions, the power losses when using the HTZ test compared to the CWB test were smaller for single coefficient tests than for multiple-contrast hypothesis tests. For single coefficient tests, the degrees

of freedom are calculated using the Satterthwaite correction, which may accurately approximate the sampling distribution. However, for multiple-contrast hypothesis tests, the degrees of freedom are calculated using an extension of the Satterthwaite correction (Tipton & Pustejovsky, 2015; Zhang, 2012, 2013). Such an extension may not approximate the sampling distribution accurately and may result in conservative tests.

Some aspects of the results of my simulation studies were also dependent on the covariate type. Particularly, tests of the within-study covariates had higher power compared to those of the between-study covariates in both simulation studies. Further, in Study 1, tests of the continuous covariates had higher power compared to those of the binary covariates. The patterns of results replicated findings from Tipton (2015). Tipton (2015) found that the study-level covariates had small degrees of freedom. In contrast, the within-study covariates had larger degrees of freedom and the continuous covariates had even larger degrees of freedom, indicating higher power levels. Tipton (2015) explained that for study-level covariates, the Satterthwaite degrees of freedom depend on the number of studies, and not on the number of effect sizes per study. However, for within-study varying covariates, the degrees of freedom increase as the number of effect sizes within studies increases (Tipton, 2015). Continuous covariates have higher variance than binary covariates and thus, generally tend to have more power. Furthermore, Tipton (2015) found that the Satterthwaite degrees of freedom were small for highly imbalanced covariates. The CWB test algorithm does not involve estimating degrees of freedom. However, the distribution of the covariates likely influences the bootstrap distribution from which the p-values are estimated. Regardless of the covariate type, the CWB test had higher power than the HTZ test across almost all conditions, especially for multiple-contrast hypothesis tests.

Moreover, the results of my simulations showed that the CWB and the CWB Adjusted tests performed very similarly in terms of Type I error rates and power. Such lack of difference counters previous findings. MacKinnon (2013) showed that multiplying the residuals by the HC2 correction when running wild bootstrapping resulted in better control of Type I error rates compared to multiplying the residuals by the HC1 correction. However, in the cluster wild bootstrap context of my studies, multiplying the residuals with the CR2 matrices resulted in no difference. The raw

residuals underestimate the error variance even when the working model is correct (Pustejovsky & Tipton, 2018). Multiplying the residuals by the CR2 adjustment matrices should correct the under-estimation of the error variance exactly when the working model is correct and approximately when the working model is incorrect (Pustejovsky & Tipton, 2018). In my simulation studies, I only examined minor deviance from the working model, i.e, specification of the ρ value to 0.5 instead of the within-study correlation value of 0.8 as used by `robumeta`. Perhaps, major deviance from the working model—for example, using a hierarchical effects working model when the actual data structure is correlated effects—may result in more meaningful differences between the performances of the CWB and the CWB Adjusted tests.

The results were generally not sensitive to values of ρ. I only examined two values of ρ, one of which deviated from the assumed value of the within-study correlation in `robumeta`. However, results from Tipton and Pustejovsky (2015) were also not sensitive to the values of ρ. Furthermore, Hedges et al. (2010) showed that estimates of standard errors of regression coefficients and of τ^2 were generally not sensitive to the values of the within-study correlation between effect sizes when using RVE. Therefore, inferring from my results and also the results of Tipton and Pustejovsky (2015) and Hedges et al. (2010), mis-specification of the value for the within-study correlation when running tests using RVE or CWB will likely yield accurate results. The results for power ratios in my simulation studies were slightly sensitive to the values of τ. However, the overall pattern of results were not sensitive to the values of τ or ρ.

6.3 Generalizability of Results

In my simulation studies, I only generated data based on a correlated effects model, and analyzed the data using a correlated effects working model. Based on the results from my simulation studies, I cannot conclude whether the performance of the Naive F-test, the HTZ test, the CWB test, and the CWB Adjusted test would be similar across different data structures and working models like the hierarchical effects model or the more complex models detailed in Pustejovsky and Tipton (2020). Furthermore, I cannot conclude whether the results are robust to major mis-specifications of the working model.

In terms of the exact specification of the CWB test, I only studied one variation—multiplying the residuals by the CR2 adjustment matrices. The performances of the CWB and the CWB Adjusted tests were similar across all conditions. I did not study different types of weights based on conclusions from simulations conducted by Webb (2013) and MacKinnon (2015) that suggested the superiority of the Rademacher weights compared to any other weights for studies with 10 or more clusters. For analyses involving fewer than 10 clusters, Webb (2013) recommended using weights with more than two points. Furthermore, I did not examine whether imposing the null hypothesis when running bootstrapping results in better performance compared to not imposing the null. However, results from simulations in Djogbenou et al. (2019) and theoretical justifications provided by MacKinnon (2012) suggested that imposing the null can result in better control of Type I error rates. Even though I only examined small variations in the CWB algorithm, based on results from the econometrics literature, the use of the Rademacher weights and imposition of the null hypothesis when running bootstrapping should be ideal in the context of meta-analysis.

Moreover, I used the `robumeta` package set-up which uses the methods-of-moments estimator for τ^2 (Fisher et al., 2017; Hedges et al., 2010). The results should generalize to analyses using different estimators of τ^2 like the restricted maximum likelihood estimator as used in the `metafor` package (Viechtbauer, 2010). The estimates of τ^2 from different estimators should not differ so drastically and the results from my simulation studies were generally not sensitive to different values of τ.

The results of my simulation studies were generalizable across different design matrices in Study 1 and Study 2. The design matrix in Study 1 contained worst-case scenario covariates—ones with high imbalance and non-normality. The design matrices in Study 2 contained balanced covariates. I expect the results of my simulation to generalize to analyses conducted with different design matrices that contain varying types of covariates. However, the design matrices in both studies did not contain missing data. Based on the simulation studies I conducted, I cannot recommend how to use CWB or RVE when the moderators have missing data. Further, I generated data based on main effects of the covariates and used main effects meta-regression analysis to estimate regression coefficients. In real meta-analysis, it would be impossible to know the underlying data generating model. Mis-specification of the analytic

model will possibly bias the estimates of the regression coefficients.

Moreover, I only studied standardized mean differences. However, Hedges (2019a) argued theoretically about the *fundamental unity of meta-analytic methods* in that the results based on standardized mean differences should generalize to analyses conducted with other types of effect size like odds ratios and correlations. Therefore, theoretically, the results from my simulations should generalize to analyses conducted with other types of effect size measures.

I also did not examine publication bias. Researchers can possibly use CWB with methods to detect and handle publication bias when the meta-analytic data contains dependent effect sizes. These methods are detailed in Rodgers and Pustejovsky (2019) and Mathur and VanderWeele (2020).

6.4 Recommendations for Applied Researchers

Based on the results of my study as well as those of Hedges et al. (2010), Tipton (2015), and Tipton and Pustejovsky (2015), I would recommend researchers not to use the Naive-F test. The HTZ test will control Type I error rates but the test may be too conservative, especially if researchers are interested in multiple-contrast hypothesis tests. In such cases, I recommend using the CWB test. The CWB test had higher power than the HTZ test for multiple-contrast hypothesis tests even in conditions with 80 studies. Therefore, I would recommend the CWB test over the HTZ test for meta-analyses with small and moderate to large sample sizes. Although I did not examine different weights, based on results from Davidson and Flachaire (2008) and MacKinnon (2015), I would recommend the use of the Rademacher weights for meta-analyses with the number of studies as low as 10. Additionally, based on suggestions by MacKinnon (2012), I recommend imposing the null hypothesis when bootstrapping. Because the results of my simulation studies did not show any differences between the performance of the CWB and the CWB Adjusted tests, I recommend the CWB test over the CWB Adjusted test as the CWB test is conceptually and algorithmically simpler. However, the CWB test should be used over the CWB Adjusted test only under the conditions examined in my simulation studies. Furthermore, in my simulation studies, I used 399 bootstrap replications. Generally, higher number of bootstrap replications will result in higher power (Davidson & MacKinnon, 2000).

However, higher number of bootstrap replications will require more computation time. For recommendations on selecting the number of bootstraps, please see Davidson and MacKinnon (2000).

6.5 Software

In addition to examining the CWB test methodologically, I have also implemented the algorithm to run the CWB and the CWB Adjusted tests in a package called `wildmeta`. The package is specifically designed for meta-analysts. I hope the package will ease the process of implementing CWB for applied meta-analysts.

6.6 Limitations and Future Directions

Due to the computationally intensive nature of bootstrapping, I only examined a limited set of conditions. Future research can examine the performance of CWB for wider ranges of number of studies, different types of contrasts, different meta-regression model specifications, and different types of effect sizes. Future studies can also examine the performance of CWB using a different design matrix, perhaps one based on real meta-analytic data. It would be particularly helpful to examine the implementation of multiple-contrast hypothesis tests, methods to handle dependence, and small sample correction methods in meta-analytic data containing missing values in the moderators. Applied meta-analysts are likely to encounter missing data in the moderators as primary studies may not provide relevant information (Pigott, 2019).

Moreover, I only examined a correlated effects data structure paired with a correlated effects working model. I followed the study designs of Tipton (2015) and Tipton and Pustejovsky (2015); both examined correlated effects data structure and working model. Further, RVE is more pertinent for the correlated effects working model as the hierarchical effects working model has the underlying assumption that the correlation between effect sizes within a study is 0 (Hedges et al., 2010). For the correlated effects working model, RVE is robust to the specification of the correlation value between effect sizes within a study (Hedges et al., 2010). Pustejovsky and Tipton (2020) introduced further types of working models that combine correlated and hierarchical effects structures. Future studies should examine the performance of the

CWB and HTZ tests for different kinds of dependent data structures, and different working models that may or may not be specified correctly.

Future studies can also examine small sample correction methods for multi-level meta-analytic models. Although I only focused on the RVE framework for my dissertation, multi-level models are another way to correct for dependence. Multi-level models can be especially useful if the meta-analytic data structure is hierarchical. Small sample correction for multi-level models include the Kenward-Rodger correction (Kenward & Roger, 1997, 2009). To my knowledge, there is no software that specifically implements small sample corrections for multi-level meta-analysis, especially in R.

Furthermore, the examination of bootstrap-based confidence intervals was out of the scope of this dissertation. MacKinnon (2015) examined bootstrap-based confidence intervals for single coefficients in the econometrics setting. Future studies can examine bootstrap-based confidence interval coverage and width in the meta-analytic context.

Additionally, I did not center the covariates following Tipton and Pustejovsky (2015) and following what applied meta-analysts likely do. Certain aspects of the results of my simulation studies were different for the within-study and the between-study covariate types. Such differences may indicate that group mean centering a within-study moderator and including study-level averages of the moderator in meta-regression analysis might influence Type I error rates and power. Future studies can look at the effects of centering on the performance of the methods examined here.

References

Anderson, T., & Girshick, M. (1944). Some extensions of the Wishart distribution. *The Annals of Mathematical Statistics*, *15*(4), 345–357.

Becker, B. J. (2000). Multivariate meta-analysis, In *Handbook of applied multivariate statistics and mathematical modeling*. Elsevier.

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin*, *144*(1), 77–110. https://doi.org/10.1037/bul0000130

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169–182.

Boos, D. D. Et al. (2003). Introduction to the bootstrap world. *Statistical science*, *18*(2), 168–174.

Borenstein, M., & Hedges, L. V. (2019). Effect sizes for meta-analysis. *The handbook of research synthesis and meta-analysis*, 207–243.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research synthesis methods*, *8*(1), 5–18.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 47.

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101–129.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (Vol. 2). Sage.

Davidson, R., & Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, *146*(1), 162–169. https://doi.org/10.1016/j.jeconom.2008.08.003

Davidson, R., & MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews*, *19*(1), 55–68.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177–188.

Djogbenou, A. A., MacKinnon, J. G., & Nielsen, M. Ø. (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics*, *212*(2), 393–412.

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Robumeta: Robust variance meta-regression* [R package version 2.0]. R package version 2.0. https://CRAN.R-project.org/package=robumeta

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, *9*(6), 1218–1228.

Freedman, D. A. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models. *The Annals of Statistics*, *12*(3), 827–842.

Gallet, C. A., & Doucouliagos, H. (2014). The income elasticity of air travel: A meta-analysis. *Annals of Tourism Research*, *49*, 141–155. https://doi.org/10.1016/j.annals.2014.09.006

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2020). *mvtnorm: Multivariate normal and t distributions* [R package version 1.1-0]. R package version 1.1-0. https://CRAN.R-project.org/package=mvtnorm

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

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171.

Hedges, L. V. (2019a). Statistical considerations. *The handbook of research synthesis and meta-analysis*, *3*, 37–48.

Hedges, L. V. (2019b). Stochastically dependent effect sizes. *The handbook of research synthesis and meta-analysis*, 281–297.

Hedges, L. V., & Cooper, H. M. (2009). Research synthesis as a scientific process. *The handbook of research synthesis and meta-analysis*, *1*.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, *21*(11), 1539–1558.

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137–159.

Hotelling, H. (1931). Hotelling t2 test. *Ann Math Stat*, *2*, 360.

Joshi, M., & Pustejovsky, J. E. (2020). *simhelpers: Helper functions for simulation studies* [R package version 0.1.0]. R package version 0.1.0. https://CRAN.R-project.org/package=simhelpers

Joshi, M., Pustejovsky, J. E., & Cappelli, P. (2020). *wildmeta: Cluster wild bootstrapping for meta-analysis* [R package version 0.0.0]. R package version 0.0.0.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983–997.

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*(7), 2583–2595.

Konstantopoulos, S., & Hedges, L. V. (2019). Statistically analyzing effect sizes: Fixed-and random-effects models. *The Handbook of Research Synthesis and Meta-Analysis,* 245–279.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., De Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? a meta-analytic review. *Psychological bulletin*, *144*(4), 394.

Liu, R. Y. Et al. (1988). Bootstrap procedures under some non-iid models. *The annals of statistics*, *16*(4), 1696–1708.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics. *Economic Record*, *82*, S2–S18.

MacKinnon, J. G. (2009). Bootstrap hypothesis testing. *Handbook of computational econometrics*, *183*, 213.

MacKinnon, J. G. (2012). Inference Based on the Wild Bootstrap, 34.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference, In *Recent advances and future directions in causality, prediction, and specification analysis*. Springer.

MacKinnon, J. G. (2015). Wild cluster bootstrap confidence intervals. *L'Actualité économique*, *91*(1-2), 11–33.

MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2019). Wild bootstrap and asymptotic inference with multiway clustering. *Journal of Business & Economic Statistics*, 1–15.

MacKinnon, J. G., & Webb, M. D. (2017). Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Applied Econometrics*, *32*(2), 233–254. https://doi.org/10.1002/jae.2508

MacKinnon, J. G., & Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, *21*(2), 114–135. https://doi.org/10.1111/ectj.12107

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, *29*(3), 305–325.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The annals of statistics*, 255–285.

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119.

McCaffrey, D. F., Bell, R. M., & Botts, C. H. (2001). Generalizations of biased reduced linearization, In *Proceedings of the annual meeting of the american statistical association*.

McEwan, P. J. (2015). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research*, *85*(3), 353–394. https://doi.org/10.3102/0034654314553127

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074–2102.

Oczkowski, E., & Doucouliagos, H. (2015). Wine Prices and Quality Ratings: A Meta-regression Analysis. *American Journal of Agricultural Economics*, *97*(1), 103–121. https://doi.org/10.1093/ajae/aau057

Ola, O., & Menapace, L. (2020). A meta-analysis understanding smallholder entry into high-value markets. *World Development*, *135*, 105079.

Olkin, I., & Gleser, L. (2009). Stochastically dependent effect sizes. *The handbook of research synthesis and meta-analysis*, 357–376.

Park, S. Y., & Beretvas, S. N. (2016). Assessing estimation of the three-level meta-analysis model: Synthesizing effect sizes for multiple outcomes per study. *Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C.*

Pedersen, T. L. (2020). *Patchwork: The composer of plots* [R package version 1.1.1]. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork

Pigott, T. D. (2012). *Advances in meta-analysis*. Springer Science & Business Media.

Pigott, T. D. (2019). Handling missing data. *The handbook of research synthesis and meta-analysis*, *3*, 367–381.

Pustejovsky, J. E. (2020a). *Clubsandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* [R package version 0.4.2]. R package version 0.4.2. https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E. (2020b). *Pusto: Pusto's miscellaneous data analysis and simulation tools* [R package version 0.3.0]. R package version 0.3.0. https://github.com/jepusto/Pusto

Pustejovsky, J. E., & Tipton, E. (2018). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics*, *36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Pustejovsky, J. E., & Tipton, E. (2020). Meta-analysis with robust variance estimation: Expanding the range of working models.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. *The handbook of research synthesis and meta-analysis*, *2*, 295–316.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*(1), 111.

Rice, K., Higgins, J. P., & Lumley, T. (2018). A re-evaluation of fixed effect (s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(1), 205–227.

Rodgers, M. A., & Pustejovsky, J. E. (2019). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes.

Sala, G., Tatlidil, K. S., & Gobet, F. (2018). Video game training does not enhance cognitive ability: A comprehensive meta-analytic investigation. *Psychological bulletin*, *144*(2), 111.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of educational research*, *84*(3), 328–364.

Swanson, H. L., Trainin, G., Necoechea, D. M., & Hammill, D. D. (2003). Rapid naming, phonological awareness, and reading: A meta-analysis of the correlation evidence. *Review of Educational Research*, *73*(4), 407–440.

Tanner-Smith, E. E., & Lipsey, M. W. (2015). Brief alcohol interventions for adolescents and young adults: A systematic review and meta-analysis. *Journal of substance abuse treatment*, *51*, 1–18.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling Complex Meta-analytic Data Structures Using Robust Variance Estimates: A Tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*(1), 85–112. https://doi.org/10.1007/s40865-016-0026-5

Thompson, T., Oram, C., Correll, C. U., Tsermentseli, S., & Stubbs, B. (2017). Analgesic effects of alcohol: A systematic review and meta-analysis of controlled experimental studies in healthy participants. *The Journal of Pain*, *18*(5), 499–510.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

Tipton, E., & Pustejovsky, J. E. (2015). Small-Sample Adjustments for Tests of Moderators and Model Fit Using Robust Variance Estimation in Meta-Regression.

Journal of Educational and Behavioral Statistics, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research synthesis methods*, *10*(2), 180–194.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, *7*(1), 55–79.

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

Viechtbauer, W. (2007). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie/Journal of Psychology*, *215*(2), 104–121.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of statistical software*, *36*(3), 1–48.

Webb, M. D. (2013). *Reworking wild bootstrap based inference for clustered errors* (tech. rep.). Queen's Economics Department Working Paper.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., Hester, J., & Chang, W. (2020). *devtools: Tools to make developing r packages easier* [R package version 2.3.2]. R package version 2.3.2. https://CRAN.R-project.org/package=devtools

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, *14*(4), 1261–1295.

Zhang, J.-T. (2012). An approximate Hotelling T2-test for heteroscedastic one-way MANOVA. *Open Journal of Statistics*, *2*(1), 1–11.

Zhang, J.-T. (2013). Tests of linear hypotheses in the ANOVA under heteroscedasticity. *International Journal of Advanced Statistics and Probability*, *1*(2), 9–24.