2006

A Comparison Of Ordinary Least Squares, Weighted Least Squares, And Other Procedures When Testing For The Equality Of Regression

Patrick J. Rosopa
University of Central Florida, prosopa@clemson.edu

A COMPARISON OF ORDINARY LEAST SQUARES, WEIGHTED LEAST SQUARES, AND OTHER PROCEDURES WHEN TESTING FOR THE EQUALITY OF REGRESSION SLOPES WITH HETEROSCEDASTICITY ACROSS GROUPS: A MONTE CARLO STUDY

by

PATRICK J. ROSOPA
B.S. Tulane University, 2000
M.S. University of Central Florida, 2003

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Fall Term
2006

Major Professor: Eugene F. Stone-Romero

ABSTRACT

When testing for the equality of regression slopes based on ordinary least squares (OLS) estimation, extant research has shown that the standard *F* performs poorly when the critical assumption of homoscedasticity is violated, resulting in increased Type I error rates and reduced statistical power (Box, 1954; DeShon & Alexander, 1996; Wilcox, 1997). Overton (2001) recommended weighted least squares estimation, demonstrating that it outperformed OLS and performed comparably to various statistical approximations. However, Overton's method was limited to two groups. In this study, a generalization of Overton's method is described. Then, using a Monte Carlo simulation, its performance was compared to three alternative weight estimators and three other methods. The results suggest that the generalization provides power levels comparable to the other methods without sacrificing control of Type I error rates. Moreover, in contrast to the statistical approximations, the generalization (a) is computationally simple, (b) can be conducted in commonly available statistical software, and (c) permits post hoc analyses. Various unique findings are discussed. In addition, implications for theory and practice in psychology and future research directions are discussed.

ACKNOWLEDGMENTS

I would like to thank my major advisor, Dr. Eugene Stone-Romero, for his career-related guidance and support. He has taught me to persevere even in the face of adversity. I would also like to thank Dr. Kim Smith-Jenstch for her kindness and continuous psychosocial support. I have softened my handshake slightly so now you cannot call me "The Crusher" anymore. I would like to thank Dr. Florian Jentsch for his valuable insights and practical advice. To Dr. Herman Aguinis, thank you for your sage advice. Your online programs were critical in completing this manuscript. I would also like to thank Dr. Janet Ruscher for "giving me a chance" when others would not. I would also like to thank her for encouraging me to take a doctoral course with Dr. Bill Dunlap (1969-2002). I believe his lucid explanations (and humor) provided a strong foundation for me to more confidently pursue graduate study (and improve my skills at using the "confuser," i.e., computer, as he would call it). Dr. Dunlap played an important role in my career development and I wish I had a chance to tell him.

I would like to thank my parents, wife, and daughters. Without them, I would be incomplete. My parents taught me the value of knowledge; because of them, I recognize the importance of education and have grown in many ways. In addition, for two years, my parents provided social support as I commuted between Jacksonville and Orlando on the weekends. Thank you Mom and Dad.

My wife, Melodee, has been amazingly patient and understanding as we juggled our work and school schedules while raising our girls (from day-care to elementary school, from soccer practice to swimming lessons, etc.) and we managed to survive three moves in five years. Somehow, "I love you" just does not seem enough. To Cecilia and Elenah, "Dr. Daddy" can't stop working, but I can make more time to play.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS/ABBREVIATIONS

ANOVA	analysis of variance
BLUE	best linear unbiased estimator
df	degrees of freedom
HCCM	heteroscedasticity consistent covariance matrix
k	number of independent groups (or levels) of a categorical variable
ML	maximum likelihood
N	sample size
n_j	subgroup sample size for a specific group (or level)
OLS	ordinary least squares
RML	restricted maximum likelihood
SSE	sum of squared errors
WLS	weighted least squares
x	continuous predictor (or independent) variable
y	continuous criterion (or dependent) variable
z	categorical predictor (or independent) variable

CHAPTER ONE: INTRODUCTION

Testing for the equality of regression slopes is frequently conducted in a variety of psychological research settings. Evidence of this can be found in research on aptitude by treatment interactions (Cronbach & Snow, 1977; Smith & Sechrest, 1991), differential prediction (Cleary, 1968; Linn, 1978; Saad & Sackett, 2002), and analysis of covariance (Huitema, 1980; Rutherford, 1992). Parameters in a linear model are typically estimated using ordinary least squares (OLS). Then, testing for the equality of regression slopes is conducted using an *F* test. As described below, research has shown that this test performs poorly when the critical assumption of homoscedasticity is violated (Aguinis & Pierce, 1998; Alexander & DeShon, 1994; Box, 1954; DeShon & Alexander, 1996; Wilcox, 1997). In this paper, I proffer an alternative data-analytic solution which extends an approach described by Overton (2001). Thus, the main purpose of this paper is to describe the extension and compare its performance to alternative procedures. First, however, the OLS-based approach for testing the equality of regression slopes is reviewed.

Testing for the Equality of Regression Slopes Using OLS

When testing for the equality of regression slopes using OLS, a continuous response (y) is modeled as a function of a continuous predictor (x), a categorical predictor (z) with k levels (indexed by $k - 1$ regressors, i.e., $z_1, z_2, \ldots, z_{k-1}$), and the two-way interaction between x and z (indexed by $k - 1$ product terms between x and the regressors). It deserves mentioning that population parameters are denoted by Greek letters such as β and σ^2 for example, to differentiate them from sample estimates indicated with a diacritic such as $\hat{\beta}$ and $\hat{\sigma}^2$, respectively.

For $k = 2$, the full linear model for the ith observed response can be expressed as:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 x_i z_{1i} + e_i \qquad (1)$$

for $i = 1, 2, \ldots, N$, where N = total sample size; β_0, β_1, β_2, and β_3 are unstandardized regression coefficients; and e_i is the ith residual, an estimate of ε_i (an unknown error). A statistically significant t for the test of the null hypothesis $\beta_3 = 0$ indicates that the two population regression slopes are unequal. Stated differently, x and z interact in estimating $E(y_i)$.[1]

More generally, for $k \geq 2$, the full linear model for the N observations (with the number of terms $p = 2k - 1$) can be compactly expressed in matrix form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2}$$

where $\mathbf{y}$ is an $N \times 1$ response vector, $\mathbf{X}$ is an $N \times (p + 1)$ model matrix, $\boldsymbol{\beta}$ is a $(p + 1) \times 1$ vector of unstandardized regression coefficients, and $\mathbf{e}$ is an $N \times 1$ residual vector. In addition, it is assumed that the first-order and second-order moments of $\mathbf{e}$ have $E(\mathbf{e}) = \mathbf{0}$ and $\text{cov}(\mathbf{e}) = \sigma^2 \mathbf{I}_N$, respectively (where $\mathbf{0}$ = a null vector, σ^2 = the common variance, and $\mathbf{I}_N$ = an identity matrix of order N) (Schott, 1997). These assumptions state that the model is linear, all relevant terms are included in the model, and ε_is are constant and uncorrelated. Note that normally distributed ε_is are neither required nor assumed for the above-noted model to be valid (Rencher, 2000). However, when the normality assumption is invoked, for hypothesis testing, this implies that the y_is (and ε_is) are independent.

The best linear unbiased estimator (BLUE) of the parameters in Equation 2 is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3}$$

Although $\mathbf{X}$, in Equation 3, can be partitioned differently, for convenience:

$$\mathbf{X} = [\mathbf{j}\ \mathbf{x}\ \mathbf{D}_z\ \mathbf{D}_{xz}] \tag{4}$$

where $\mathbf{j}$ is an $N \times 1$ vector of 1s, $\mathbf{x}$ is an $N \times 1$ vector of the continuous predictor, $\mathbf{D}_z$ is an $N \times (k - 1)$ matrix of regressors, and $\mathbf{D}_{xz}$ is an $N \times (k - 1)$ matrix of product terms between $\mathbf{x}$ and the

regressors in $\mathbf{D}_z$. Based on Equation 3 and the constant variance assumption, an unbiased estimator of σ^2 can be expressed as:

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{N - p - 1} = \frac{SSE}{N - p - 1} \tag{5}$$

where *SSE* = sum of squared errors. Moreover, when **e** is normally distributed, maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are given by Equation 3 and SSE / N, respectively (Rencher, 2000).

Although **X** in Equation 4 represents the full model matrix, a full-and-reduced linear model approach can be used to construct the test of whether the *k* population regression slopes are equal (Rencher, 2000). The reduced model matrix ($\mathbf{X}_{\text{Reduced}}$) excludes $\mathbf{D}_{xz}$. Then, $\boldsymbol{\beta}_{\text{Reduced}}$ becomes a $(k + 1) \times 1$ vector of regression coefficients.

Assuming that **e** is normally distributed, the test for the equality of regression slopes is conducted using an *F* ratio. It assesses whether the decrease in the *SSE* from a reduced (SSE_{Reduced}) to a full (SSE_{Full}) model is statistically significant. The *F* random variable can be expressed as:

$$F = \frac{(SSE_{\text{Reduced}} - SSE_{\text{Full}}) / df_O}{SSE_{\text{Full}} / df_E} \tag{6}$$

where df_O = the number of terms omitted from the full model and df_E = the error degrees of freedom for the full model. An equivalent general linear hypothesis test can be conducted using the full model (see Equation 8.27 in Rencher, 2000). If $F > F(1 - \alpha, df_O, df_E)$ (where α = Type I error rate), then the hypothesis of equality of regression slopes is rejected; stated differently, *x* and *z* interact in estimating $E(y_i)$. Otherwise, the null hypothesis of equal regression slopes cannot be rejected. These procedures are described in greater detail in numerous texts (see

Cohen, Cohen, West, & Aiken, 2003; Draper & Smith, 1966; Maxwell & Delaney, 2000; Neter, Kutner, Nachtsheim, & Wasserman, 1996). Hereinafter, the *F* test based on OLS estimation is denoted by F_{OLS}.

As mentioned above, when the important assumption of homoscedasticity is violated, the F_{OLS} test of the interaction between *x* and *z* no longer performs accurately. In particular, Monte Carlo studies indicate that Type I error rates are biased and statistical power is reduced (Alexander & DeShon, 1994; DeShon & Alexander, 1996). Moreover, this assumption is not uncommon to violate (Aguinis, 2004; Alexander & DeShon, 1994; DeShon & Alexander, 1996; Luh & Guo, 2002; Overton, 2001; Wilcox, 1997). For example, based on a review of three journals (*Academy of Management Journal, Journal of Applied Psychology,* and *Personnel Psychology*) from 1987 to 1999, Aguinis, Peterson, and Pierce (1999) identified 87 articles that reported at least one test for the equality of regression slopes. Out of 117 tests, Aguinis and his colleagues found that at least 39% of these violated the assumption. The implication of this finding is that researchers might have wrongly concluded that an interaction *exists* in the population *when it does not* (Type I error) or that an interaction *does not exist* in the population *when it does* (Type II error). In either case, "substantive research conclusions can be erroneous, theory development can be hindered, and incorrect decisions can be made…" (Aguinis et al., p. 319).

In the following sections, three main issues are discussed: (a) homoscedasticity in linear models when testing for the equality of regression slopes, (b) proposed remedies when homoscedasticity is violated in such models, and (c) weighted least squares regression as an alternative solution.

Homoscedasticity in Linear Models

A number of psychological disciplines, such as clinical psychology, educational psychology, and industrial and organizational psychology, use general linear models to analyze data (e.g., analysis of variance [ANOVA] and OLS regression). Hypothesis tests in such models require that the conditional variance of *y* remain constant about the fitted surface. This homoscedasticity assumption can be denoted by $\text{var}(e_i) = \sigma^2$ for all *i*. In linear models with all categorical predictors (viz., ANOVA), this assumption is also referred to as homogeneity of variance (Wilcox, 1996). In the following paragraphs, I first provide a clarification of the homoscedasticity assumption, along with the form of nonconstant variance (i.e., heteroscedasticity) addressed in this paper. Then, the effects of heteroscedasticity in OLS multiple regression, when testing for the equality of regression slopes, are discussed.

A Clarification of Homoscedasticity

When defining homoscedasticity, and depending on the type of predictors in a model, some researchers employ different terminology claiming that these terms are distinctly different concepts. For example, it has been asserted that *homogeneity of error variance* applies to linear models testing for the equality of regression slopes and that *homoscedasticity* applies more generally to all OLS regression models (Aguinis & Pierce, 1998). However, the variation of the response should be the same around the fitted surface whether dealing with a single line or multiple higher dimensional hyperplanes. If the conditional variance of *y* in the fitted model differs depending on (a) the value of a continuous predictor, (b) the level of a categorical predictor, (c) the fitted values, or (d) a linear combination of the predictors in the model, these *all* represent nonconstant variance, namely, heterogeneity of variance or heteroscedasticity (Carroll & Ruppert, 1988; Casella & Berger, 2002; Cook & Weisberg, 1983, 1999; Fox, 1997; Neter et

al., 1996; Rencher, 2000). However, in fitting any model, specifying the correct features of the error structure is left to the researcher to identify accurately. Are errors normally distributed? Are errors constant or nonconstant? Do errors increase in variability as a function of a continuous predictor and/or a categorical predictor? Do errors vary as a function of an omitted predictor? Two examples are next described to illustrate heteroscedasticity in linear models.

The first example is a one-way ANOVA with $k = 3$. Using two regressors, the ith observed response can be parameterized as $y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + e_i$. This model is assumed to have a diagonal covariance matrix for the error term denoted by:

$$\text{cov}(\mathbf{e}) = \sigma^2 \mathbf{I}_N = \begin{pmatrix} \sigma^2 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sigma^2 \end{pmatrix} \tag{7}$$

Namely, errors are homoscedastic. However, heteroscedasticity can occur such that:

$$\text{cov}(\mathbf{e}) = \sigma_i^2 \mathbf{I}_N = \begin{pmatrix} \sigma_1^2 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sigma_N^2 \end{pmatrix} \tag{8}$$

where $\sigma_i^2 \neq \sigma_{i'}^2$ for some i and i'. For convenience, however, assume the data are ordered such that the first n_1 observations are from Group 1 (where n_j denotes the jth sample size for $j = 1, 2, \ldots, k$). Assume further that the next n_2 observations are from Group 2 and the remaining observations are from Group 3. Heteroscedasticity could result such that $(\sigma_1^2 = \ldots = \sigma_{n_1}^2) \neq (\sigma_{n_1+1}^2 = \ldots = \sigma_{n_1+n_2}^2) \neq (\sigma_{n_1+n_2+1}^2 = \ldots = \sigma_N^2)$. Therefore, the N diagonal elements of Equation 8 would assume three different values—one for each group. Thus, the error variance changes depending on the level of the categorical predictor.

The second example is the model in Equation 1 which fits an intercept and slope for each of two groups. Again, Equation 7 is assumed to be true and the data are ordered such that the first n_1 observations are from Group 1 and the remaining n_2 observations are from Group 2. Heteroscedasticity might exist in the population as a function of the categorical predictor only. Namely, the diagonal elements of Equation 8 could have the form ($\sigma_1^2 = \ldots = \sigma_{n_1}^2$) $\neq$ ($\sigma_{n_1+1}^2 = \ldots = \sigma_N^2$), assuming two different values—one for each group. This also represents heteroscedasticity because the error variance changes depending on the level of the categorical predictor.

Note that both examples were linear models which assumed homoscedasticity. When this was violated, the result was heteroscedasticity. For pedagogical purposes, it is averred that any form of nonconstant variance is heteroscedasticity and, therefore, the exclusivity of terminology, such as (a) *heteroscedasticity* for models with continuous predictors or (b) *heterogeneity of error variance* for models with continuous and categorical predictors, is not used in the present paper. Instead, consistent with many research literatures (Breusch & Pagan, 1979; Carroll & Ruppert, 1988; Casella & Berger, 2002; Cohen et al., 2003; Cook & Weisberg, 1983; Fry, 1994; Greene, 2003, Jolicoeur, 1999; Neter et al., 1996; Pinheiro & Bates, 2000; Rencher, 2000; White, 1980; Wilcox, 1996), nonconstant variance, heteroscedasticity, and heterogeneity of error variance are used interchangeably because these terms are synonymous. However, different forms of heteroscedasticity can occur and researchers should identify the form detected in the fitted model. For example, based on the model in Equation 1, Wilcox (1997) distinguished between Type I heteroscedasticity (i.e., nonconstant variance across the values of x within groups), Type II heteroscedasticity (i.e., nonconstant variance between groups), and complete heteroscedasticity (i.e., where both Type I and Type II exist).

The present investigation focuses on Type II heteroscedasticity. As research by Aguinis et al. (1999) suggests, this form is likely to occur because bivariate data (i.e., *y* and *x*) collected across independent groups (e.g., female versus male, ethnic group, etc.) may exhibit different error variances because differential validity tends to increase the likelihood of heteroscedasticity. Because this paper addresses only this form of heteroscedasticity, for convenience, the modifier "Type II" hereinafter was omitted.

Effects of Heteroscedasticity When Testing for the Equality of Regression Slopes

The biasing effects of heteroscedasticity on statistical inferences (i.e., hypothesis tests, confidence intervals, joint confidence regions, etc.) in OLS multiple regression analyses are nontrivial. These effects and their implications for both theory and practice in psychology are next discussed.

The error variance in the *j*th group can be expressed as:

$$\sigma_{e_j}^2 = \sigma_{y_j}^2(1-\rho_{y_j x_j}^2) \qquad (9)$$

where $\sigma_{y_j}^2$ and $\rho_{y_j x_j}$, respectively, are the variance of *y* and correlation coefficient between *y* and *x* in the *j*th group (Cook & Weisberg, 1999). Homoscedasticity obtains when $\sigma_{e_1}^2 = \ldots = \sigma_{e_k}^2$.

As noted above, research suggests that the statistical assumption of homoscedasticity is likely to be violated. Inspection of Equation 9 shows that if $\sigma_{y_j}^2$ (or $\rho_{y_j x_j}$) is constant across the *k* groups, then any difference in $\rho_{y_j x_j}$ (or $\sigma_{y_j}^2$) across the *k* groups will result in heteroscedasticity, unless values for $\sigma_{y_j}^2$ and $\rho_{y_j x_j}$ are such that they "balance out" so as to satisfy the homoscedasticity assumption. Moreover, when population regression slopes *actually*

differ, the assumption is likely to be violated (Overton, 2001). This is evident in the following expression, after substituting $\rho_{y_j x_j} = \beta_{y_j \cdot x_j} (\sigma_{x_j} / \sigma_{y_j})$ into Equation 9:

$$\sigma_{e_j}^2 = \sigma_{y_j}^2 - (\beta_{y_j \cdot x_j})^2 \sigma_{x_j}^2 \quad (10)$$

where $\beta_{y_j \cdot x_j}$ and $\sigma_{x_j}^2$, respectively, are the regression coefficient and variance of *x* in the *j*th group. Thus, when slopes are unequal, $\sigma_{y_j}^2$ and $\sigma_{x_j}^2$ in each group must have va**lue**s that offset one another so as to allow $\sigma_{e_1}^2 = \ldots = \sigma_{e_k}^2$. In short, when testing for the equality of regression slopes, violating the homoscedasticity assumption is not uncommon (Aguinis, 2004; Aguinis et al., 1999; Alexander & DeShon, 1994; DeShon & Alexander, 1996; Luh & Guo, 2002; Overton, 2001; Wilcox, 1997).

Violating the homoscedasticity assumption has biasing effects on Type I error rates and statistical power in OLS regression analyses when n_js are equal and unequal. Although with equal n_js and equal $\sigma_{x_j}^2$ across groups, some argue that Type I error rates perform "acceptably well" (Dretzke, Levin, & Serlin, 1982, p. 376), when equal $\sigma_{x_j}^2$ across groups is untenable, Type I error rates become conservative which can reduce the power of F_{OLS} (DeShon & Alexander, 1996). However, power does not suffer greatly when n_js are equal and error variances do not differ considerably (Alexander & DeShon, 1994).

With heteroscedasticity and unequal n_js, however, the effects are much more severe. Type I error rates and statistical power "can be either gross underestimates or severe overestimates depending on the pattern of sample sizes relative to the pattern of error variances" (DeShon & Alexander, 1996, p. 270). More precisely, when the larger $\sigma_{e_j}^2$ is paired with the larger n_j (direct pairing), F_{OLS} tests become conservative. This results in actual Type I error rates

less than the nominal level and, ceteris paribus, power is decreased. Conversely, when the larger $\sigma_{e_j}^2$ is paired with the smaller n_j (indirect pairing), F_{OLS} tests become liberal. This results in actual Type I error rates greater than the nominal level and, ceteris paribus, power is increased (albeit illegitimately) (see e.g., DeShon & Alexander, 1996, p. 265; Overton, 2001, p. 227).

Consistent with the above-described simulation results, Box (1954) demonstrated mathematically that with unequal n_js and unequal variances, under the null hypothesis, the F_{OLS} test is unstable. Recall that a central F random variable is the ratio of two independent central χ^2 variables each divided by their respective means or, equivalently, the ratio of two independent mean squares, each estimating σ^2. Box showed that with large variances paired with large n_js, the expectation of the numerator mean squares will be less than the expectation of the denominator mean squares, resulting in the approximate expected F ratio being shifted to < 1. Therefore, actual Type I error rates would be less than the nominal level (see e.g., rows 5, 9, and 13 of Table 4 in Box, 1954). Conversely, with large variances paired with small n_js, the expectation of the numerator mean squares will be greater than the expectation of the denominator mean squares, resulting in the approximate expected F ratio being shifted to > 1. Therefore, actual Type I error rates would be greater than the nominal level (see e.g., rows 4, 8, and 12 of Table 4 in Box, 1954). Stated differently, with heteroscedasticity, the diagonal elements of the $(p + 1) \times (p + 1)$ covariance matrix among the regression coefficients, i.e., $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$, will generally be too large or too small (viz., inefficient). Thus, hypothesis tests will be based on biased standard errors, and the estimate of σ^2 in Equation 5 will be incorrect due to the different error variances. Hence, previous simulation and mathematical results, together, provide cogent evidence to show that F_{OLS} performs poorly when the homoscedasticity assumption is violated, especially with unequal n_js.

To exacerbate matters, unequal n_js are quite common in psychological research for a number of reasons, particularly in nonexperimental and quasi-experimental studies. One reason is that attrition may result in unbalanced data (Shadish, Cook, & Campbell, 2002), for example in clinical studies comparing alternative interventions (Vanable, Carey, Carey, & Maisto, 2002). Another reason is that the population from which a researcher purposively samples could be disproportionate across subpopulations of the characteristic of interest (e.g., race) (Shadish et al.). This commonly occurs in the validation of personnel selection instruments (see e.g., Hattrup & Schmitt, 1990; Hunter, Schmidt, & Hunter, 1979). In addition, in longitudinal studies or in the analysis of archival data, missing values can lead to unequal n_js across variables of interest (Schafer & Graham, 2002).

This can have considerable implications on both theory development and practice in psychology. For example, consider a model with various causal links and interaction effects among variables. When Type II error rates are inflated due to the influence of heteroscedasticity in empirical studies that test for interactions, the inferences derived from the results of OLS regression analyses would lead researchers to conclude that the interaction effects in the model are untenable. However, the failure to detect such effects in the model would be illusory, that is, due to the influence of heteroscedasticity. This situation seems plausible considering that, for decades, researchers have underscored the problem of failing to detect hypothesized interactions using OLS regression (Aiken & West, 1991; McClelland & Judd, 1993; Stone-Romero & Liakhovitski, 2002; Zedeck, 1971).

On the other hand, tests of interactions in the just-noted model could lead to contradictory findings. Although, across studies, such findings could be due to a number of factors, including the unreliability of measures or heterogeneity of units (Shadish et al., 2002). However, ceteris

paribus, tests of the model could provide support for it in some instances (due to inflated Type I error rates) and lack of support for it in other instances (due to inflated Type II error rates). Moreover, the model could spawn additional flawed theoretical frameworks or cause practitioners to render erroneous decisions based on the model. This would have serious implications in such settings as assessing (a) the effectiveness of alternative therapies, or (b) whether a personnel selection instrument adversely affects an ethnic minority group.

To provide an illustration, Aguinis et al. (1999) demonstrated how substantive conclusions from two independent studies published in the *Journal of Applied Psychology* changed (likely due to the biasing effects of heteroscedasticity). Specifically, Aguinis et al. recognized that the homoscedasticity assumption was violated in these studies and reanalyzed the data using two statistical approximations (to be discussed below). From one study, the published finding which reported a *statistically significant interaction* was, in fact, not statistically significant when the more appropriate statistical approximations were used. Aguinis et al. concluded that the reported interaction was likely a Type I error because an inspection of the subgroup descriptive statistics indicated that the larger $\hat{\sigma}_{e_j}^2$ was paired with the smaller n_j. From another study, the published finding which reported *no statistically significant interaction* was, in fact, statistically significant when the more appropriate statistical approximations were used. Aguinis et al. concluded that the researchers likely failed to detect the hypothesized interaction because of low statistical power. That is, an inspection of the subgroup descriptive statistics indicated that the larger $\hat{\sigma}_{e_j}^2$ was paired with the larger n_j.

Hence, substantive conclusions can change when the homoscedasticity assumption is violated. Clearly, research in psychology and the accumulation of knowledge can be negatively affected by studies that violate this assumption.

Proposed Remedies When Homoscedasticity Is Violated

In an effort to mitigate the problems caused by heteroscedasticity in models that test for the equality of regression slopes, numerous data-analytic strategies have been investigated. Although some nonparametric and distribution-free alternatives are available, they tend to still perform sub-optimally relative to parametric approaches (DeShon & Alexander, 1996). Thus, these techniques are not discussed here. Alternatively, some examples of parametric approaches include the Welch-Aspin F approximation (F^*; Aspin, 1948; Welch, 1938), a generalization of James' (1951) second-order approximation (J; DeShon & Alexander, 1994), and the normalized t approximation (A; Alexander & Govern, 1994). In addition, Luh and Guo (2002) recently applied Johnson's (1978) transformations to the A approximation and Welch's (1951) approximation. Researchers have also advocated that statistical tests and confidence intervals be based on a heteroscedasticity consistent covariance matrix (HCCM; Long & Ervin, 2000; White, 1980). As another alternative, many statistical packages can fit linear (or nonlinear) mixed models which allow for the modeling of correlated and/or nonconstant errors (Littell, Milliken, Stroup, & Wolfinger, 1996; Pinheiro & Bates, 2000). Finally, Overton (2001) recommended weighted least squares (WLS) regression.

Of the above-noted procedures, the F^*, J, and A approximations and those recommended by Luh and Guo (2002) do not permit post hoc probing of statistically significant interactions. In contrast, HCCMs, mixed models, and WLS regression do permit post hoc analyses. These procedures are next discussed.

General Findings on Remedial Procedures: Post Hoc Analyses Not Permitted

Simulation research on the utility of the various approximation procedures has led to a number of conclusions. As expected, across various manipulated conditions, the F^*, J, and A

approximations result in more stable performance than the standard F_{OLS} (DeShon & Alexander, 1994, 1996). For small *N*s, the *J* approximation slightly outperforms the F^* and *A* approximations. However, because the Type I and Type II error rates of the *J* and *A* approximations are nearly identical and the *A* approximation is easier to compute, DeShon and Alexander (1996) recommended the *A* approximation for general use. Finally, when Luh and Guo (2002) applied Johnson's (1978) transformations in conjunction with two approximate methods, the approximations were more robust than without the transformations.

These alternative procedures are computationally intensive and are *not* incorporated into standard statistical software (Overton, 2001).[2] Furthermore, as Overton noted, if the F^*, *J*, or *A* approximations indicate that the *k* population regression slopes are not all equal, further probing of a significant interaction is precluded. Stated differently, these procedures *do not permit post hoc analyses* (e.g., tests of simple slopes, joint confidence regions). This is unfortunate because such analyses play an important role in understanding the nature of an interaction (Aiken & West, 1991; Bauer & Curran, 2005).

General Findings on Remedial Procedures: Post Hoc Analyses Permitted

In the following sections, HCCMs, mixed models, and WLS regression are discussed.

HCCM. Four HCCMs have been recommended and, in the statistics and econometrics literature, they are referred to as HC0, HC1, HC2, and HC3 (Long & Ervin, 2000; MacKinnon & White, 1985). It has been argued that a "HCCM allows a researcher to easily avoid the adverse effects of heteroscedasticity even when nothing is known about the form of heteroscedasticity" (Long & Ervin, 2000, p. 217). Because the OLS parameter estimates are unbiased, these would still be used in hypothesis testing and when computing confidence intervals. However, the

HCCM would provide the appropriate squared standard errors and covariances among the estimated regression coefficients.

Eicker (1963), Huber (1967), and White (1980) provided evidence suggesting that a HCCM provides asymptotically correct statistical inference on regression coefficients. However, HC0 (White, 1980) has been shown to perform very poorly when N is small (Long & Ervin, 2000; Mackinnon & White, 1985). When N is large (e.g., ≥ 500), there is little difference in the performance among the four HCCMs (Long & Ervin, 2000). Compared to other HCCMs, however, HC3 performed well even when N was small (e.g., 25) and when testing coefficients that were greatly affected by heteroscedasticity. Consequently, it was recommended for general use (Long & Ervin, 2000). HC3 can be denoted by:

$$\text{HC3} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\text{diag}\left[\frac{e_i^2}{\left(1-h_{ii}\right)^2}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{11}$$

where h_{ii} is the ith leverage value found on the diagonal of the projection matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ (Neter et al., 1996). Thus, with the estimated squared standard errors and covariances among the regression coefficients, respectively, on the diagonal and off-diagonal of HC3, testing regression coefficients, constructing confidence intervals, and performing post hoc analyses can be conducted as usual.

It deserves stressing that research has found some conditions where HC3 did not perform well (e.g., when $p = 1$ and small N, Wilcox, 2001). Although previous research has evaluated the performance of HC3 using different forms of heteroscedasticity, only the main effects of continuous predictors were part of the model (Long & Ervin, 2000). The inclusion of HC3 in the present paper is unique because (a) the model considered here includes a categorical predictor and the interaction between it and a continuous predictor, and (b) it appears that the effect of

unbalanced data on the performance of HC3 has never been investigated. Therefore, it would be useful to compare the performance of the *F* test for the equality of regression slopes based on HC3 (hereinafter, referred to as F_{HC3}) to other procedures.

Mixed Models. Linear mixed models provide another data-analytic alternative when heteroscedasticity exists (Snijders & Bosker, 1999). Such models are referred to as mixed when both fixed and random effects are included in the model (see McCulloch & Searle, 2001; Pinheiro & Bates, 2000; Searle, Casella, & McCulloch, 1992). With no random effects (**u**) and no corresponding random effects model matrix (**Z**), the linear mixed model ($\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$) reduces to Equation 2 (Searle et al., 1992, pp. 138-139). Examples of applications of mixed models include multilevel modeling (Raudenbush & Bryk, 2002; Singer, 1998; Snijders & Bosker, 1999), growth curve modeling (Singer, 1998; Singer & Willett, 2003), estimation of intraclass correlation coefficients (Bliese, 2000; Shrout & Fleiss, 1979), and generalizability theory (Brennan, 2001; Cronbach, 1972; DeShon, 2002).

As noted above, with heteroscedasticity, cov(**e**) no longer has the form shown in Equation 7. Rather, the diagonal elements differ depending on the level of the categorical predictor. Mixed models can be used in such instances. That is, in fitting a mixed model, researchers can specify that the error structure has different variances for each level of a categorical variable (Littell et al., 1996; Pinheiro & Bates, 2000).

Although several parameter estimation methods have been suggested (e.g., ANOVA), maximum likelihood (ML) and restricted maximum likelihood (RML) (a) are used in many mixed modeling procedures (e.g., S-PLUS, SAS, SPSS), (b) have asymptotically optimal properties (Casella & Berger, 2002), (c) in contrast to ANOVA estimators, result in estimated variance components (e.g., $\hat{\sigma}^2$) that can never be negative (Searle et al., 1992), and (d) in

contrast to ANOVA estimators, can be used with unbalanced data. ML estimation does not account for the *df* used in estimating fixed effects (Searle et al., 1992). "By failing to allocate some degrees of freedom to the estimation of fixed effects, . . . [ML] overstates the degrees of freedom left for estimating variance components and underestimates the variance components themselves, leading to biased estimates when samples are small" (Singer & Willett, 2003, p. 88). In contrast, RML estimation accounts for the *df* used in estimating fixed effects and, when the data are balanced, RML "solutions are identical to ANOVA estimators" (Searle et al., 1992, p. 255). According to Kreft and de Leeuw (1998), it is unclear which estimation method is preferred.

ML and RML estimation, unlike OLS (or WLS), *require* that the underlying probability distribution for the data follow a normal distribution (Searle et al., 1992). Stated differently, although ML and RML estimators are derived assuming normality, the OLS estimator in Equation 3 is derived *without the normality assumption* (Searle et al., 1992). As noted above, however, for statistical inference (e.g., hypothesis tests, confidence intervals), the normality assumption is invoked.

In mixed models, statistical inference involving fixed effects can be performed using likelihood ratio tests or *F* tests (Searle et al., 1992). However, because likelihood ratio tests based on nested models with different fixed effects terms can only be defined when fitted by ML estimation (Snijders & Bosker, 1999) and because simulations suggest that such tests perform poorly for fixed effects (Pinheiro & Bates, 2000, pp. 88-91), it has been recommended that statistical tests and estimated confidence intervals involving fixed effects be based on the *F* test (Pinheiro & Bates, 2000). Moreover, the *F* test is defined regardless of ML (hereinafter, referred to as F_{ML}) or RML (hereinafter, referred to as F_{RML}) estimation.

WLS Regression. For heteroscedasticity in models like that of Equation 1 (viz., $k = 2$), Overton (2001) recommended WLS regression as another practical solution for a number of reasons. First, it performed comparably to the F^*, J, and A approximations in terms of Type I error and power. Second, it was computationally simple. Third, it can be readily conducted in commonly available statistical software. Fourth, post hoc analyses (see e.g., Aiken & West, 1991; Rogosa, 1980) can be conducted as usual. Hereinafter, the F test based on Overton's approach is referred to as $F_{WLS(O)}$.

In summary, when heteroscedasticity exists, there are several important reasons for considering HCCMs, mixed models (using ML or RML), and WLS regression. First, compared to statistical approximations (e.g., A or J), HCCMs and WLS regression are relatively easy to compute (cf. Alexander & Govern, 1994, pp. 93-94; DeShon & Alexander, 1994, pp. 330-331). Second, compared to statistical approximations, HCCMs, mixed models, and WLS regression are available in many major statistical programs (either by default or as part of an add-on library, e.g., S-PLUS, SAS, SYSTAT). Third, perhaps most importantly, researchers can perform post hoc analyses with these procedures. However, it is very important to note that Overton (2001) *cautioned against* the use of the WLS approach for $k > 2$ because Type I error rates were inflated.

In the following section, WLS regression is described and I explicate how Overton's (2001) method can be extended to $k \geq 2$ based on an approach I developed.

WLS Regression as an Alternative Solution

In linear models with heteroscedasticity and when no variance-stabilizing transformation can be found, WLS regression is an often-recommended data-analytic technique (Carroll & Ruppert, 1988; Cohen et al., 2003; Cook & Weisberg, 1982, 1999; Draper & Smith, 1966; Fox,

1997; Greene, 2003; Mak, 1992; Neter et al., 1996).[3] According to Cohen et al. (2003), it "is the most commonly used remedial procedure for heteroscedasticity" (p. 146). The following paragraphs describe (a) WLS regression, (b) the WLS approach for testing the equality of two regression slopes, (c) a generalization of WLS regression for testing the equality of $k \geq 2$ regression slopes, and (d) alternative weight estimators in WLS regression.

WLS Regression

OLS and WLS regression minimize a similar function. Parameter estimation by OLS minimizes the following objective function (Q):

$$Q = \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi} \right)^2$$

with the OLS normal equations denoted by $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$. Solving the OLS normal equations for $\boldsymbol{\beta}$ results in the estimator in Equation 3. To estimate parameters using WLS, the following objective function is minimized:

$$Q^* = \sum_{i=1}^{N} w_i \left(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi} \right)^2$$

where w_i is a weight associated with the ith observation. The weights are rarely known and must be estimated (Neter et al., 1996). Note that when all weights are equal to unity, Q^* reduces to Q. That is, OLS is a special case of WLS (Greene, 2003; Neter et al., 1996; Rencher, 2000). The WLS normal equations are denoted by $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, where $\mathbf{V}$ is an $N \times N$ diagonal matrix with $v_{ii} > 0$.

When heteroscedasticity exists, OLS regression still results in a linear unbiased estimator, i.e., $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. However, the estimator is no longer the *best*; viz., it will not have minimum variance (see Gauss-Markov Theorem, Greene, 2003; Neter et al., 1996; Rencher, 2000; Schott,

1997). Stated another way, the OLS-based parameter estimates will no longer have the smallest standard errors among the class of methods that can be used to estimate the same parameters.

Alternatively, WLS regression is a transformation of the model matrix (**X**) and response vector (**y**) that not only accounts for heteroscedasticity, but also restores the minimum variance property of the parameter estimates (Cook & Weisberg, 1999; Draper & Smith, 1966; Greene, 2003; Neter et al., 1996; Rencher, 2000). Solving the WLS normal equations results in a WLS estimator for the parameters expressed as:

$$\hat{\boldsymbol{\beta}}^{*} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad (12)$$

with σ_i^2 (or more commonly, $\hat{\sigma}_i^2$) on the diagonal of **V** (Schott, 1997). Note that when errors are normally distributed, $\hat{\boldsymbol{\beta}}^{*}$ is also the maximum likelihood estimator (Rencher, 2000). With Equation 12, homoscedasticity obtains and $\hat{\boldsymbol{\beta}}^{*}$ has the optimal property of minimum variance. In other words, they are the most efficient unbiased estimators among all unbiased estimators that are linear functions of **y** (i.e., BLUE) (Rencher, 2000).

Statistical inferences (e.g., significance tests, confidence intervals) involving $\hat{\boldsymbol{\beta}}^{*}$ will provide more accurate Type I error rates and greater statistical power, compared to OLS regression. Noteworthy, estimates of population regression coefficients based on OLS or WLS will be identical.[4] However, when heteroscedasticity exists, WLS regression results in accurate standard errors compared to OLS regression. That is, by restoring homoscedasticity, WLS regression has smaller unbiased variances on the diagonal of $\text{cov}(\hat{\boldsymbol{\beta}}^{*}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ (Greene, 2003; Draper & Smith, 1966; Neter et al., 1996; Rencher, 2000). Consistent with this, the *F* test based on WLS can be viewed as a corrected *F* (see Moser & Lin, 1992).

The Two-Group WLS Approach

When heteroscedasticity exists, Overton (2001) recommended WLS regression for $k = 2$ and detailed how to estimate the weights for each group. One weight applied to each observation in Group 1; the second weight, to each observation in Group 2.

Initially, the residual variances for the *j*th group were computed as follows: (a) calculate SSE_j based on the OLS simple regression of y_j on x_j, and (b) divide SSE_j by $df_j = (n_j - 2)$. However, Overton (2001) found that the reciprocals of these estimated variances were biased and should not be used for weight estimation, arguing that they were biased by 2 *df*. Thus, it was concluded that the error variances ($_{\text{wls}}\sigma_{e_j}^{\ 2}$) should be estimated by dividing SSE_j by an adjusted df_j^* to account for the bias, namely, $df_j^* = (n_j - 2 - 2) = (n_j - 4)$. Using df_j^* in the estimation of $_{\text{wls}}\sigma_{e_1}^{\ 2}$ and $_{\text{wls}}\sigma_{e_2}^{\ 2}$ and taking their reciprocals, the weights in the two-group case were unbiased (Overton, 2001).

In summary, the two-group WLS approach is simple to conduct. The weights described by Overton (2001) ($_{\text{O}}w_j$) can be calculated by the following expression:

$$_{\text{O}}w_j = (_{\text{wls}}\hat{\sigma}_{e_j}^{\ 2})^{-1} = \frac{n_j - 4}{SSE_j} \tag{13}$$

where each observation in the *j*th group is assigned $_{\text{O}}w_j$. Specifically, the n_1 and n_2 weights equal $_{\text{O}}w_1$ and $_{\text{O}}w_2$, respectively. These *N* weights can be entered into a statistical software package and specified as a weight variable when fitting a general linear model. Equivalently, in terms of $\mathbf{V}^{-1}$ in Equation 12 and assuming the observations are ordered by group membership, the first n_1 diagonal elements would equal $_{\text{O}}w_1$ and the remaining n_2 diagonal elements would equal $_{\text{O}}w_2$. Based on Overton's research using $_{\text{O}}w_j$ in WLS, $F_{WLS(O)}$ performed well with α at

the nominal level and statistical power was comparable to more computationally intensive alternatives.

Overton's (2001) weight estimator in Equation 13 was limited to $k = 2$. In the following section, a generalization is delineated.

A Generalization of WLS Regression for $k \geq 2$

As with any linear model, an estimate of σ^2 is required for hypothesis testing and for constructing confidence intervals. Note that Equation 5 provides an unbiased estimator of σ^2 if homoscedasticity is satisfied. That is, $\hat{\sigma}^2 = SSE / df_E$, where df_E = (number of observations – number of estimated model parameters) (Maxwell & Delaney, 2000; Neter et al., 1996). For example, in the one-way ANOVA described above in the section titled *A Clarification of Homoscedasticity*, the $df_E = (N - 3)$ because there were three estimated model parameters.

Note that Overton (2001) fitted the linear model in Equation 1. Because four parameters were estimated in the full model (viz., β_0, β_1, β_2, and β_3), this indicates that df_E necessarily equals $(N - 4)$. For the purposes of the two-group WLS approach with n_j replications in group j, the degrees of freedom for the jth group (df_{E_j}) equaled $(n_j - 4)$ which led to ${}_{\text{wls}}\hat{\sigma}_{e_1}{}^2 = SSE_1 / (n_1 - 4)$ and ${}_{\text{wls}}\hat{\sigma}_{e_2}{}^2 = SSE_2 / (n_2 - 4)$ whose reciprocals follow Equation 13. However, a more general approach for estimating weights can be used.

Instead of estimating errors separately within each group, they can be estimated using the model. For example, in a fitted model, the estimated errors from it are used in estimating the weights (see e.g., Cohen et al., 2003; Greene, 2003; Neter et al., 1996). Note that "the magnitudes of σ_i^2 . . . often vary in a regular fashion with one or several predictor variables" (Neter et al., 1996, p. 403). When heteroscedasticity exists across groups, the *residuals for*

observations within a group provide the best estimates of error for that group. Therefore, by computing the variances of the residuals within groups and taking the reciprocal, these serve as reasonable weights for each group. Specifically, the sum of the squared residuals in the *j*th group is the SSE_j.

With n_j observations in each group, a nontrivial issue is that of how to reasonably estimate the variances of the residuals within each group whose reciprocals can serve as weights. Some researchers use the average of the SSE_j in the *j*th group and take the reciprocal to serve as the weight for observations in the respective group. However, research has demonstrated that using n_j performs poorly even in instances where the average is taken at each unique *x* (Carroll & Cline, 1988). It has also been recommended that SSE_j be divided by $(n_j - 1)$ and use the reciprocal as the weight for observations in the *j*th group (Carroll & Cline, 1988). This seems reasonable considering that an unbiased estimate of the variance of a single variable uses $(N - 1)$ as the denominator because only one parameter (i.e., β_0) is estimated. However, it merits noting that to estimate the errors, $q = (p + 1)$ parameters were estimated in the model (i.e., $\beta_0, \beta_1, \ldots, \beta_p$), using up *q df*. Therefore, an unbiased estimate of the error variance in the *j*th group should reduce the denominator of the estimated variance. Namely, the variance in each of the *k* groups can be estimated using the following expression:

$$_{\text{wls}}\hat{\sigma}_{e_j}{}^2 = \frac{SSE_j}{n_j - q} \tag{14}$$

For example, in the full linear model when testing for the equality of regression slopes with one continuous predictor and $k = 4$, there is an intercept and slope estimated for each group. Therefore, the denominator in Equation 14 would be $df_{E_j} = (n_j - 8)$. Note that when $k = 2$ with

an intercept and slope estimated for each group, this equation reduces to Overton's (2001) variance estimator for the two-group WLS approach.

In summary, with estimates of ${}_{\text{wls}}\sigma_{e_j}{}^2$ for each of the *k* groups, the weights are the reciprocals of the variances in Equation 14. In the general WLS regression model for testing the equality of regression slopes, the weights (w_j^*) are:

$$w_j^* = ({}_{\text{wls}}\hat{\sigma}_{e_j}{}^2)^{-1} = \frac{n_j - q}{SSE_j} \tag{15}$$

where each observation in the *j*th group is assigned w_j^*. Using commonly available statistical software packages, the *N* weights can be specified as a weighting variable in a general linear model procedure. Using w_j^* in WLS, the *F* test can be conducted as usual (hereinafter, referred to as F_{WLS*}).

Note that OLS regression will result in a larger estimated σ^2 compared to WLS regression. Stated differently, using WLS regression results in a smaller residual variance (approximately unity due to the transformation of the *N* observations) than OLS regression (Cook & Weisberg, 1999, p. 208; Neter et al., 1996, p. 409). This is mathematically expected because when heteroscedasticity exists, WLS restores the optimality property of the least squares estimator. Recall that with heteroscedasticity in OLS regression, the test statistic in Equation 6, under the null hypothesis, is no longer distributed as a central *F* random variable with a ratio of expectations equal to 1. As a consequence, with heteroscedasticity, the F_{OLS} test performs poorly (Box, 1954; DeShon & Alexander, 1996; Overton, 2001). Using WLS regression with estimated weights, on the other hand, results in a test statistic that *is* asymptotically distributed as a central *F* random variable with a ratio of expectations approximately equal to 1 when the null is true, providing a more robust test for the equality of regression slopes when heteroscedasticity exists.

Alternative Weighting Methods

Researchers have recommended other weighting methods. As noted above, the reciprocal of the average of the estimated errors (n_j / SSE_j) is known to perform poorly (Carroll & Cline, 1988). Therefore, it is not considered further. Another is based on sample-variances of estimated errors which can be expressed as (Carroll & Cline, 1988):

$$ {}_1w_j = \frac{n_j - 1}{SSE_j} \tag{16} $$

Hereinafter, the *F* test for the equality of regression slopes using ${}_1w_j$ in WLS is referred to as $F_{WLS(1)}$. Because the groups are independent and there is an intercept and slope estimated for each group, it would seem reasonable that $df_{E_j} = (n_j - 2)$ which can be expressed as (see e.g., Overton, 2001, p. 221):

$$ {}_2w_j = \frac{n_j - 2}{SSE_j} \tag{17} $$

Hereinafter, the *F* test for the equality of regression slopes using ${}_2w_j$ in WLS is referred to as $F_{WLS(2)}$. It is important to note that Equations 16 and 17 fail to consider that *q* parameters were estimated to compute the fitted values. These estimated fitted values were then used to calculate the squared estimated errors $\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_p x_{pi}\right)^2$ on which the weighting methods rely. Furthermore, it is important to realize that weight estimators may not perform optimally as sample size decreases. As is well known in the literature on linear regression, it is preferable to have small *q* and large *N*. Consistent with this, Bement and Williams (1969) stated that "the most practical conclusion drawn from the results . . . [of their study] . . . is that each estimated weight should be based on at least 10 *df*" (p. 1369).

The Present Study

As noted above, the main purpose of the present study was to describe a proposed extension of Overton's (2001) method and compare its performance (specifically, Type I error rates and statistical power) to alternative procedures under various conditions (including that of heteroscedasticity). These alternatives are computationally simple, available in many common statistical software packages, and permit post hoc analyses of an interaction. To compare the performance of the various methods, a Monte Carlo simulation was conducted. Overall, I expected:

(a) F_{RML} to control Type I error rates better than F_{ML} and F_{OLS};

(b) the WLS methods to control Type I error rates better than F_{OLS}. More specifically, I expected F_{WLS*} to maintain control of Type I error rates better than $F_{WLS(1)}$, $F_{WLS(2)}$, and $F_{WLS(O)}$;

(c) all methods to be more powerful than F_{OLS}.

Note that I did not have any specific prediction of how F_{HC3} would perform because it has not been investigated under the simulated conditions considered below.

CHAPTER TWO: METHOD

Description of Simulation

Overview

Using the *S* language in S-PLUS (Chambers, 1998; S-PLUS, 2002), a Monte Carlo simulation was conducted to compare the Type I error rates and statistical power of F_{OLS}, F_{HC3}, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*}. In addition to manipulating the degree of heteroscedasticity (using Equations 9 and 10), the performance of these procedures was compared at different values of k, N, and n_j/N. The nominal α was set at .05.

Manipulated Parameters

Number of groups. To demonstrate that WLS can be accurately applied to tests for equality of regression slopes for more than two independent groups, k was equal to 3 and 4.

Total sample size. Seven *N*s were used in the present study, ranging from 48 to 336. This range overlaps with *N*s considered by previous research on two groups (60 and 300, Aguinis & Stone-Romero, 1997; 140 to 430, DeShon & Alexander, 1996; 30 to 200, Dretzke et al., 1982; 40 to 360, Overton, 2001). In addition, this range brackets *N*s typically found in validation studies (Lent, Aurbach, & Levin, 1971; Salgado, 1998).

Proportion within groups. The proportion of observations within groups (${}_kP_j = n_j/N$) assumed three levels—equal, moderately unequal (i.e., for $k = 3$, ${}_3P_j$ s = .50, .25, .25; for $k = 4$, ${}_4P_j$ s = .375, .208$\overline{3}$, .208$\overline{3}$, .208$\overline{3}$), and very unequal (i.e., for $k = 3$, ${}_3P_j$ s = .$\overline{6}$, .1$\overline{6}$, .1$\overline{6}$; for $k = 4$, ${}_4P_j$ s = .50, .1$\overline{6}$, .1$\overline{6}$, .1$\overline{6}$). Based on the manipulated *k*s, *N*s, and ${}_kP_j$s, this resulted in the subgroup sample sizes shown in Table 1. (Text continues on p. 29.)

Table 1

Subgroup Sample Sizes When k = 3 and 4

	$k = 3$			$k = 4$			
N	n_1	n_2	n_3	n_1	n_2	n_3	n_4
48	16	16	16	12	12	12	12
48	24	12	12	18	10	10	10
48	32	8	8	24	8	8	8
96	32	32	32	24	24	24	24
96	48	24	24	36	20	20	20
96	64	16	16	48	16	16	16
144	48	48	48	36	36	36	36
144	72	36	36	54	30	30	30
144	96	24	24	72	24	24	24
192	64	64	64	48	48	48	48
192	96	48	48	72	40	40	40
192	128	32	32	96	32	32	32
240	80	80	80	60	60	60	60
240	120	60	60	90	50	50	50
240	160	40	40	120	40	40	40
288	96	96	96	72	72	72	72
288	144	72	72	108	60	60	60
288	192	48	48	144	48	48	48

336	112	112	112	84	84	84	84
336	168	84	84	126	70	70	70
336	224	56	56	168	56	56	56

Note. N = total sample size. k = number of groups. n_1, n_2, n_3, and n_4 = subgroup sample sizes.

Heteroscedasticity. When $k = 3$ ($k = 4$), the ratio of error variances included homoscedasticity, 1:1:1 (1:1:1:1), and three levels of heteroscedasticity, 4:1:1 (4:1:1:1), 16:1:1 (16:1:1:1), and 64:1:1 (64:1:1:1). Consistent with Overton's (2001) argument, "this wide range from homogeneity . . . to extreme heterogeneity . . . was chosen because if [these procedures perform] . . . well under these conditions, . . . [they] . . . would be expected to perform well in most research settings" (p. 223). Note that the manipulated conditions included pairing the larger $\sigma_{e_j}^2$ with the (a) larger n_j (direct pairing), and (b) smaller n_j (indirect pairing).

Type I error and statistical power. For the conditions that assessed the empirical Type I error rates, the null hypothesis of equal population slopes across the k groups was true. Noteworthy, in Overton's (2001) simulation, σ_{y_j} was set equal to σ_{x_j}, so necessarily $\sigma_{y_j} / \sigma_{x_j} = 1$. Consequently, in Overton's study, the test for the equality of regression slopes was equivalent to the test for the equality of correlation coefficients. However, because these tests are not equivalent when the ratios of σ_{y_j} and σ_{x_j} are not the same across groups (Alexander & DeShon, 1994), in the present study, I also simulated conditions where the k population correlation coefficients differ yet population slopes were equal.[1] Bradley's (1978) liberal criterion [0.025, 0.075] for robustness was used.

Numerous alternative hypotheses can be tested, such as the population slopes for two groups are steeper than that of the remaining population slopes or all pairwise population differences between slopes are ≠ 0. For simplicity and without loss of generality, five different values of effect size were selected such that the population slope of a focal group (hereinafter, Group 2) was allowed to differ from the remaining groups (which assumed a common slope).

As noted by Aguinis, Beaty, Boik, and Pierce (2005), the effect size metric (f^2) to describe the strength of an interaction effect in multiple regression (Aiken & West, 1991; Cohen, 1988) is not appropriate when heteroscedasticity exists. Therefore, I used the modified f^2 (hereinafter, referred to as f^2) derived by Aguinis et al. (2005, p. 105). Because the calculations are complex, Aguinis and Pierce (in press) provide a computer program available at http://carbon.cudenver.edu/~haguinis/mmr/ which performs the required computations online.

The five levels of non-zero effect sizes were 0.002, 0.01, 0.02, 0.05, and 0.08. Although 0.002 is markedly lower than Cohen's (1988) convention for small interactions, a 30-year review of literature in applied psychology and management found 0.002 to be the median effect size in tests for the equality of regression slopes (Aguinis et al., 2005). I did not include effect sizes greater than 0.08 because, based on initial testing of the *S* code on a subset of simulated conditions, the statistical power of F_{OLS} and F_{WLS*} was near unity at $f^2 = 0.08$. Beyond this, there would be little opportunity to compare techniques.

Research Design Summary

Based on the manipulated parameters, overall, there were 1,764 conditions (i.e., 2 (k) × 7 (N) × 3 ($_kP_j$) × 7 (heteroscedasticity) × 6 (f^2)). However, 252 conditions were excluded because they cannot occur. To clarify, Table 2 provides a layout of a subset of the study's design. In it, the rows are the three levels of $_kP_j$s, labeled Equal, Moderately Unequal, and Very Unequal. At

the top of the same table is the heading Heteroscedasticity with various levels beneath it. For example, the heading Absent represents conditions where heteroscedasticity does not exist (i.e., homoscedasticity). All remaining columns fall under the heading Present; that is, conditions where some form of heteroscedasticity existed. The three columns under the heading No Pairing have the numbers 4, 16, and 64. These represent conditions of increasing heteroscedasticity (i.e., $\sigma_{e_j}^2$) when no pairing exists which can occur *only* when ${}_kP_j$s are *equal*. Thus, in the corresponding cells, there is a symbol "●" indicating that such a condition exists.

Table 2

Research Design Sub-table for the ${}_kP_j$ × Heteroscedasticity Conditions

	Heteroscedasticity								
	Absent	Present							
		No Pairing			Pairing				
					Direct			Indirect	
${}_kP_j$		4	16	64	4	16	64	4	16	64
Equal	●	●	●	●						
Moderately Unequal	●				●	●	●	●	●	●
Very Unequal	●				●	●	●	●	●	●

Note. ${}_kP_j$ = degree of disproportionate subgroup sample sizes.

For example, when $k = 3$ and ${}_3P_j$s are *equal*, the condition where the larger $\sigma_{e_j}^2$ is in Group 1 (i.e., 16, 1, 1) is equivalent to the condition where the larger $\sigma_{e_j}^2$ is in Group 3 (i.e., 1, 1, 16). Therefore, when ${}_kP_j$s are equal, pairing is not an issue. The remaining columns are under the heading termed Pairing which has two levels (i.e., Direct and Indirect). For both, they also include increasing degrees of heteroscedasticity (i.e., 4, 16, and 64). It is very important to note, however, that direct and indirect pairing exist *only* when ${}_kP_j$s are *unequal*. For example, when a larger error variance is paired with a larger subgroup sample size, this is direct pairing. In the same table, there are 18 cells indicating conditions that can be simulated. With 18 × 2 (k) × 7 (N) × 6 (f^2) cells, in the present study, there were 1,512 simulated conditions.

Data Generation

For the jth group, n_j pairs (y and x) of independent normal random numbers were generated from populations with means of zero and standard deviations of σ_{y_j} and σ_{x_j}, respectively. The n_j observations of bivariate normal data were generated using the S-PLUS function `rmvnorm` which permits either user-specified population covariances or correlation coefficients among variables. Using this functionality, $\rho_{y_j x_j}$ was set between y and x. The same procedure was used to generate the data for the remaining groups. The values of the manipulated variables were selected so as to result in various degrees of heteroscedasticity and f^2s in the population.[2]

On each simulated dataset, the equality of regression slopes was tested using eight F tests. These were F_{OLS}, F_{HC3}, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*}.[3] For each condition, there were 1,000 replications. The proportion of times the null hypothesis was rejected within a condition was recorded for all tests.

Analyses

The results of the Monte Carlo simulation produced 1,512 (conditions) × 8 (tests) = 12,096 proportions. The dependent variables were Type I error rates and statistical power. To provide a summary of the Type I error rates, OLS multiple regression was used. For these analyses, $\alpha = .05$. Note that continuous predictors (e.g., N) were standardized prior to computing interaction terms and dummy regressors were used to index categorical predictors (e.g., ${}_kP_j$) (see Aguinis, 2004; Cohen et al., 2003). Note that, for each of the eight tests, the functional form of how each of the predictors was related to power was not known nor would it be reasonable to assume that the same linear and interactive terms could be used to model the relations between power (of each test) and each of the predictors (or functions of them, e.g., ln (f^2), power terms). If separate analyses were conducted, this would have resulted in different regression analyses with potentially different predictors and functions of predictors in each equation. Because this would have detracted from the purpose of understanding which test was the most powerful and yet provided control of Type I error rates across various conditions, regression analyses were not conducted. However, various power tables and figures with power curves from numerous study conditions are presented.

Because a completely crossed design could not be used, resulting in 252 empty cells, some effects cannot be unambiguously computed (see e.g., the literature on aliased effects in confounded factorial designs and fractional factorial designs, Dean & Voss, 1999; Kirk, 1995). To simplify the analyses and interpretation, conditions involving homoscedasticity were not included in the regression analyses. Because it is not suggested that F_{OLS} be supplanted when the homoscedasticity assumption is met, I focused on understanding how the various procedures performed when heteroscedasticity exists.

When using Type I error rates as a dependent variable, the interpretation of the multiple regression analyses deserve note. Type I error rates across conditions should be controlled (i.e., a constant) at α (viz., .05 in the present study). If a particular method is *effective* at controlling Type I error rates, only an intercept would be needed in the regression equation (i.e., $\hat{y} = \hat{\beta}_0$), with $\hat{\beta}_0$ approximately equal to .05. In other words, the regression equation cannot explain any variance beyond that explained by the mean of the Type I error rates. Specifically, the *MSE* would simply be the variance of the Type I error rates and R^2 would be a function of the usual one-sample *t*-test and its associated *df*. Therefore, if a method is *ineffective* at controlling Type I error rates, variability in the rejection rates *increase* and more terms beyond $\hat{\beta}_0$ are needed in the regression equation, thus, explaining additional (albeit, undesirable) variance.

Because Type I error rates should be controlled at the nominal α regardless of the manipulated variables (e.g., ${}_kP_j$, heteroscedasticity) the marginal distributions of the Type I error rates for each of the tests should be symmetric about their mean (ideally, .05). Therefore, for Type I error rates, in addition to standard descriptive statistics, I present skewness statistics. Values with an absolute magnitude near zero are indicative of symmetry.

CHAPTER THREE: RESULTS

The eight *F* tests are next compared in terms of their empirical Type I error rates and statistical power.[1] Note that in presenting the study's findings, I refer to Tables 3 – 23 and Figures 1 – 20.

Type I Error Rates

In the following sections, the overall robustness of the tests under the null hypothesis is compared. Then, their performance with heteroscedasticity and equal and unequal ${}_kP_j$s is considered.

Overall Robustness

In evaluating the ability of the eight tests to control Type I error rates at the nominal α, Bradley's (1978) liberal criterion was used. Table 3 presents the results of applying this criterion to each of the tests across various conditions. Note that the column labeled ALL contains all conditions (i.e., both homoscedasticity and heteroscedasticity). The column labeled HO contains only conditions where homoscedasticity existed. These are provided for descriptive purposes. For example, with homoscedasticity, not surprisingly, 100% of the Type I error rates using F_{OLS} met the criterion. However, the WLS methods also performed well. For example, $F_{WLS(O)}$ and F_{WLS*}, respectively, had 100% and 97.56% of their Type I error rates satisfy the robustness criterion.

The last four columns, in the same table, are of particular importance. Across all the heteroscedasticity conditions (denoted by H), F_{OLS} had 17.62% of its Type I error rates satisfy the robustness criterion. F_{HC3} did very poorly; only 5.71% of its Type I error rates met the criterion. $F_{WLS(O)}$ and F_{WLS*} performed well with values of 95.24% and 98.53%, respectively.

(Text continues on p. 37.)

Table 3

Percentage of Empirical Type I Error Rates Satisfying Bradley's (1978) Liberal Criterion for Robustness Across Various Conditions

Test	ALL	HO	H	H(E)	H(D)	H(I)
F_{OLS}	31.35%	100.00%	17.62%	38.10%	22.62%	2.38%
F_{HC3}	6.35%	9.52%	5.71%	9.52%	4.76%	4.76%
F_{ML}	88.49%	90.48%	88.10%	95.24%	90.48%	82.14%
F_{RML}	91.67%	92.86%	91.43%	95.24%	95.24%	85.71%
$F_{WLS(1)}$	89.68%	90.48%	89.52%	95.24%	91.67%	84.52%
$F_{WLS(2)}$	91.67%	92.86%	91.43%	95.24%	95.24%	85.71%
$F_{WLS(O)}$	96.03%	100.00%	95.24%	95.24%	96.43%	94.04%
F_{WLS*}	98.37%	97.56%	98.53%	95.24%	98.77%	100.00%

Note. An *F* statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). ALL = all homoscedasticity and heteroscedasticity conditions. HO = homoscedasticity conditions only. H = all heteroscedasticity conditions only. H(E) = heteroscedasticity conditions with equal subgroup sample sizes only. H(D) = heteroscedasticity conditions with direct pairing only. H(I) = heteroscedasticity conditions with indirect pairing only. For all tests, the number of conditions on which the percentages were based for ALL, HO, H, H(E), H(D), and H(I) were 252, 42, 210, 42,

84, and 84, respectively, except for F_{WLS*} which was based on 245, 41, 204, 42, 81, and 81, respectively.

For heteroscedasticity conditions where ${}_kP_j$s were equal (denoted by H(E)), F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*} performed identically. Overall, when heteroscedasticity existed, F_{WLS*} tended to outperform all the methods in controlling Type I error rates. For example, with direct pairing (denoted by H(D)), F_{HC3} had a value of 4.76%. In comparison, F_{RML} performed well with 95.24% of its Type I error rates satisfying the criterion. However, F_{WLS*} performed better with a value of 98.77%.

In the following sections, Tables 4 and 5 present the results of multiple regression analyses for Type I error rates with equal and unequal ${}_kP_j$s, respectively.[2] In these, main effects, two-way interactions, and a single three-way interaction term are presented because all remaining higher-order terms explained very little additional variance in the dependent variable. Instead of referencing each statistically significant term in each of the regression analyses below, common relations across the tests are described as well as key differences among them.

Heteroscedasticity and Equal ${}_kP_j$s

In Table 4, for equal ${}_kP_j$s, it is quite clear that F_{OLS} ($R^2 = .73$, $p < .01$) and F_{HC3} ($R^2 = .85$, $p < .01$) were most affected by the manipulated variables. For example, F_{OLS} was affected by k ($\hat{\beta} = 0.007$) and heteroscedasticity ($\hat{\beta} = 0.011$). That is, as k or heteroscedasticity increased, Type I error rates increased. Most notably, it was the *only* method with Type I error rates that were affected by heteroscedasticity. Not surprisingly, N did not predict Type I error rates. Most of the alternative methods performed similarly with one exception—F_{HC3}.

(Text continues on p. 43.)

Table 4

Multiple Regression Analysis of Empirical Type I Error Rates when Testing for the Equality of Regression Slopes with Equal ${}_kP_j$s and Heteroscedasticity Exists

Predictor / R^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
K	**0.007****	**0.150****	**0.003****	**0.003****	**0.003****	**0.003****	**0.003****	**0.003****
	(4.692)	(4.703)	(2.913)	(2.872)	(2.913)	(2.913)	(2.913)	(2.913)
N	-0.001	**-0.396****	**-0.004****	**-0.004****	**-0.004****	**-0.004****	**-0.004****	**-0.004****
	(-0.852)	(-12.395)	(-3.428)	(-3.395)	(-3.428)	(-3.428)	(-3.428)	(-3.428)
$\sigma_{e_j}^2$	**0.011****	-0.001	0.000	0.000	0.000	0.000	0.000	0.000
	(7.638)	(-0.034)	(0.120)	(0.129)	(0.120)	(0.120)	(0.120)	(0.120)
$k \times N$	-0.000	**-0.073***	**-0.003***	**-0.003****	**-0.003****	**-0.003****	**-0.003****	**-0.003****
	(-0.280)	(-2.269)	(-2.440)	(-2.413)	(-2.440)	(-2.440)	(-2.440)	(-2.440)
$k \times \sigma_{e_j}^2$	**0.003***	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	(2.328)	(0.038)	(0.219)	(0.228)	(0.219)	(0.219)	(0.219)	(0.219)
$N \times \sigma_{e_j}^2$	-0.001	0.005	0.000	0.000	0.000	0.000	0.000	0.000
	(-0.626)	(0.153)	(0.374)	(0.367)	(0.374)	(0.374)	(0.374)	(0.374)
$k \times N \times \sigma_{e_j}^2$	-0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000
	(-0.100)	(0.147)	(0.055)	(0.050)	(0.055)	(0.055)	(0.055)	(0.055)
R^2	.73	.85	.44	.44	.44	.44	.44	.44
F (7, 34)	13.13**	26.52**	3.85**	3.76**	3.85**	3.85**	3.85**	3.85**

Note. An F statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}),

weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). k = number of groups. N = total sample size. ${}_kP_j$ = degree of disproportionate subgroup sample sizes. $\sigma_{e_j}{}^2$ = degree of heteroscedasticity. The values for each row of predictors are estimated regression coefficients with associated *t*-statistics in parentheses. To facilitate the interpretation of the eight analyses, statistically significant regression coefficients are boldfaced.

* $p < .05$. ** $p < .01$.

Table 5

Multiple Regression Analysis of Empirical Type I Error Rates when Testing for the Equality of Regression Slopes with Unequal ${}_kP_j$s and Heteroscedasticity Exists

Predictor / R^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
K	**0.011****	**0.164****	**0.004***	**0.003****	**0.003****	**0.003****	**0.003****	0.001
	(4.519)	(6.418)	(2.599)	(2.763)	(2.625)	(2.635)	(2.864)	(1.316)
N	0.002	**-0.419****	**-0.003***	**-0.003***	**-0.003***	**-0.003***	**-0.003****	-0.001
	(0.931)	(-16.398)	(-2.410)	(-2.238)	(-2.296)	(-2.406)	(-2.917)	(-1.154)
${}_kP_j$ (D_1)	**-0.019****	**0.162****	**0.007****	**0.005***	**0.006****	**0.005***	0.001	**-0.004***
	(-4.781)	(3.860)	(3.302)	(2.304)	(2.785)	(2.292)	(0.792)	(-2.071)
Pairing (D_2)	**0.099****	0.061	0.004	0.003	0.003	0.003	0.002	-0.002
	(25.722)	(1.447)	(1.706)	(1.425)	(1.535)	(1.405)	(1.129)	(-1.262)
$\sigma_{e_j}^2$	-0.000	-0.021	0.000	-0.000	-0.000	-0.000	-0.000	0.001
	(-0.037)	(-0.733)	(-0.014)	(-0.170)	(-0.074)	(-0.170)	(-0.194)	(0.727)
$k \times N$	0.001	**-0.038***	-0.001	**-0.002****	**-0.002***	**-0.002***	**-0.002****	-0.001
	(0.584)	(-2.572)	(-1.825)	(-2.871)	(-2.030)	(-2.578)	(-3.622)	(-1.930)
$k \times D_1$	**-0.012****	-0.045	-0.003	-0.002	-0.002	-0.002	-0.001	0.000
	(-4.511)	(-1.539)	(-1.619)	(-1.569)	(-1.449)	(-1.335)	(-0.933)	(0.335)
$k \times D_2$	**-0.020****	-0.038	**-0.003***	-0.002	-0.003	-0.002	-0.001	-0.002
	(-7.442)	(-1.299)	(-2.123)	(-1.587)	(-1.674)	(-1.370)	(-1.017)	(-1.354)
$k \times \sigma_{e_j}^2$	-0.001	0.006	0.000	0.000	-0.000	-0.000	0.000	-0.000
	(-0.400)	(0.374)	(0.247)	(0.118)	(-0.082)	(-0.040)	(-0.034)	(-0.123)

$N \times D_1$	-0.004	-0.036	**-0.008****	**-0.005****	**-0.007****	**-0.005****	-0.001	**0.003****
	(-1.395)	(-1.216)	(-4.979)	(-3.818)	(-4.376)	(-3.494)	(-0.613)	(2.666)
$N \times D_2$	**-0.006***	-0.033	**-0.007****	**-0.006****	**-0.006****	**-0.005****	**-0.004****	0.001
	(-2.054)	(-1.136)	(-4.375)	(-4.134)	(-4.056)	(-3.834)	(-3.077)	(0.833)
$N \times \sigma_{e_j}^2$	-0.001	-0.002	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(-0.972)	(-0.145)	(-0.110)	(-0.162)	(-0.156)	(-0.321)	(-0.070)	(-0.583)
$D_1 \times D_2$	**0.081****	0.099	**0.005***	0.004	**0.006***	0.005	0.002	0.002
	(14.809)	(1.658)	(1.672)	(1.610)	(2.026)	(1.854)	(0.893)	(0.585)
$D_1 \times \sigma_{e_j}^2$	-0.001	0.004	-0.001	-0.000	-0.001	-0.000	-0.001	-0.001
	(-0.215)	(0.091)	(-0.455)	(-0.184)	(-0.367)	(-0.183)	(-0.446)	(-0.523)
$D_2 \times \sigma_{e_j}^2$	**0.025****	0.041	0.001	0.001	0.001	0.001	0.001	-0.000
	(6.640)	(0.998)	(0.355)	(0.597)	(0.419)	(0.600)	(0.487)	(-0.456)
$D_1 \times D_2 \times \sigma_{e_j}^2$	**0.017****	-0.020	0.001	0.000	0.000	-0.000	0.001	0.001
	(3.158)	(-0.336)	(0.339)	(0.028)	(0.104)	(-0.137)	(0.393)	(0.288)
R^2	.96	.88	.67	.60	.63	.59	.45	.16
F (16, 151)[a]	229.80**	69.47**	18.90**	14.44**	16.36**	13.82**	7.81**	1.71

Note. An F statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). k = number of groups. N = total sample size. ${}_kP_j$ = degree of disproportionate subgroup sample sizes. D_1 = dummy variable indexing conditions with very unequal subgroup sample sizes. D_2 = dummy

variable indexing conditions with indirect pairing. $\sigma_{e_j}^2$ = degree of heteroscedasticity. The values for each row of predictors are estimated regression coefficients with associated *t*-statistics in parentheses. To facilitate the interpretation of the eight analyses, statistically significant regression coefficients are boldfaced.

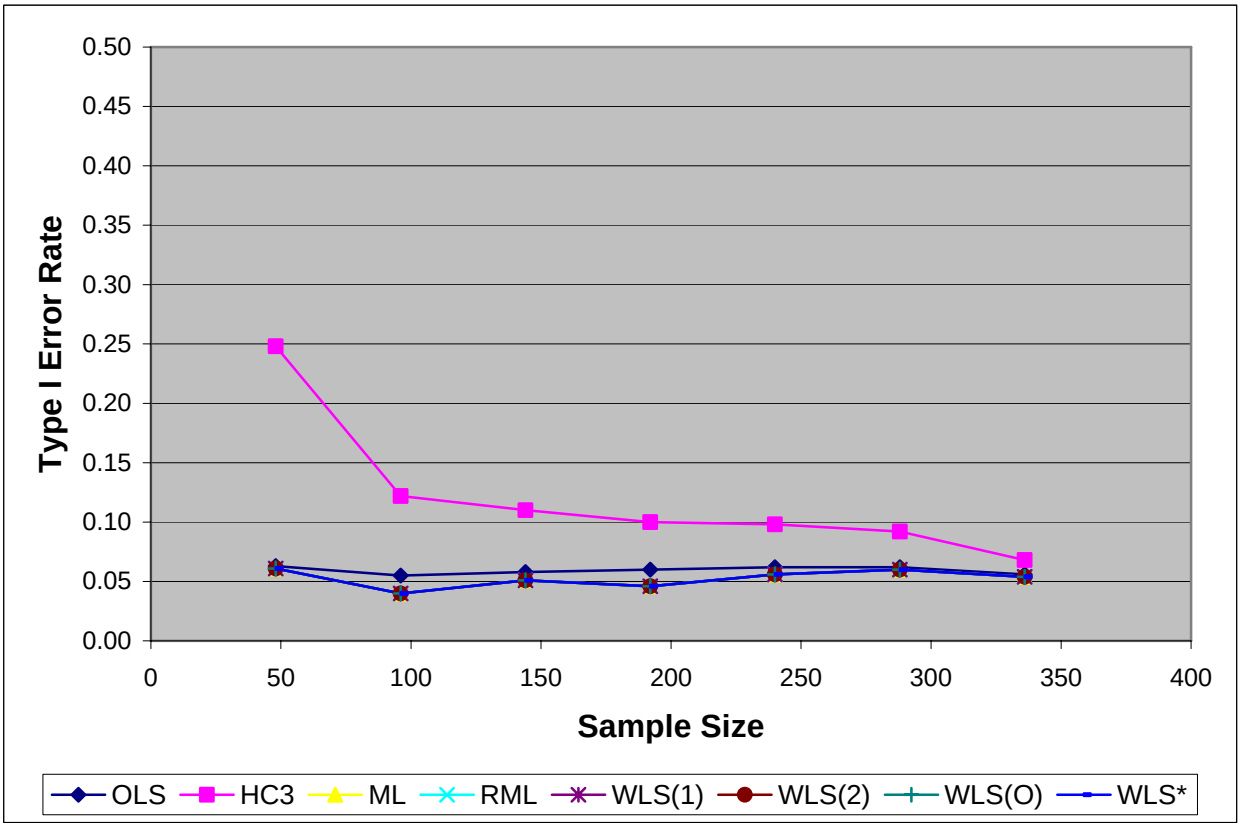
* $p < .05$. ** $p < .01$.

[a] F (16, 145) for the model predicting the Type I error rates based on F_{WLS*}.

Its Type I error rates were greatly affected by N ($\hat{\beta}$ = -0.396). More precisely, as N increased, Type I error rates decreased.

The performance of all the alternative methods had Type I error rates that were affected by the two-way interaction between k and N. That is, although Type I error rates decreased as N increased, this depended on the value of k. Holding N constant, as k increases, the alternative methods may not perform well. For example, in Figure 1, k = 3 and $\sigma_{e_j}^2$ = 4, 1, 1 (see Figure 1A), $\sigma_{e_j}^2$ = 16, 1, 1 (see Figure 1B), and $\sigma_{e_j}^2$ = 64, 1, 1 (see Figure 1C). Type I error rates for F_{OLS} were inflated, which increased with heteroscedasticity. In addition, consistent with the results from the regression analysis, F_{HC3} had greatly inflated Type I error rates which tended to converge towards the nominal α as N increased. Overall, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*} performed identically with empirical Type I error rates within Bradley's (1978) liberal criterion. Now consider similar conditions in Figure 2, but when k = 4 and $\sigma_{e_j}^2$ = 4, 1, 1, 1 (see Figure 2A), $\sigma_{e_j}^2$ = 16, 1, 1, 1 (see Figure 2B), and $\sigma_{e_j}^2$ = 64, 1, 1, 1 (see Figure 2C). Type I error rates for F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*} were slightly more inflated compared to when k = 3. However, these Type I error rates are still closer to the nominal α than those of F_{OLS} or F_{HC3}. (Text continues on p. 50.)

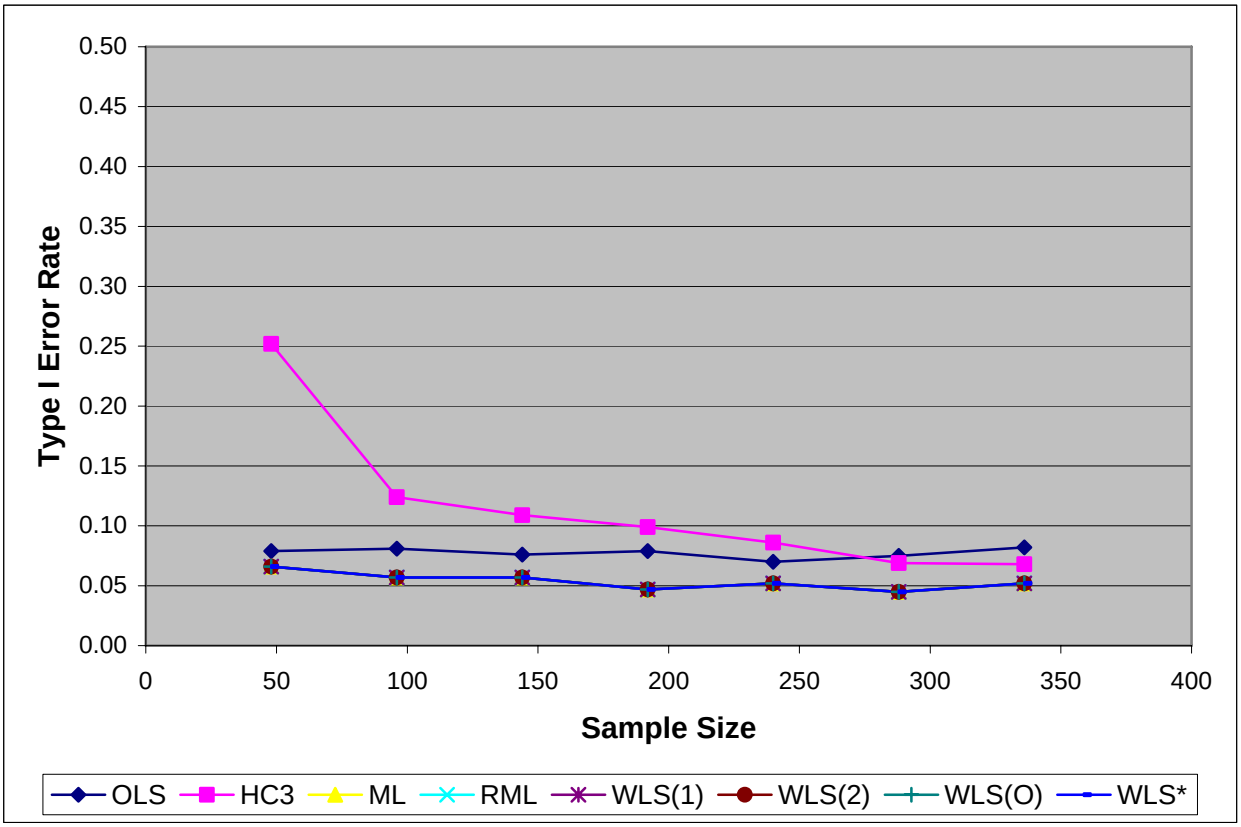
Panel A



Note. The Type I error rates for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

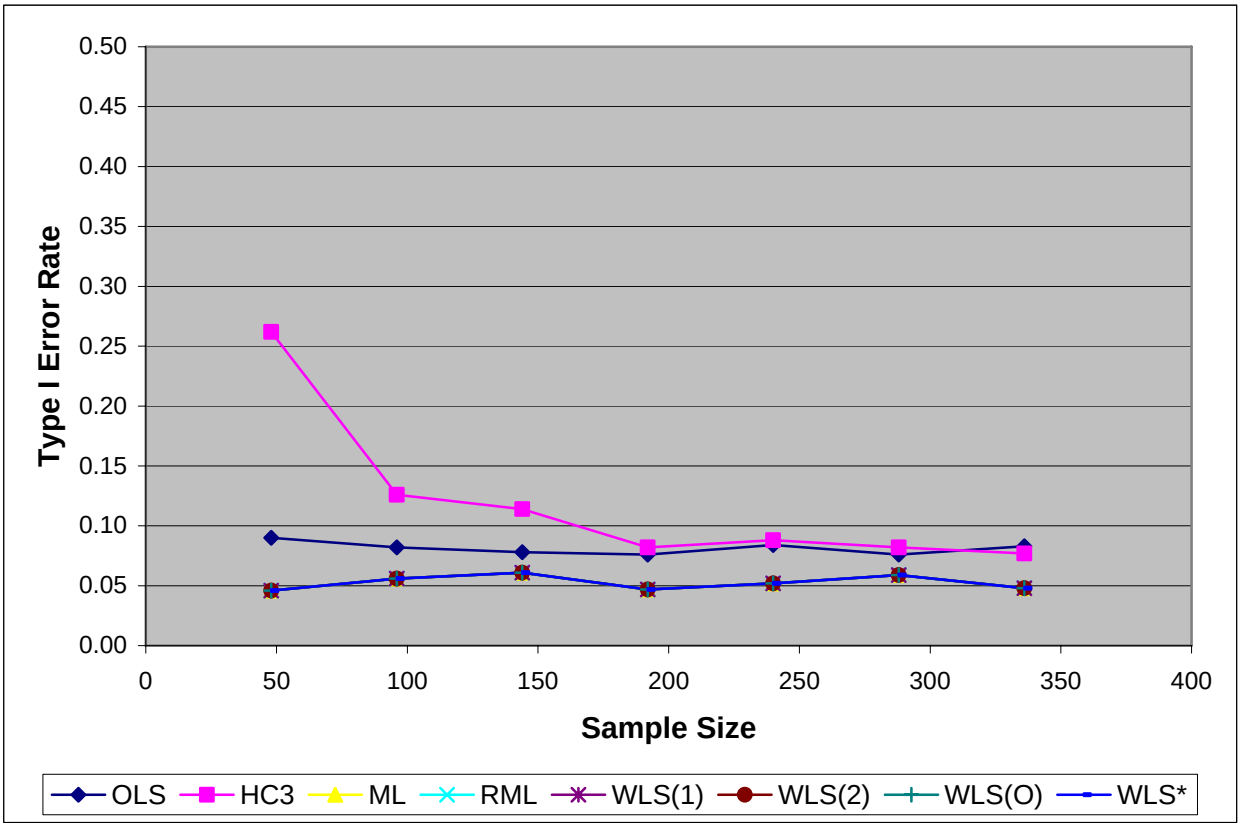
Figure 1. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, equal subgroup sample sizes, and (A) $\sigma_{e_j}^2$ s = 4, 1, 1, (B) $\sigma_{e_j}^2$ s = 16, 1, 1, and (C) $\sigma_{e_j}^2$ s = 64, 1, 1.

Panel B



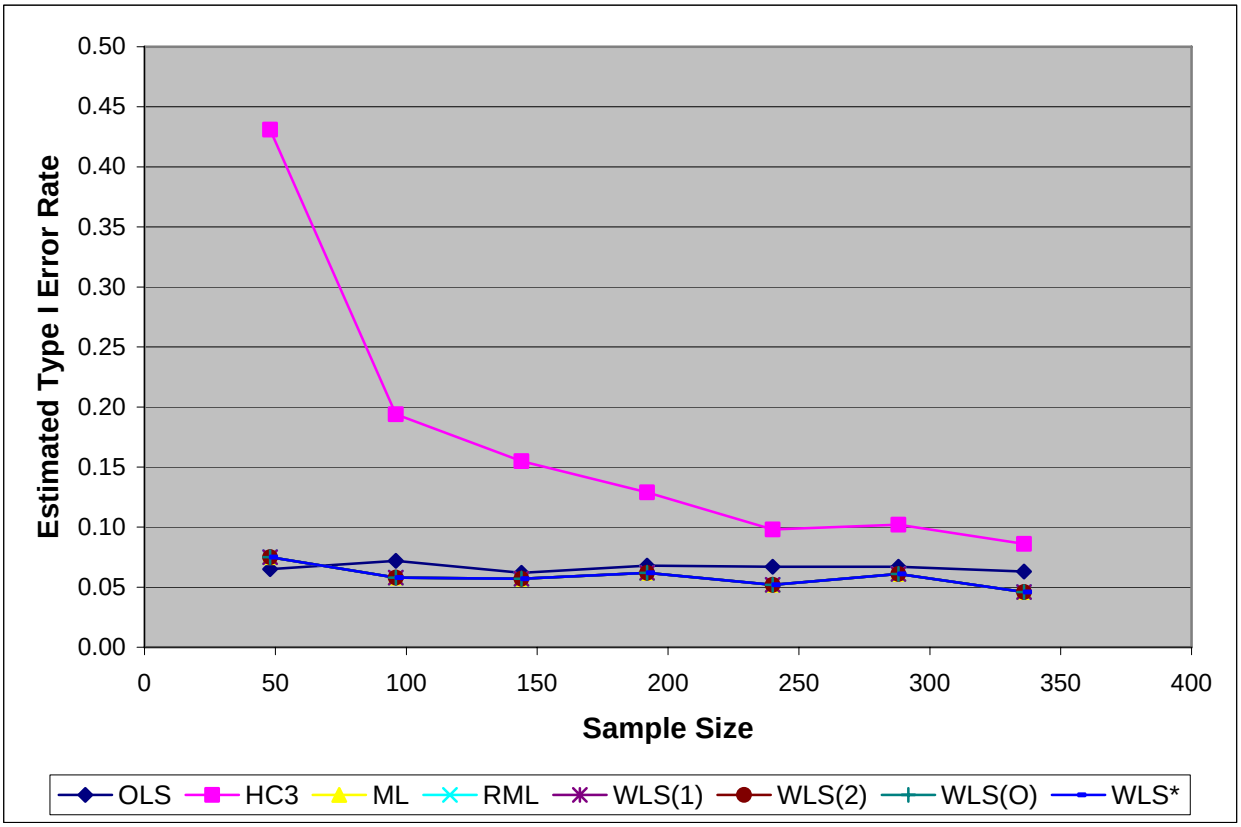
Note. The Type I error rates for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

Panel C



Note. The Type I error rates for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

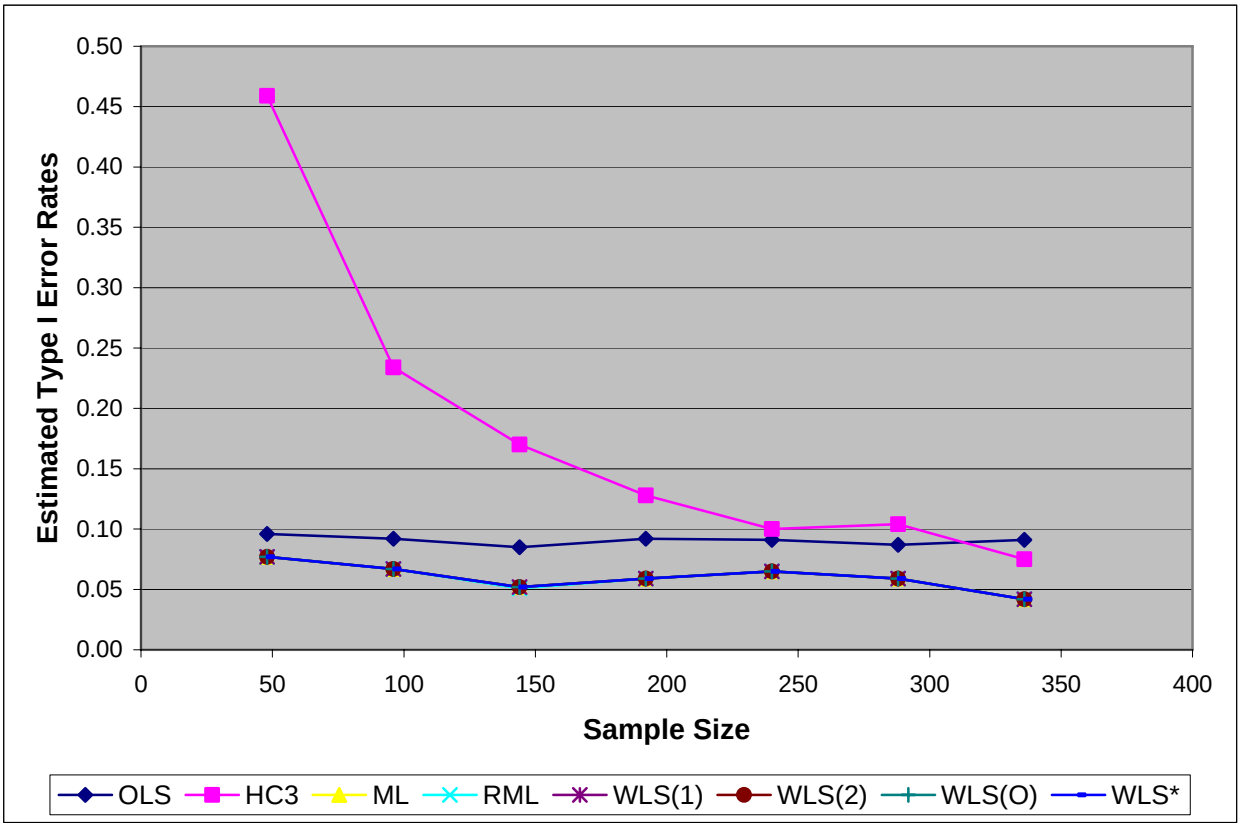
Panel A



Note. The Type I error rates for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

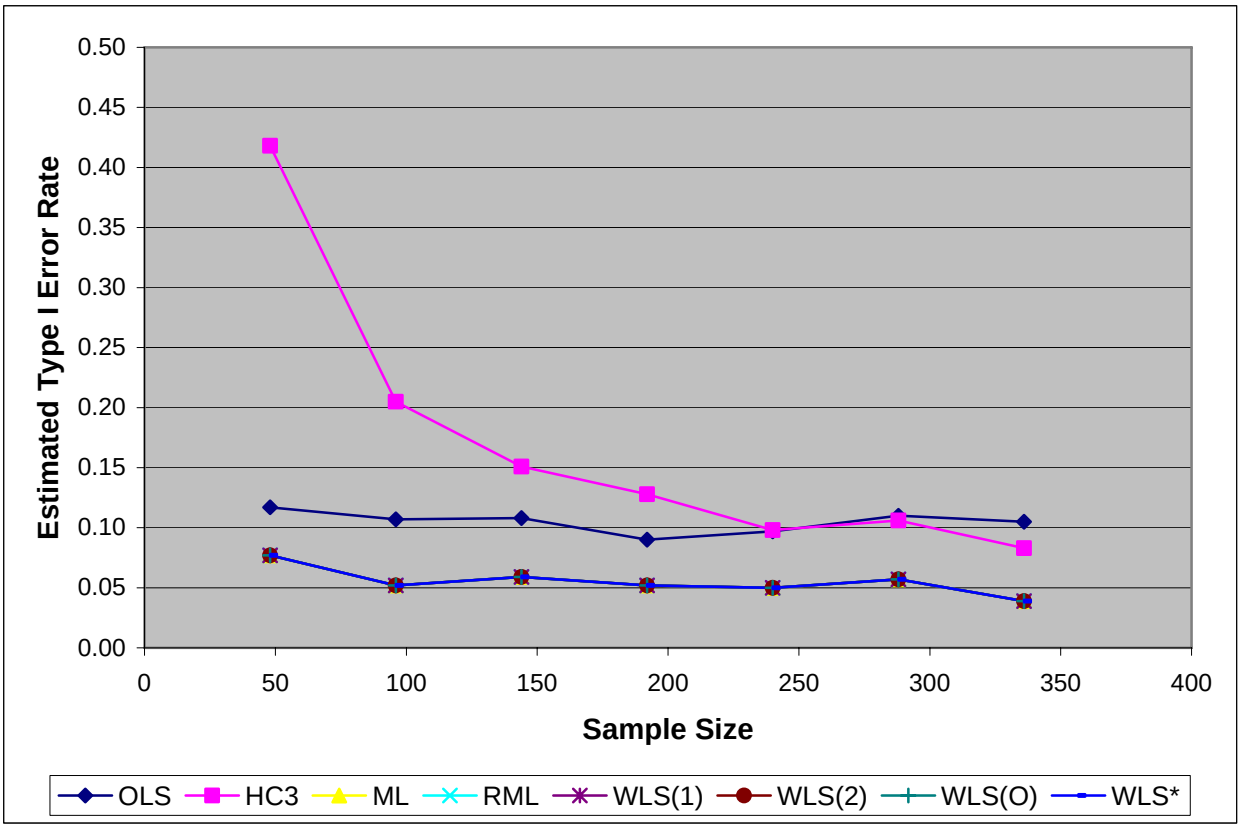
Figure 2. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with four groups, equal subgroup sample sizes, and (A) $\sigma_{e_j}^2$ s = 4, 1, 1, 1, (B) $\sigma_{e_j}^2$ s = 16, 1, 1, 1, and (C) $\sigma_{e_j}^2$ s = 64, 1, 1, 1.

Panel B



Note. The Type I error rates for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

Panel C



Note. The Type I error rates for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

Table 6 presents descriptive statistics for the eight tests across various conditions which deserve noting. With heteroscedasticity and equal ${}_kP_j$s (denoted by H(E)), the average Type I error rate for F_{OLS} and F_{HC3} were inflated with values of .0799 and .1464, respectively. F_{HC3} had a very positively skewed distribution (2.1573). The other six methods performed identically with an average Type I error rate of .0563.

In the same table, across all of the conditions where heteroscedasticity existed (denoted by H), F_{OLS} and F_{HC3} had an average Type I error rate of .0848 and .1716, respectively. Most notably, for F_{HC3} only, the distribution of its Type I error rates *did not* include the value of .05! Its minimum and maximum were .061 and .587, respectively. The other six methods had average Type I error rates ranging from .0509 for F_{WLS*} to .0608 for F_{ML}. Note that F_{WLS*} not only had an average Type I error rate closer to the nominal α than all the other methods, but it also had a (a) small standard deviation (i.e., .0006), and (b) symmetric distribution (i.e., -0.1665).

Heteroscedasticity and Unequal ${}_kP_j$s

As can be seen in Table 5, F_{OLS} had Type I error rates that were greatly affected by the main and interactive effects of the study variables ($R^2 = .96$, $p < .01$). The same was true for the other methods with R^2 values ranging from .45 to .88 (ps < .01). Notably, the Type I error rates based on F_{WLS*} appeared to be less affected by the manipulated variables ($R^2 = .16$, $p > .05$).

Because main effects should be interpreted tentatively in the presence of interactions which indicate relations that are conditional on the values or levels of other variables, I focus primarily on the interactions. For F_{OLS}, k interacted with ${}_kP_j$ ($\hat{\beta} = -0.012$) and pairing ($\hat{\beta} = -0.02$). The two-way interaction that applied to most of the methods but not F_{OLS} or F_{HC3} was that between N and ${}_kP_j$. (Text continues on p. 54.)

Table 6

Descriptive Statistics for Empirical Type I Error Rates (at α = .05) when Testing for the Equality of Regression Slopes Across Conditions of Heteroscedasticity

Conditions	Test	Mean (S.E.)	SD	Min	Max	Skewness (S.E.)
ALL	F_{OLS}	.0788 (.0044)	.0697	.000	.305	1.0478 (.1534)
ALL	F_{HC3}	.1709 (.0074)	.1179	.061	.587	1.8264 (.1534)
ALL	F_{ML}	.0606 (.0010)	.0154	.032	.131	1.8593 (.1534)
ALL	F_{RML}	.0582 (.0008)	.0123	.032	.108	1.3441 (.1534)
ALL	$F_{WLS(1)}$	.0595 (.0009)	.0140	.032	.120	1.7515 (.1534)
ALL	$F_{WLS(2)}$	.0583 (.0008)	.0122	.032	.108	1.3570 (.1534)
ALL	$F_{WLS(O)}$	.0555 (.0006)	.0091	.032	.082	0.4363 (.1534)
ALL	F_{WLS*}	.0507 (.0006)	.0088	.020	.077	-0.2749 (.1555)
H	F_{OLS}	.0848 (.0052)	.0749	.000	.305	0.7890 (.1678)
H	F_{HC3}	.1716 (.0081)	.1180	.061	.587	1.8416 (.1678)
H	F_{ML}	.0608 (.0011)	.0155	.032	.131	1.8412 (.1678)
H	F_{RML}	.0583 (.0009)	.0126	.032	.108	1.3598 (.1678)
H	$F_{WLS(1)}$	.0597 (.0010)	.0142	.032	.120	1.7496 (.1678)
H	$F_{WLS(2)}$	.0585 (.0009)	.0125	.032	.108	1.3768 (.1678)
H	$F_{WLS(O)}$	.0558 (.0006)	.0093	.032	.082	0.4799 (.1678)
H	F_{WLS*}	.0509 (.0006)	.0087	.024	.077	-0.1665 (.1703)
H(E)	F_{OLS}	.0799 (.0024)	.0157	.055	.119	0.4016 (.3654)
H(E)	F_{HC3}	.1464 (.0150)	.0975	.068	.459	2.1573 (.3654)

H(E)	F_{ML}	.0563 (.0013)	.0087	.040	.077	0.5254 (.3654)
H(E)	F_{RML}	.0563 (.0013)	.0087	.040	.077	0.5289 (.3654)
H(E)	$F_{WLS(1)}$	.0563 (.0013)	.0087	.040	.077	0.5254 (.3654)
H(E)	$F_{WLS(2)}$	.0563 (.0013)	.0087	.040	.077	0.5254 (.3654)
H(E)	$F_{WLS(O)}$	.0563 (.0013)	.0087	.040	.077	0.5254 (.3654)
H(E)	F_{WLS*}	.0563 (.0013)	.0087	.040	.077	0.5254 (.3654)
H(D)	F_{OLS}	.0133 (.0012)	.0108	.000	.033	0.4991 (.2627)
H(D)	F_{HC3}	.1661 (.0123)	.1132	.069	.535	1.8482 (.2627)
H(D)	F_{ML}	.0586 (.0014)	.0127	.040	.106	1.4166 (.2627)
H(D)	F_{RML}	.0563 (.0011)	.0105	.038	.091	0.8411 (.2627)
H(D)	$F_{WLS(1)}$	.0573 (.0013)	.0117	.040	.104	1.2726 (.2627)
H(D)	$F_{WLS(2)}$	.0563 (.0011)	.0105	.039	.091	0.8547 (.2627)
H(D)	$F_{WLS(O)}$	.0541 (.0009)	.0086	.040	.078	0.6153 (.2627)
H(D)	F_{WLS*}	.0502 (.0009)	.0078	.024	.066	-0.4077 (.2673)
H(I)	F_{OLS}	.1588 (.0061)	.0563	.071	.305	0.6403 (.2627)
H(I)	F_{HC3}	.1898 (.0142)	.1299	.061	.587	1.7026 (.2627)
H(I)	F_{ML}	.0653 (.0021)	.0194	.032	.131	1.5211 (.2627)
H(I)	F_{RML}	.0615 (.0017)	.0153	.032	.108	1.2460 (.2627)
H(I)	$F_{WLS(1)}$	.0638 (.0019)	.0174	.032	.120	1.5345 (.2627)
H(I)	$F_{WLS(2)}$	.0617 (.0016)	.0150	.032	.108	1.2823 (.2627)
H(I)	$F_{WLS(O)}$	.0572 (.0011)	.0100	.032	.082	0.3070 (.2627)
H(I)	F_{WLS*}	.0488 (.0009)	.0084	.026	.067	-0.6688 (.2673)

Note. An F statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). ALL = all homoscedasticity and heteroscedasticity conditions. H = all heteroscedasticity conditions only. H(E) = heteroscedasticity conditions with equal subgroup sample sizes only. H(D) = heteroscedasticity conditions with direct pairing only. H(I) = heteroscedasticity conditions with indirect pairing only. Min = minimum. Max = maximum. For all tests, the number of conditions on which the descriptive statistics were based for ALL, H, H(E), H(D), and H(I) were 252, 210, 42, 84, and 84, respectively, except for F_{WLS*} which was based on 245, 204, 42, 81, and 81, respectively.

In addition, only for F_{OLS}, there was a three-way interaction among ${}_kP_j$, pairing, and heteroscedasticity ($\hat{\beta} = 0.017$). To explore these relations further, the following sections consider direct and indirect pairing by referring to Figures 3 – 10 and Table 6.

Direct pairing. In Figures 3 and 4, $k = 3$ and direct pairing exists. In Figure 3, $\sigma_{e_j}^2 = 4, 1, 1$. Consistent with previous research, with direct pairing, F_{OLS} had conservative Type I error rates which became increasingly conservative as ${}_3P_j$s became very unequal as shown in Figure 3A (${}_3P_j = .50, .25, .25$) and Figure 3B (${}_3P_j = .\overline{6}, .1\overline{6}, .1\overline{6}$). In contrast, F_{HC3} had inflated Type I error rates which became more inflated as ${}_3P_j$s became very unequal. Overall, compared to F_{OLS} and F_{HC3}, the other six methods performed well. However, they were slightly affected by N, particularly when it was small and when ${}_3P_j$s were very unequal (see Figure 3B). In Figure 4, $\sigma_{e_j}^2 = 64, 1, 1$. That is, heteroscedasticity was increased. F_{OLS} became increasingly conservative and, not surprisingly, performed worse when coupled with very unequal ${}_3P_j$s (see Figure 4B). In the same figure, $F_{WLS(O)}$ and F_{WLS*} appeared to perform well at controlling Type I error rates. This is consistent with the descriptive statistics in Table 6 where heteroscedasticity existed with direct pairing (denoted by H(D)). $F_{WLS(O)}$ and F_{WLS*} had average Type I error rates of .0541 and .0502, respectively. Their distributions were also approximately symmetric. Note the average Type I error rates of F_{OLS} and F_{HC3} were .0133 and .1661, respectively. For F_{OLS}, its conservative Type I error rates were as low as .0. For F_{HC3}, its inflated Type I error rates were as high as .535. Note that Figures 5 and 6 are analogous to Figures 3 and 4, respectively, but when $k = 4$. In these, the trends are similar. However, the Type I error rates for F_{HC3} are more inflated compared to when $k = 3$. Notably, the performance of F_{ML}, F_{RML}, $F_{WLS(1)}$, and $F_{WLS(2)}$, was generally similar across Figures 3 – 6. (Text continues on p. 63.)

Panel A

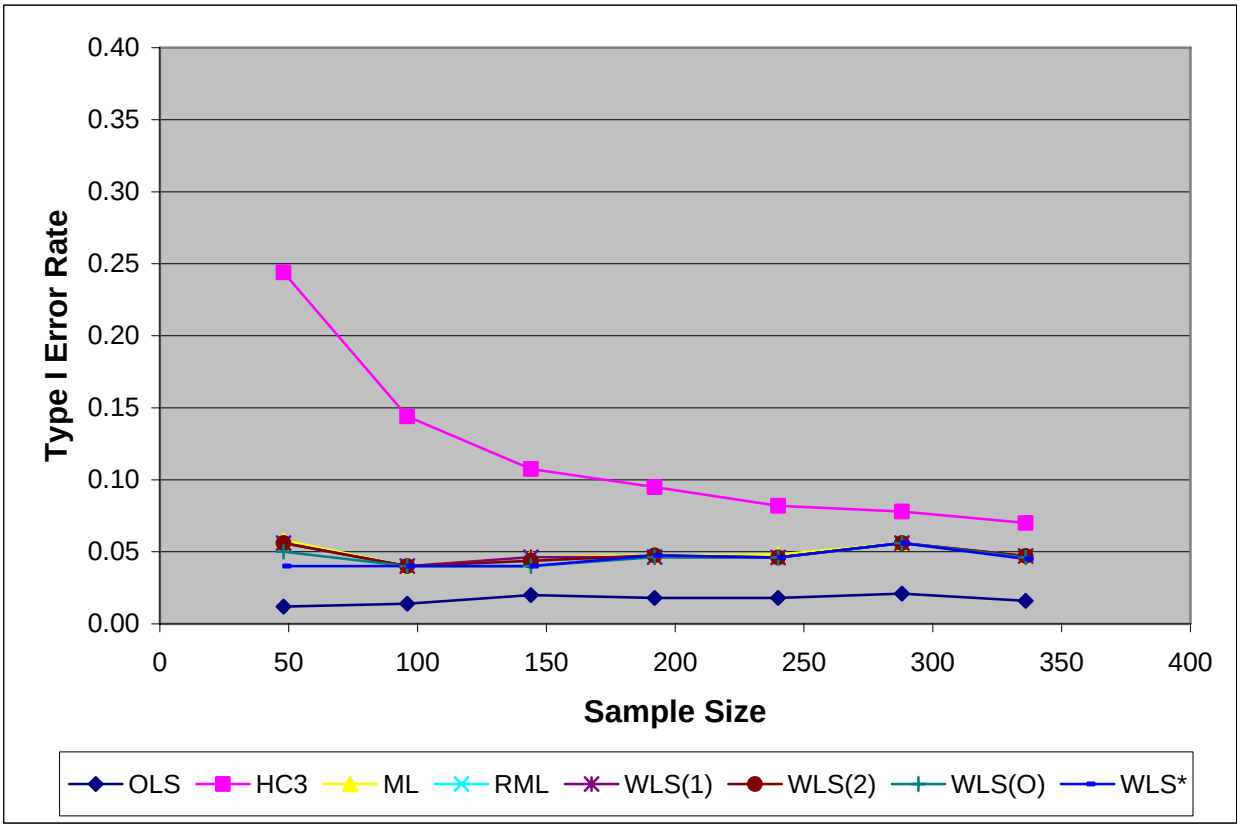
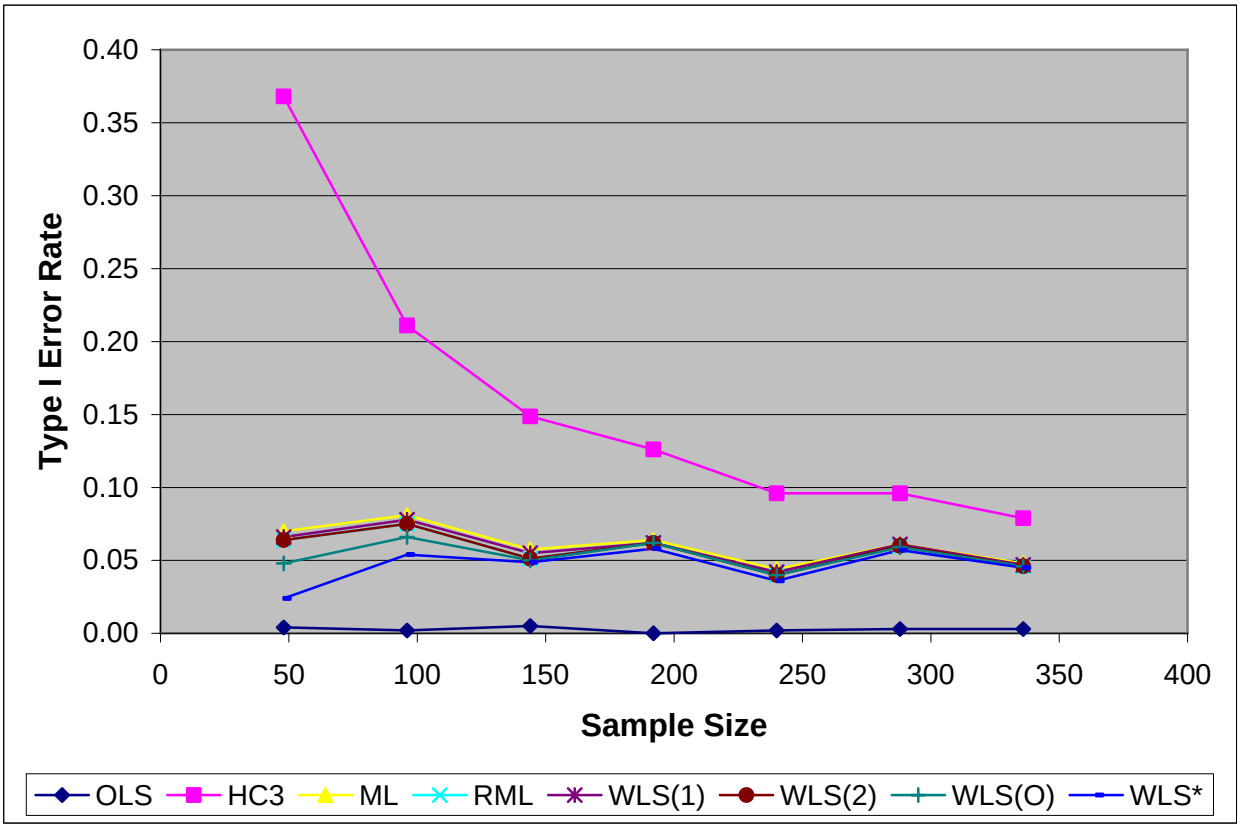


Figure 3. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, $\sigma_{e_j}{}^2$ s = 4, 1, 1 (direct pairing), and proportion within groups equal to (A) .50, .25, .25, and (B) $.\overline{6}$, $.1\overline{6}$, $.1\overline{6}$.

Panel B


Type I Error Rate
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
250
300
350
400
Sample Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

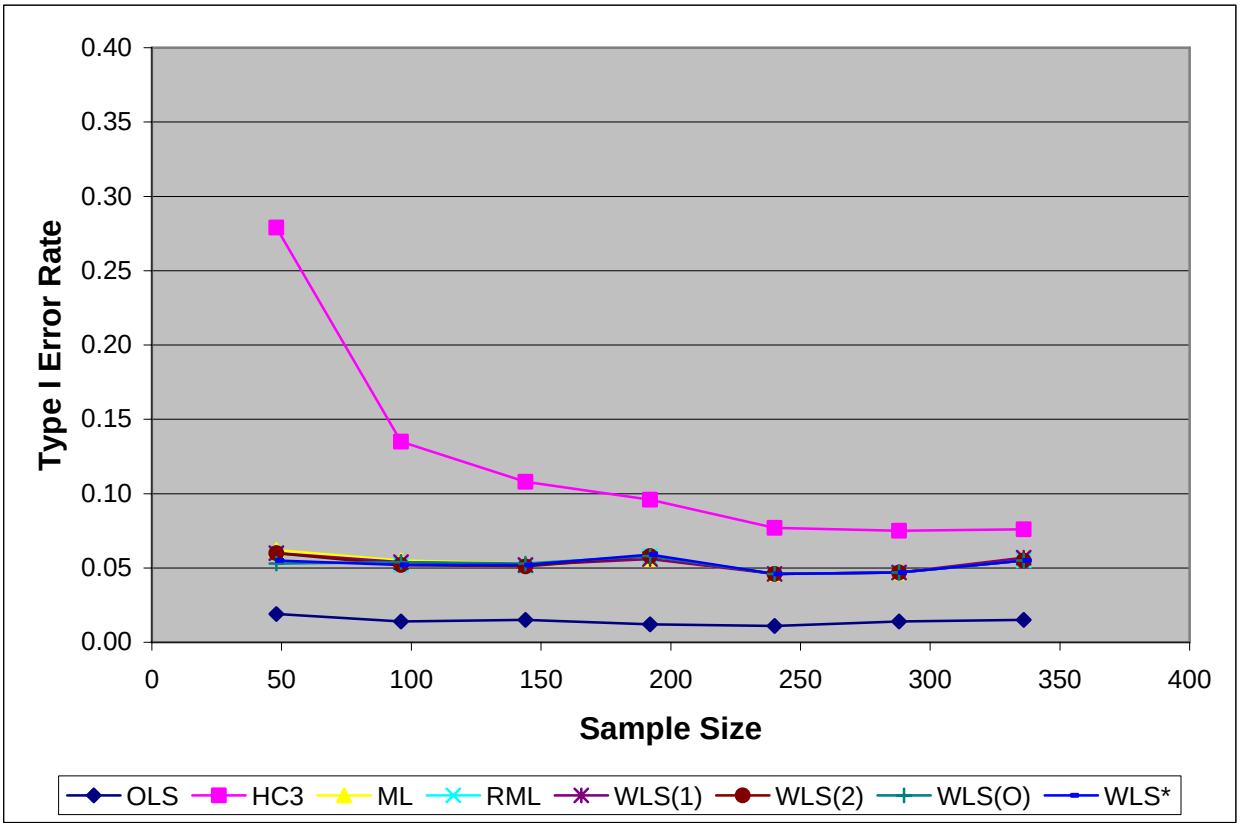
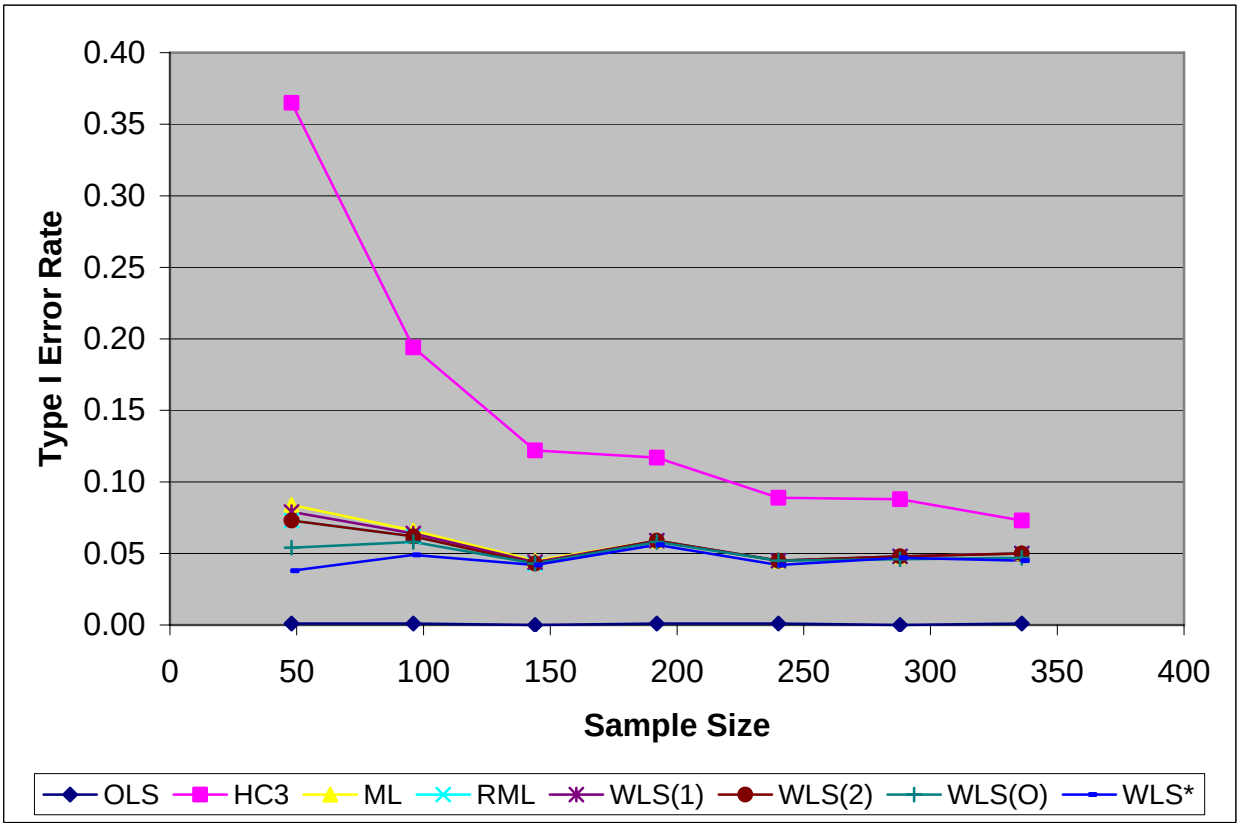


Figure 4. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, $\sigma_{e_j}{}^2$ s = 64, 1, 1 (direct pairing), and proportion within groups equal to (A) .50, .25, .25, and (B) $.\overline{6}$, $.1\overline{6}$, $.1\overline{6}$.

Panel B


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Type I Error Rate
0
50
100
150
200
250
300
350
400
Sample Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

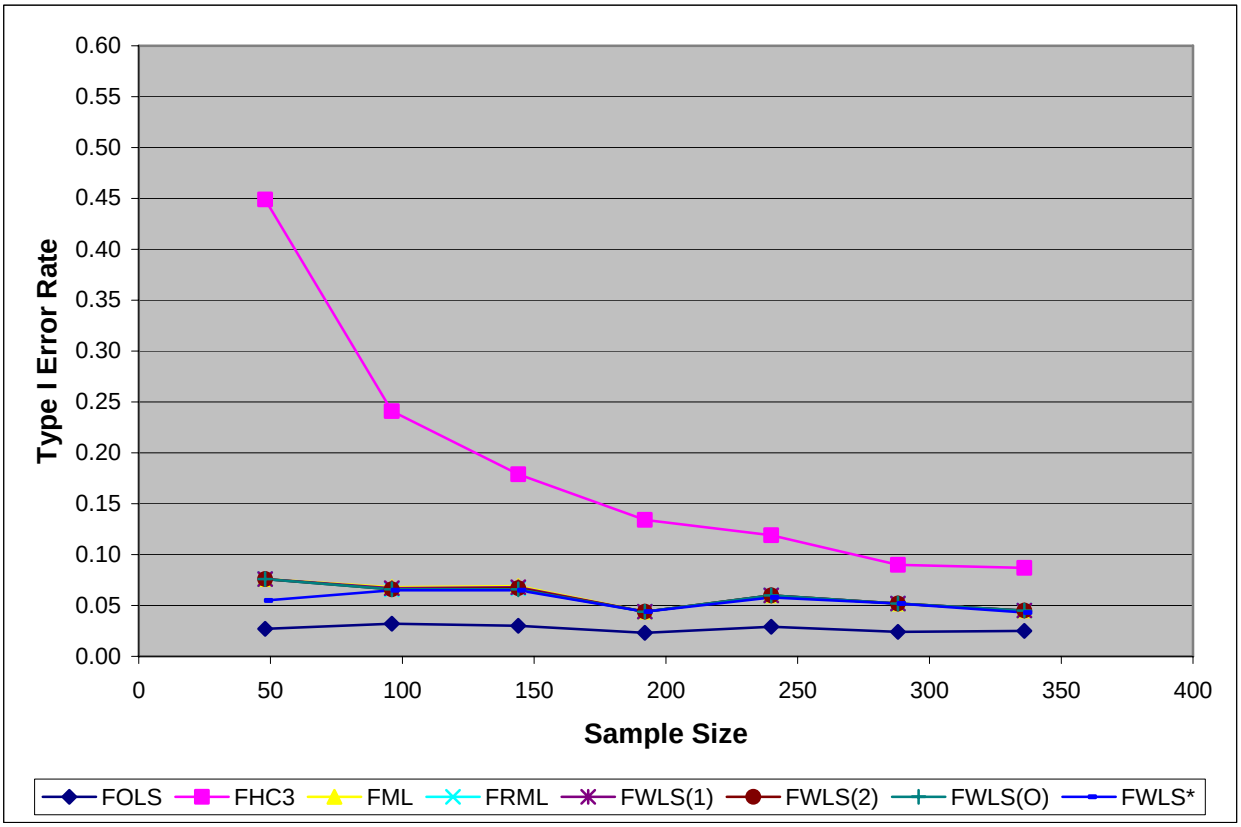
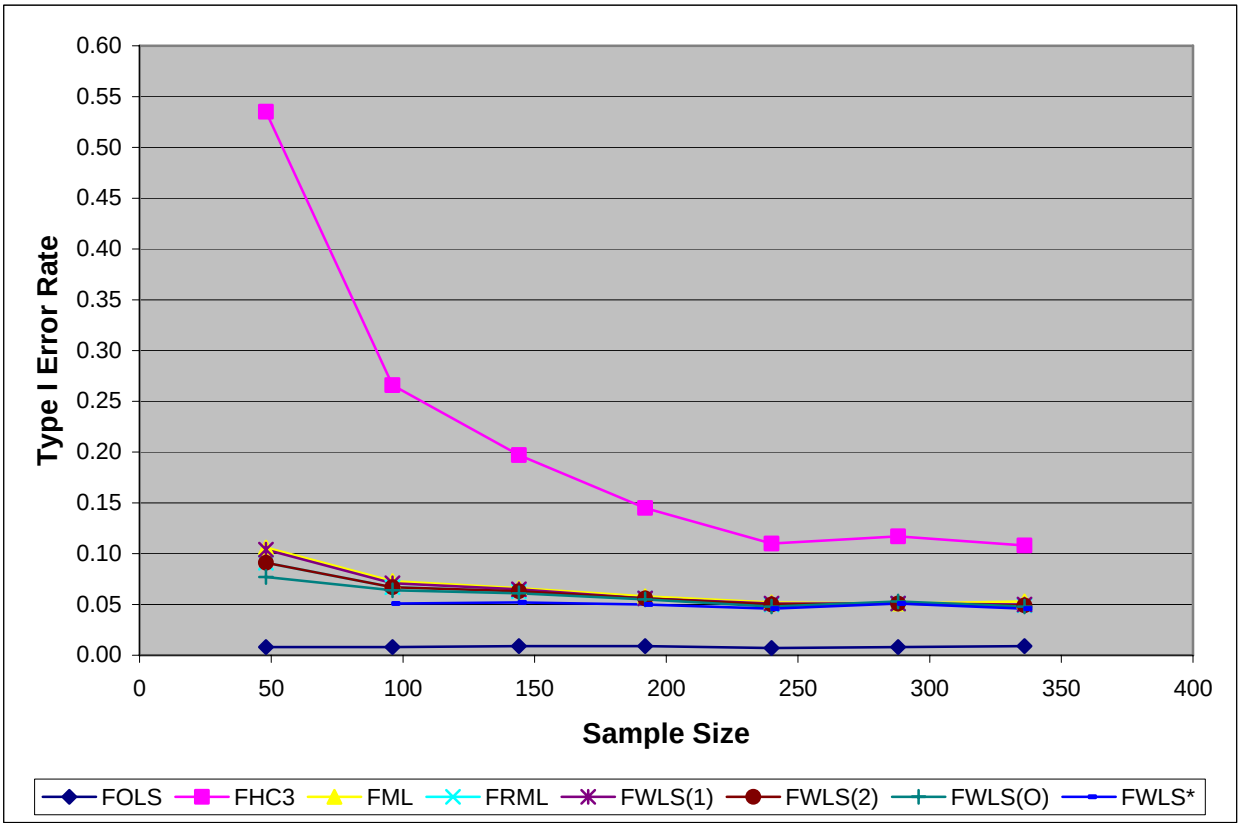


Figure 5. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with four groups, $\sigma_{e_j}^2$ s = 4, 1, 1, 1 (direct pairing), and proportion within groups equal to (A) .375, $.208\overline{3}$, $.208\overline{3}$, $.208\overline{3}$, and (B) .50, $.1\overline{6}$, $.1\overline{6}$, $.1\overline{6}$.

Panel B


0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Type I Error Rate
0
50
100
150
200
250
300
350
400
Sample Size
FOLS
FHC3
FML
FRML
FWLS(1)
FWLS(2)
FWLS(O)
FWLS*

Panel A

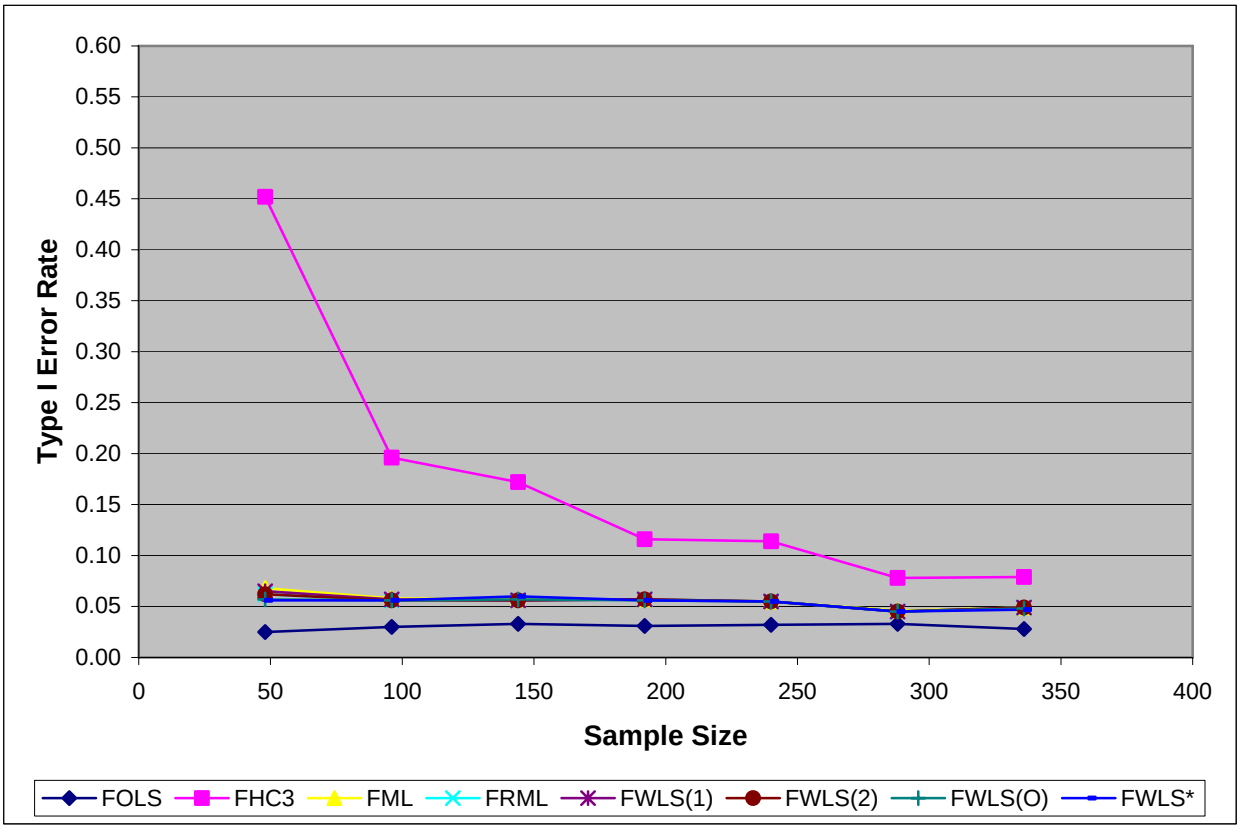
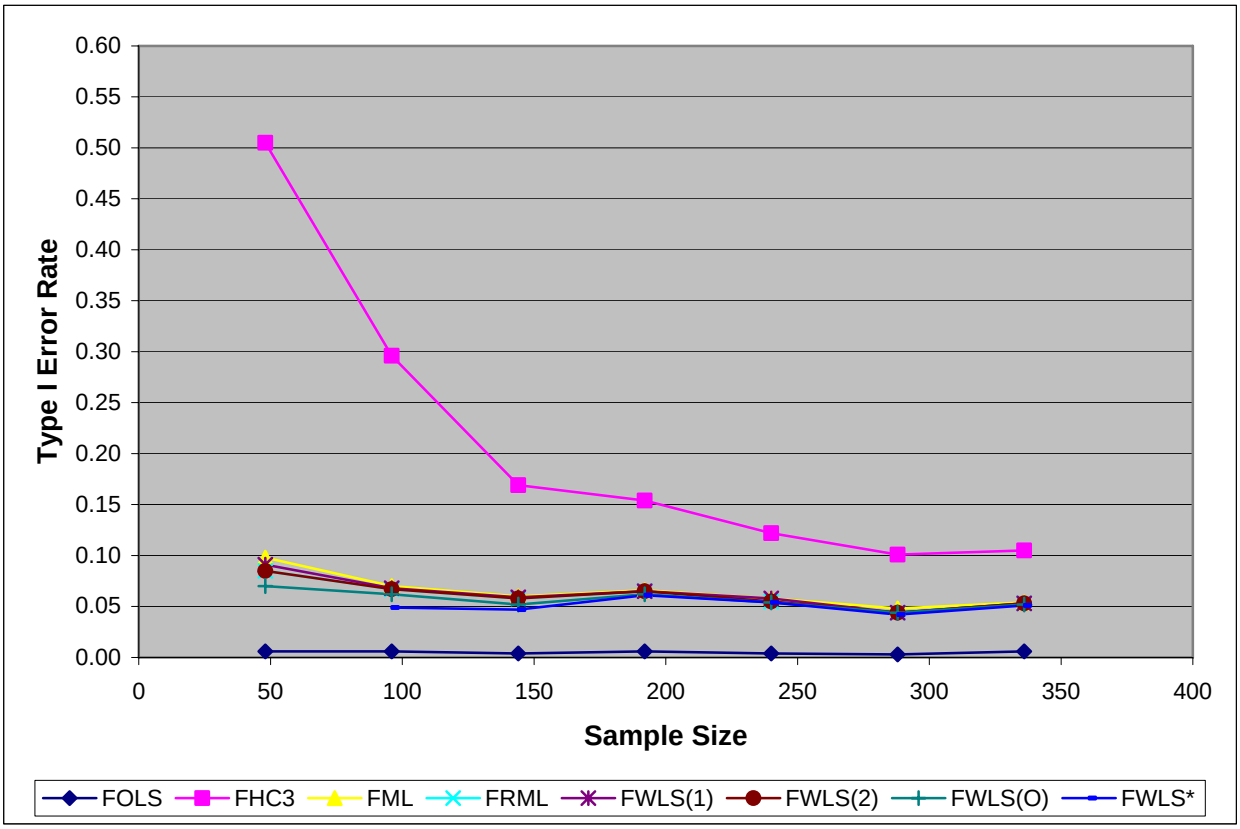


Figure 6. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with four groups, $\sigma_{e_j}{}^2$ s = 64, 1, 1, 1 (direct pairing), and proportion within groups equal to (A) .375, $.208\overline{3}$, $.208\overline{3}$, $.208\overline{3}$, and (B) .50, $.1\overline{6}$, $.1\overline{6}$, $.1\overline{6}$.

Panel B


0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Type I Error Rate
0
50
100
150
200
250
300
350
400
Sample Size
FOLS
FHC3
FML
FRML
FWLS(1)
FWLS(2)
FWLS(O)
FWLS*

Indirect pairing. In Figures 7 and 8, $k = 3$ and indirect pairing exists. In Figure 7, $\sigma_{e_j}^2$ = 1, 1, 4. Consistent with previous research, with indirect pairing, F_{OLS} had inflated Type I error rates which became increasingly inflated as ${}_3P_j$s became more unequal as shown in Figure 7A (${}_3P_j$ = .50, .25, .25) and Figure 7B (${}_3P_j = .\overline{6}, .1\overline{6}, .1\overline{6}$). For F_{HC3}, it continued to have inflated Type I error rates which became increasingly inflated as ${}_3P_j$s became more unequal. Interestingly, for the same amount of heteroscedasticity, Figure 7 (i.e., indirect pairing) compared to Figure 3 (i.e., direct pairing) suggests that F_{HC3} may produce Type I error rates that are more inflated in the indirect pairing than the direct pairing condition. This is consistent with the descriptive statistics in Table 6 for heteroscedasticity with indirect pairing (denoted by H(I)). More specifically, across all indirect pairing conditions, the average Type I error rate for F_{HC3} was .1898 which is greater than that obtained for the direct pairing condition.

Of the remaining six methods, $F_{WLS(O)}$ and F_{WLS*} continued to maintain control of Type I error rates across conditions. For example, in Figure 7, their Type I error rates appeared to be closer to the nominal α than the other methods. The same was true in Figure 8 as heteroscedasticity increased to $\sigma_{e_j}^2$ = 1, 1, 64. In contrast, F_{OLS} had extremely inflated Type I error rates which became increasingly inflated as ${}_3P_j$s became more unequal as shown in Figure 8A (${}_3P_j$ = .50, .25, .25) and Figure 8B (${}_3P_j = .\overline{6}, .1\overline{6}, .1\overline{6}$). F_{HC3} continued to demonstrate inflated Type I error rates but with a slower rate of convergence to the nominal α when ${}_3P_j$s were greatly unequal. An inspection of the descriptive statistics in the rows denoted by H(I) in Table 6 show that the average Type I error rates for $F_{WLS(O)}$ and F_{WLS*} are closer to the nominal level (i.e., .0572, and .0488, respectively) than the other methods and have distributions that are approximately symmetric. (Text continues on p. 68.)

Panel A

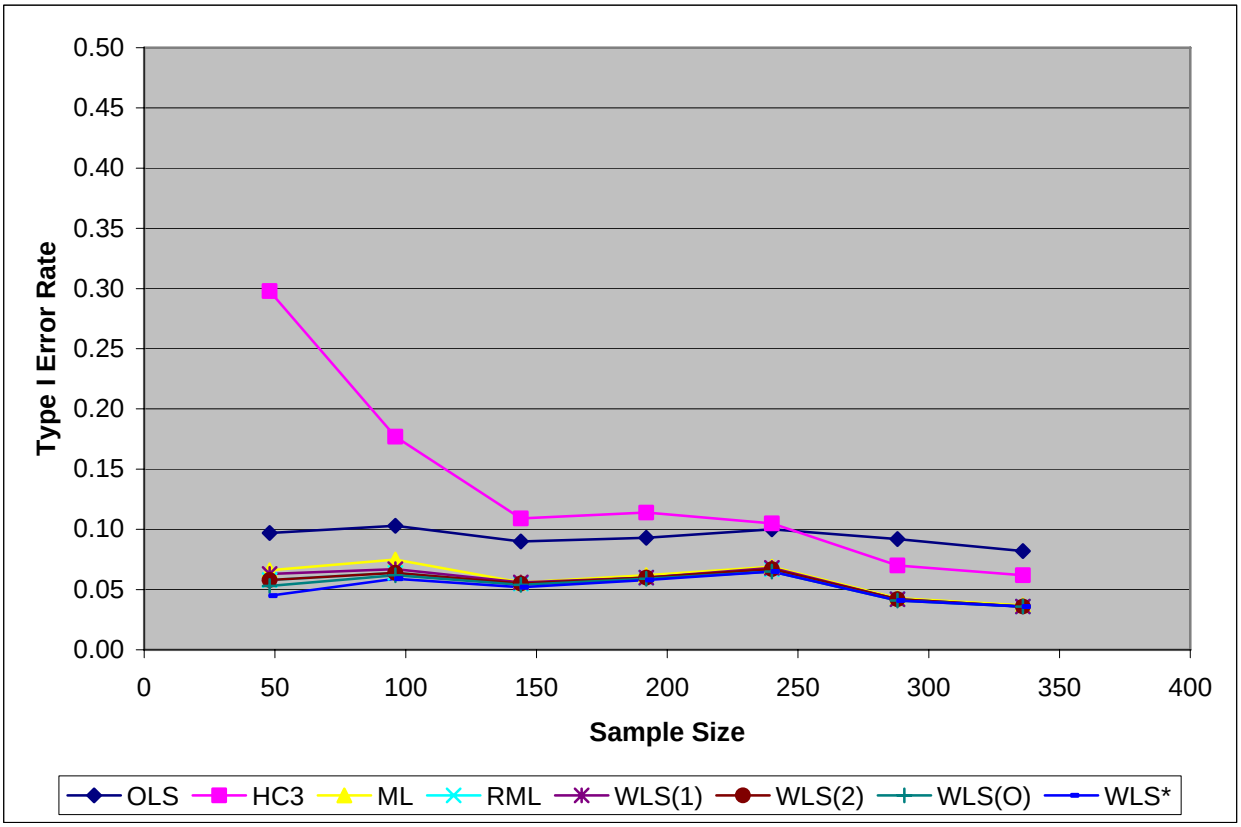
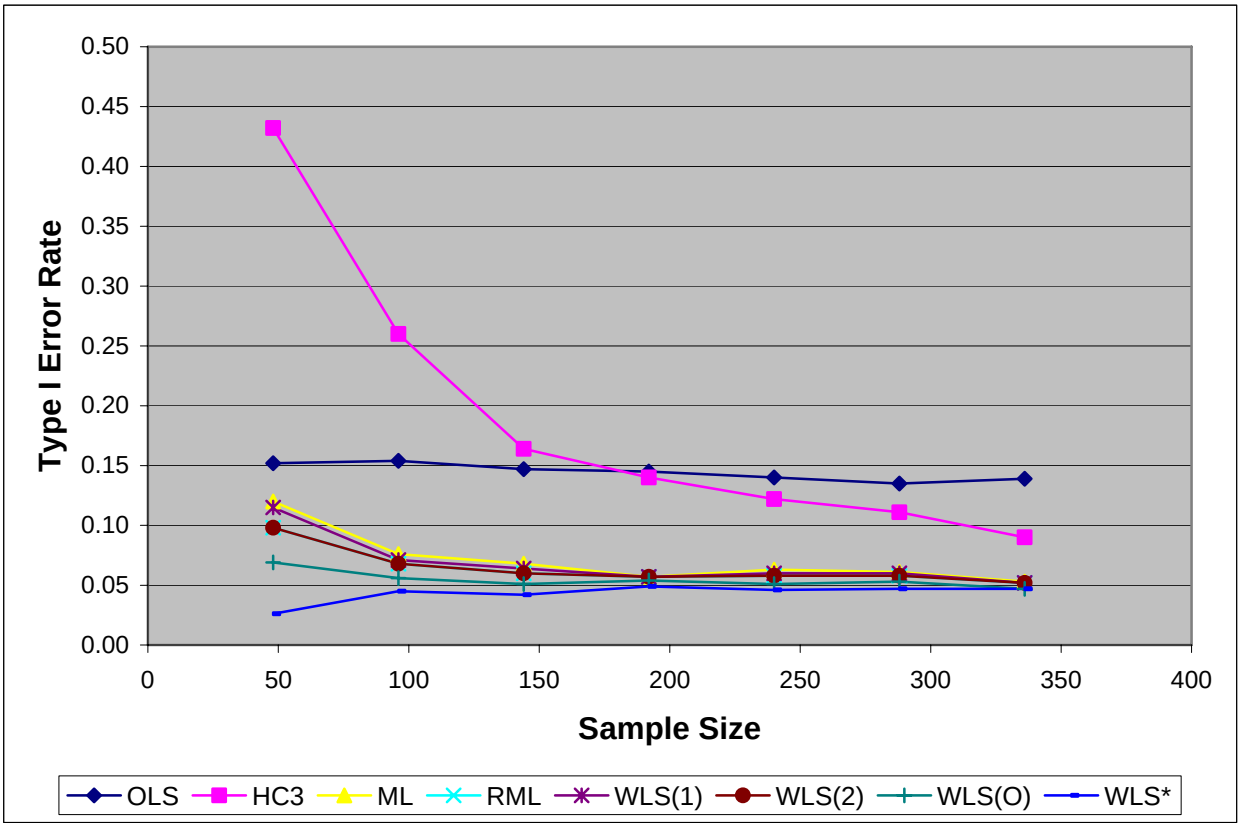


Figure 7. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, $\sigma_{e_j}{}^2$ s = 1, 1, 4 (indirect pairing), and proportion within groups equal to (A) .50, .25, .25, and (B) $.\overline{6}$, $.1\overline{6}$, $.1\overline{6}$.

Panel B


0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Type I Error Rate
0
50
100
150
200
250
300
350
400
Sample Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

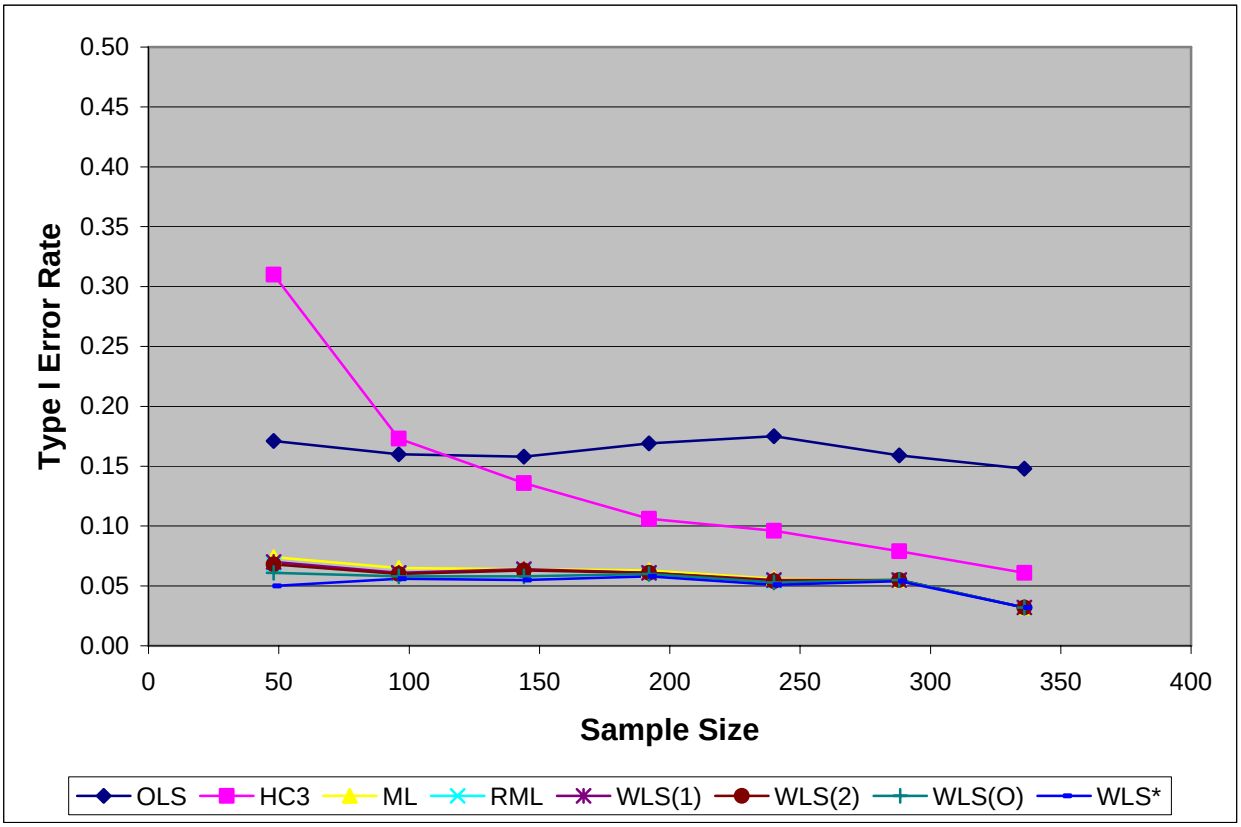
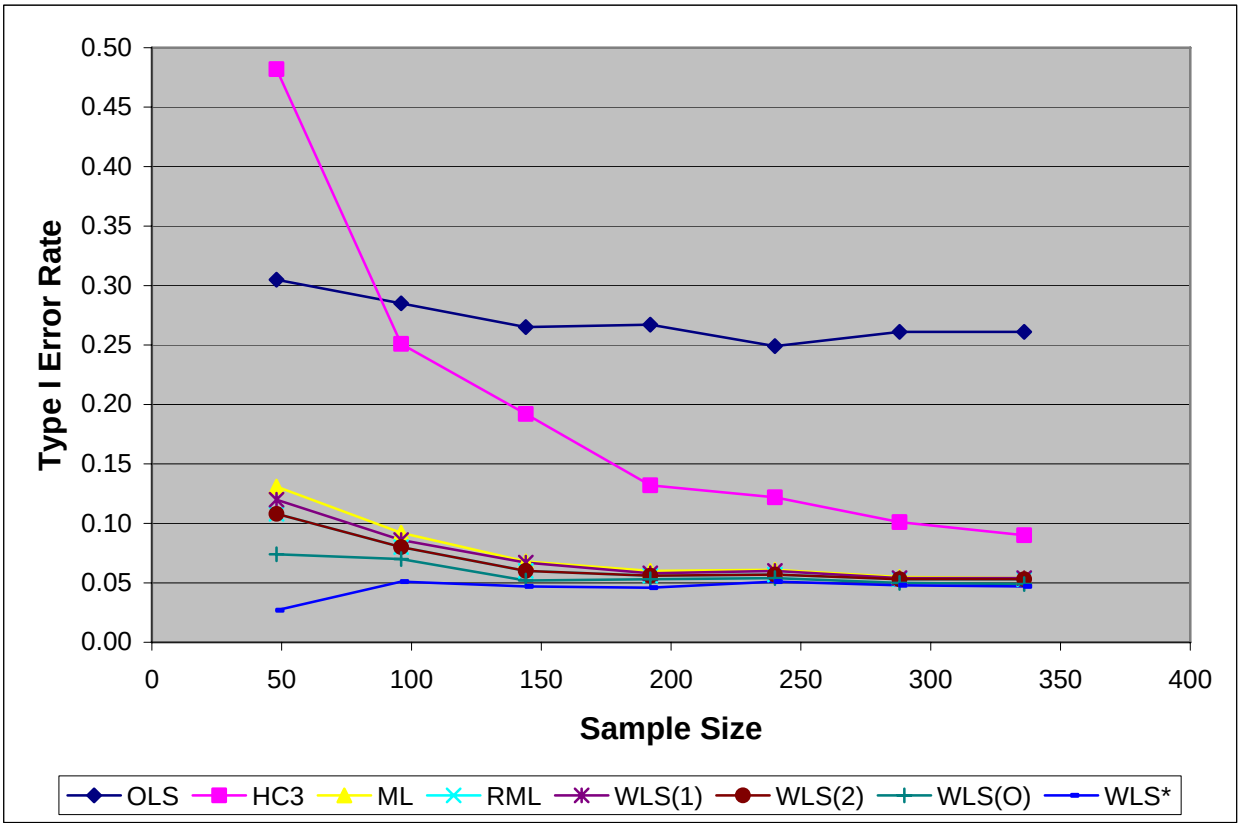


Figure 8. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, $\sigma_{e_j}{}^2$ s = 1, 1, 64 (indirect pairing), and proportion within groups equal to (A) .50, .25, .25, and (B) $.\overline{6}$, $.1\overline{6}$, $.1\overline{6}$.

Panel B


0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Type I Error Rate
0
50
100
150
200
250
300
350
400
Sample Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

In the same table, with indirect pairing, of the alternative methods, F_{HC3}, F_{ML}, and $F_{WLS(1)}$ had Type I error rates as high as .587, .131, and .12, respectively. Note that the effects of indirect pairing when $k = 4$ were similar. In Figure 9, $\sigma_{e_j}^2 = 1, 1, 1, 4$. In Figure 10, $\sigma_{e_j}^2 = 1, 1, 1, 64$. The Type I error rates for F_{ML}, F_{RML}, $F_{WLS(1)}$, and $F_{WLS(2)}$ increased at smaller Ns and very unequal ${}_kP_j$s (see e.g., Figures 9B and 10B).

Comment. To complement the analyses, average Type I error rates are also presented as a function of (a) k and pairing in Table 7, and (b) ${}_kP_j$s, pairing, and degree of heteroscedasticity in Table 8. In these, it appears that F_{OLS} and F_{HC3} are unable to control Type I error rates across conditions. In addition, F_{ML} also demonstrates inflated average Type I error rates with indirect pairing and when ${}_kP_j$s are very unequal (see Table 8).

Statistical Power

As a general summary, Table 9 presents the average power of the eight tests as a function of f^2 and ${}_kP_j$s when heteroscedasticity existed. Five trends deserve noting. First, as f^2 increased, power increased. Second, across conditions, F_{OLS} had the lowest average power compared to the alternative methods. For example, with moderately unequal ${}_kP_j$s, F_{OLS} had an average power of .09 to detect an $f^2 = .002$. F_{HC3} had an average power of .4261 and the other methods ranged between .2709 and .2844. When f^2 increased to .02, F_{OLS} had an average power of .3299 to detect it. F_{HC3} had an average power of .8726 and the other methods ranged between .7679 and .7853. Third, on average, F_{HC3} was the most powerful test. Fourth, when ${}_kP_j$s were equal, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had the same power. Fifth, differences in relative power among the tests diminished at very large f^2s. For example, with very unequal ${}_kP_j$s and $f^2 = .002$, F_{HC3} was (.4678/.2731) = 1.71 times more powerful than F_{WLS*} which decreased to 1.06 when $f^2 = .05$; their relative power became virtually indistinguishable. (Text continues on p. 78.)

Panel A

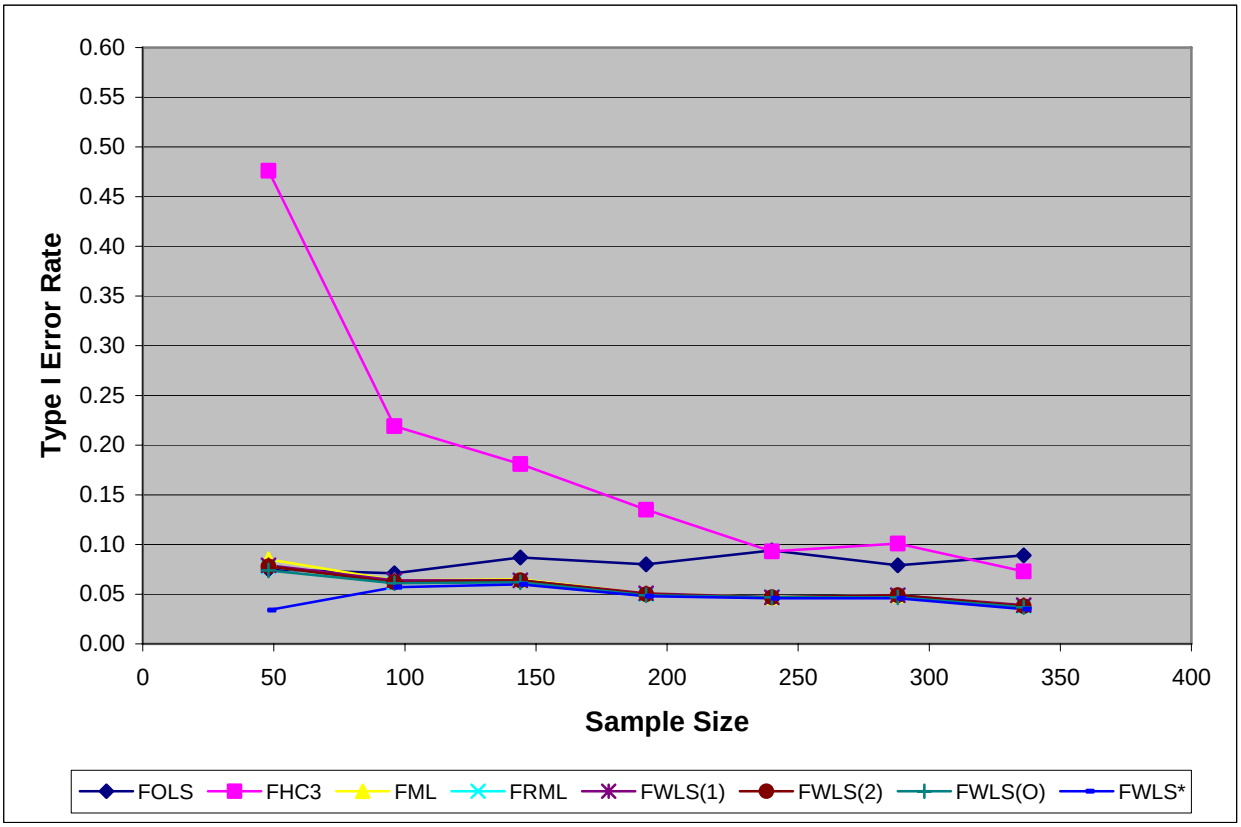
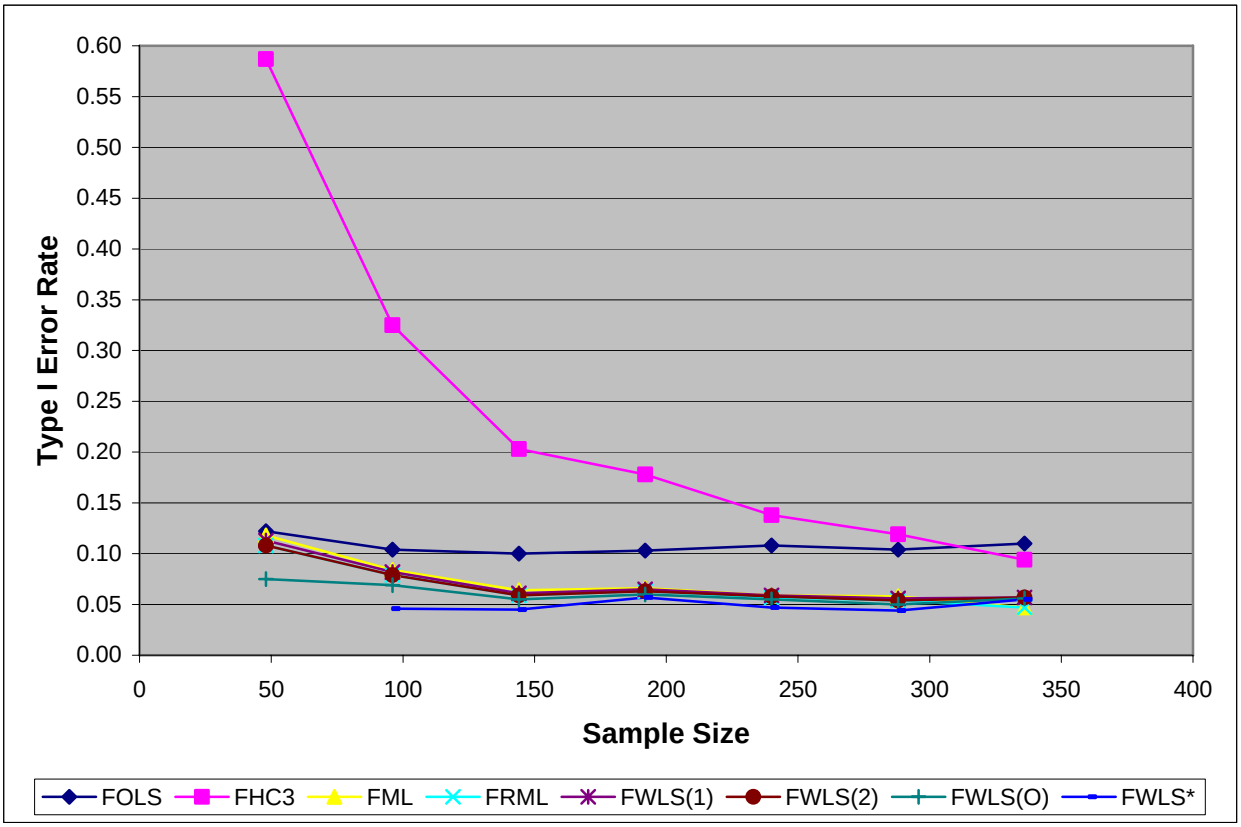


Figure 9. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with four groups, $\sigma_{e_j}{}^2$ s = 1, 1, 1, 4 (indirect pairing), and proportion within groups equal to (A) .375, $.208\overline{3}$, $.208\overline{3}$, $.208\overline{3}$, and (B) .50, $.1\overline{6}$, $.1\overline{6}$, $.1\overline{6}$.

Panel B


Type I Error Rate
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
250
300
350
400
Sample Size
FOLS
FHC3
FML
FRML
FWLS(1)
FWLS(2)
FWLS(O)
FWLS*

Panel A

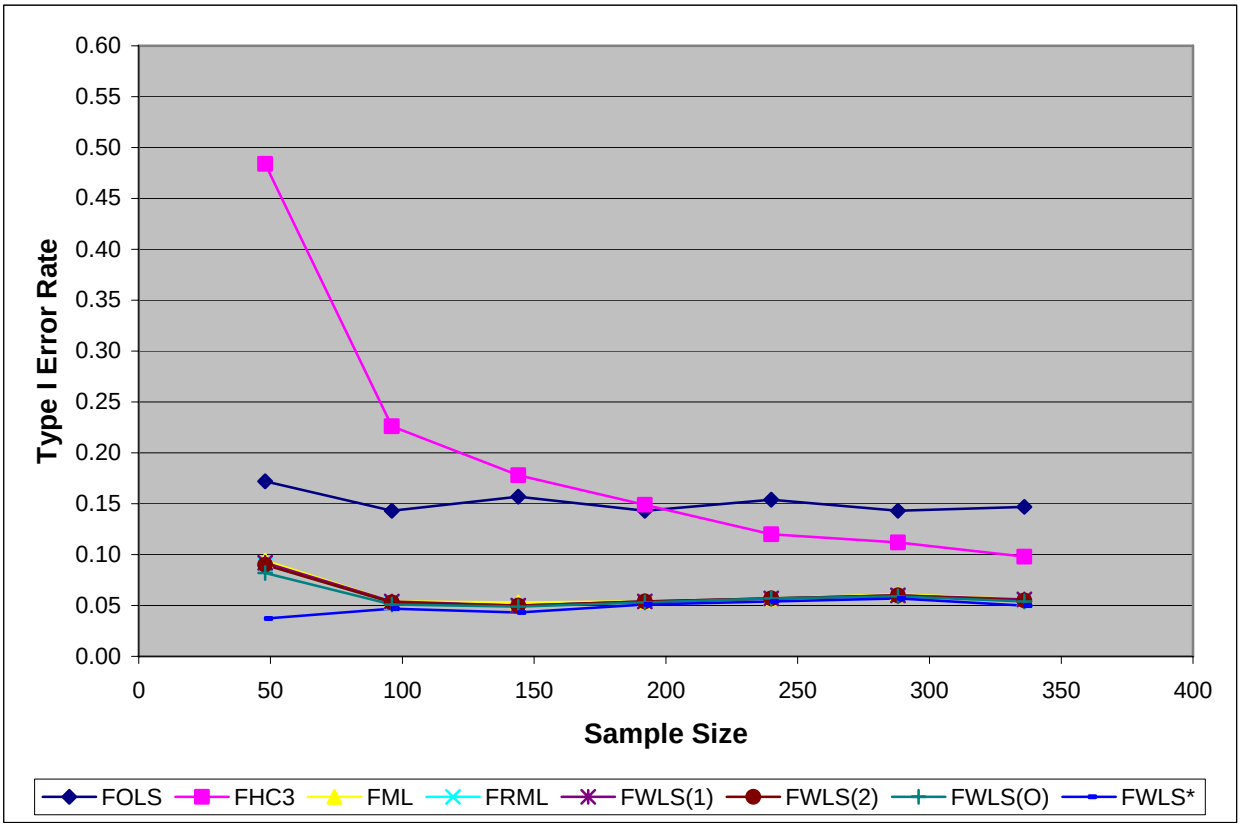


Figure 10. Type I error rates as a function of total sample size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with four groups, $\sigma_{e_j}^2$ s = 1, 1, 1, 64 (indirect pairing), and proportion within groups equal to (A) .375, $.208\overline{3}$, $.208\overline{3}$, $.208\overline{3}$, and (B) .50, $.1\overline{6}$, $.1\overline{6}$, $.1\overline{6}$.

Panel B

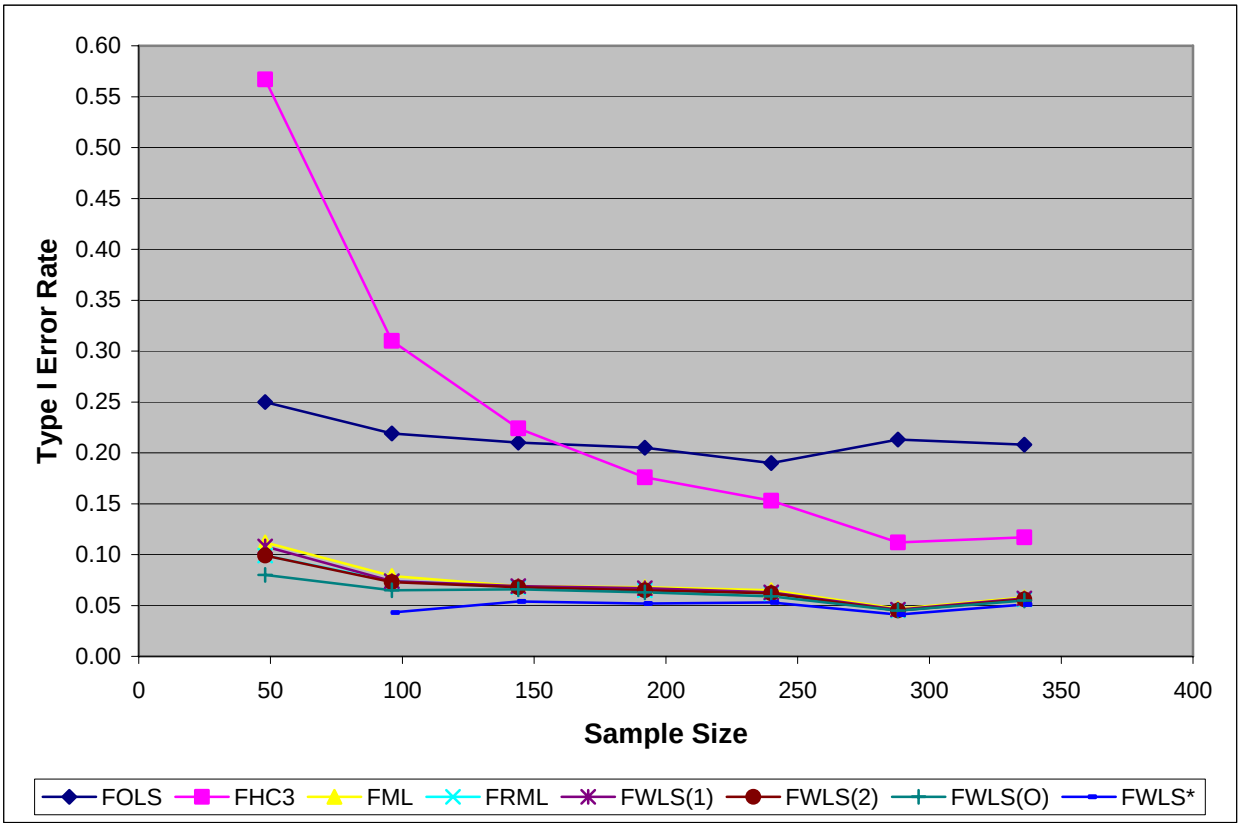

0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Type I Error Rate
0
50
100
150
200
250
300
350
400
Sample Size
FOLS
FHC3
FML
FRML
FWLS(1)
FWLS(2)
FWLS(O)
FWLS*

Table 7

Average Type I Error Rate (at α = .05) as a Function of k and Pairing when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists

K	Pairing	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
3	Direct	.0089	.1373	.0563	.0540	.0550	.0540	.0518	.0485
3	Indirect	.1744	.1669	.0663	.0614	.0640	.0614	.0560	.0487
4	Direct	.0177	.1949	.0609	.0585	.0596	.0585	.0564	.0520
4	Indirect	.1432	.2127	.0643	.0615	.0636	.0621	.0583	.0488

Note. An *F* statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). k = number of groups. For all tests, averages were based on 42 conditions except for F_{WLS*} when $k = 4$ which was based on 39 conditions.

Table 8

Average Type I Error Rate (at $\alpha = .05$) as a Function of ${}_kP_j$s, Pairing, and Amount of Heteroscedasticity when Testing for the Equality of Regression Slopes

Pairing	$\sigma_{e_j}^2$	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
					Moderately Unequal ${}_kP_j$s				
Direct	4	.0221	.1514	.0541	.0534	.0535	.0534	.0524	.0498
Direct	16	.0234	.1527	.0561	.0553	.0557	.0553	.0553	.0535
Direct	64	.0223	.1466	.0547	.0535	.0540	.0535	.0529	.0529
Indirect	4	.0880	.1581	.0577	.0550	.0561	.0550	.0534	.0487
Indirect	16	.1333	.1674	.0592	.0581	.0588	.0580	.0572	.0509
Indirect	64	.1574	.1663	.0599	.0580	.0586	.0580	.0559	.0496
					Very Unequal ${}_kP_j$s				
Direct	4	.0054	.1859	.0632	.0589	.0614	.0589	.0555	.0476
Direct	16	.0036	.1814	.0626	.0589	.0602	.0589	.0551	.0490
Direct	64	.0029	.1786	.0608	.0576	.0591	.0576	.0534	.0479
Indirect	4	.1259	.2116	.0710	.0656	.0694	.0664	.0572	.0458
Indirect	16	.2064	.2191	.0711	.0653	.0698	.0664	.0597	.0502

Indirect	64	.2420	.2164	.0728	.0669	.0702	.0668	.0596	.0470

Note. An F statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). ${}_kP_j$ = degree of disproportionate subgroup sample sizes. $\sigma_{e_j}^2$ = degree of heteroscedasticity. For all tests, averages were based on 14 conditions except for F_{WLS*} when ${}_kP_j$s were Very Unequal which was based on 13 conditions.

Table 9

Average Empirical Power (at α = .05) as a Function of f^2 and ${}_kP_j$s when Testing for the Equality of Regression Slopes and Heteroscedasticity Exists

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
				Equal ${}_kP_j$s				
.002	.0985	.3928	.2653	.2653	.2653	.2653	.2653	.2653
.01	.1814	.7250	.6130	.6130	.6130	.6130	.6130	.6130
.02	.3359	.8565	.7679	.7679	.7679	.7679	.7679	.7679
.05	.6955	.9598	.9123	.9123	.9123	.9123	.9123	.9123
.08	.8237	.9838	.9544	.9544	.9544	.9544	.9544	.9544
				Moderately Unequal ${}_kP_j$s				
.002	.0900	.4261	.2844	.2816	.2829	.2827	.2783	.2709
.01	.1699	.7514	.6340	.6307	.6324	.6307	.6268	.6158
.02	.3299	.8726	.7853	.7825	.7840	.7825	.7792	.7679
.05	.6929	.9668	.9217	.9196	.9208	.9196	.9175	.9078
.08	.8176	.9864	.9598	.9583	.9591	.9583	.9567	.9484
				Very Unequal ${}_kP_j$s				
.002	.1108	.4678	.2982	.2888	.2938	.2887	.2774	.2731
.01	.1864	.7759	.6416	.6320	.6373	.6320	.6181	.6237
.02	.3382	.8900	.7888	.7800	.7849	.7800	.7677	.7765
.05	.6953	.9729	.9234	.9176	.9208	.9176	.9087	.9169

.08	.8188	.9894	.9598	.9555	.9578	.9555	.9490	.9547

Note. An F statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). f^2 = modified effect size (Aguinis et al., 2005). ${}_kP_j$ = degree of disproportionate subgroup sample sizes. When ${}_kP_j$s were Equal, for all tests, the means were based on 42 conditions. When ${}_kP_j$s were Moderately Unequal and Very Unequal, for all tests, the means were based on 84 conditions except for F_{WLS*} when ${}_kP_j$s were Very Unequal which was based on 78 conditions.

A Critical Note Regarding F_{HC3}

Recall from the previous section that the Type I error rates for F_{HC3} were *greatly affected* by *N* and it had a positively skewed distribution that *did not* overlap with the nominal α. This is a very important issue because this has implications for power. That is, power functions should have their minimum at α and this should not vary with *N* (Casella & Berger, 2002; Gentle, 2002) However, F_{HC3} will *not* have a power function with a minimum at α. Instead, it will be inflated and the degree of inflation will vary with *N* (i.e., the degree of inflation decreases as *N* increases). Of course, power should increase monotonically as f^2 increases. Overall, comparing any method against F_{HC3} must be viewed in light of *N* and the degree to which its Type I error rate is inflated for a given *N*.

Heteroscedasticity With Equal ${}_kP_j$s

For heteroscedasticity with equal ${}_kP_j$s, Table 10 presents power of the eight tests when $N = 48$ and $k = 3$ and 4. In it, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had the same power. For example, when $k = 3$, $\sigma_{e_1}^2 = 16$, and $f^2 = .02$, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had power equal to .36 and F_{OLS} had power equal .108. In contrast, F_{HC3} had power equal to .681. Note that for a fixed *k*, as heteroscedasticity increased, power increased for all tests. For example, when $k = 4$, $\sigma_{e_1}^2 = 4$, and $f^2 = .05$, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had power equal to .298 and F_{OLS} had power equal .159. For a similar condition, but with $\sigma_{e_1}^2 = 64$, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had power equal to .92 and F_{OLS} had power equal .191. Although this phenomenon is predictable for F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ because they explicitly account for heteroscedasticity, it is less intuitive for F_{OLS}. Recall that F_{OLS} had inflated Type I error rates when ${}_kP_j$s were equal and heteroscedasticity exists. (Text continues on p. 82)

Table 10

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes with Equal $_kP_j$s and Heteroscedasticity Exists (N = 48, k = 3 and 4)

k	f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
				$n_j = 16;\ \sigma_{e_i}^2 = 4$					
3	.002	.066	.271	.067	.067	.067	.067	.067	.067
3	.01	.083	.357	.116	.116	.116	.116	.116	.116
3	.02	.125	.465	.181	.181	.181	.181	.181	.181
3	.05	.215	.651	.345	.345	.345	.345	.345	.345
3	.08	.319	.767	.492	.492	.492	.492	.492	.492
				$n_j = 16;\ \sigma_{e_i}^2 = 16$					
3	.002	.080	.299	.084	.084	.084	.084	.084	.084
3	.01	.090	.508	.223	.223	.223	.223	.223	.223
3	.02	.108	.681	.360	.360	.360	.360	.360	.360
3	.05	.212	.912	.712	.712	.712	.712	.712	.712
3	.08	.331	.980	.898	.898	.898	.898	.898	.898
				$n_j = 16;\ \sigma_{e_i}^2 = 64$					
3	.002	.091	.443	.160	.160	.160	.160	.160	.160
3	.01	.096	.840	.584	.584	.584	.584	.584	.584
3	.02	.110	.962	.853	.853	.853	.853	.853	.853
3	.05	.216	1.000	.991	.991	.991	.991	.991	.991

3	.08	.337	1.000	1.000	1.000	1.000	1.000	1.000	1.000
					$n_j = 12;\ \sigma_{e_l}^2 = 4$				
4	.002	.075	.441	.077	.077	.077	.077	.077	.077
4	.01	.088	.495	.115	.115	.115	.115	.115	.115
4	.02	.108	.551	.162	.162	.162	.162	.162	.162
4	.05	.159	.703	.298	.298	.298	.298	.298	.298
4	.08	.261	.834	.438	.438	.438	.438	.438	.438
					$n_j = 12;\ \sigma_{e_l}^2 = 16$				
4	.002	.104	.485	.100	.100	.100	.100	.100	.100
4	.01	.119	.620	.201	.201	.201	.201	.201	.201
4	.02	.130	.751	.309	.309	.309	.309	.309	.309
4	.05	.201	.910	.627	.627	.627	.627	.627	.627
4	.08	.269	.964	.802	.802	.802	.802	.802	.802
					$n_j = 12;\ \sigma_{e_l}^2 = 64$				
4	.002	.120	.554	.144	.144	.144	.144	.144	.144
4	.01	.129	.837	.473	.473	.473	.473	.473	.473
4	.02	.149	.949	.680	.680	.680	.680	.680	.680
4	.05	.191	.998	.920	.920	.920	.920	.920	.920
4	.08	.262	1.000	.989	.989	.989	.989	.989	.989

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}),

restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). ${}_kP_j$ = degree of disproportionate subgroup sample sizes. For all conditions, population standard deviations on *x* equal 1.5. *N* denotes total sample size and, in each of the *k* groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on *y* equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on *y* equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

As heteroscedasticity increased, Type I error rates increased. Thus, the minimum of its power function is shifting upwards (i.e., illusory gains in power) as heteroscedasticity increases.

Similar to Table 10, Tables 11, 12, and 13 present power for the eights tests when $N = 96$, 144, and 192, respectively. The just-noted trends are evident in these tables. However, it deserves stressing an important trend regarding the power of F_{HC3}. Note that for a fixed k, $\sigma_{e_1}^2$, and f^2, as N increases, power increases for all tests. However, for F_{HC3}, because the minimum of its power function is shifting downwards as N increases, at very small f^2s, its power may be greater when N is small (e.g., 48) than when N is large (e.g., 96). For example, in Table 10 when $k = 3$, $\sigma_{e_1}^2 = 4$, and $f^2 = .002$, F_{HC3} had power equal to .271. Note that its Type I error rate for the analogous condition (i.e., under the null hypothesis) was .248. Therefore, under the alternative hypothesis, power increased monotonically as f^2 increased, with .248 as the minimum of the power function. Interestingly, when N doubled to 96, for the same conditions but in Table 11, F_{HC3} had power equal to .158; a decrease in power. Note that its Type I error rate for the analogous condition (i.e., under the null hypothesis) was .122; the Type I error rate decreased because N increased. In Table 13, for the analogous condition, but with N now equal to 192, F_{HC3} had power equal to .182; a modest increase in power. However, its Type I error rate for the analogous condition (i.e., under the null hypothesis) was .1; that is, its Type I error rate is converging towards α.

To further investigate the power of the tests, I plotted the rejection rates for various conditions. Note that in all curves, the estimated Type I error rates were included in the plots, allowing for power to be depicted from $f^2 = 0$ to .08. Note also that although straight lines were used to connect all points this does not suggest that the actual power curve will follow such a pattern. (Text continues on p. 92.)

Table 11

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes with Equal ${}_kP_j$s and Heteroscedasticity Exists (N = 96, k = 3 and 4)

k	f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
					$n_j = 32;\ \sigma_{e_1}^2 = 4$				
3	.002	.064	.158	.075	.075	.075	.075	.075	.075
3	.01	.106	.342	.194	.194	.194	.194	.194	.194
3	.02	.163	.495	.335	.335	.335	.335	.335	.335
3	.05	.427	.787	.651	.651	.651	.651	.651	.651
3	.08	.654	.932	.847	.847	.847	.847	.847	.847
					$n_j = 32;\ \sigma_{e_1}^2 = 16$				
3	.002	.089	.260	.125	.125	.125	.125	.125	.125
3	.01	.111	.617	.418	.418	.418	.418	.418	.418
3	.02	.177	.851	.728	.728	.728	.728	.728	.728
3	.05	.401	.991	.970	.970	.970	.970	.970	.970
3	.08	.690	1.000	.998	.998	.998	.998	.998	.998
					$n_j = 32;\ \sigma_{e_1}^2 = 64$				
3	.002	.091	.472	.298	.298	.298	.298	.298	.298
3	.01	.120	.960	.921	.921	.921	.921	.921	.921
3	.02	.172	.993	.992	.992	.992	.992	.992	.992

3	.05	.371	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	.08	.713	1.000	1.000	1.000	1.000	1.000	1.000	1.000
					$n_j = 24;\ \sigma_{e_l}^{\ 2} = 4$				
4	.002	.077	.263	.088	.088	.088	.088	.088	.088
4	.01	.109	.371	.151	.151	.151	.151	.151	.151
4	.02	.157	.517	.262	.262	.262	.262	.262	.262
4	.05	.365	.787	.573	.573	.573	.573	.573	.573
4	.08	.562	.903	.754	.754	.754	.754	.754	.754
					$n_j = 24;\ \sigma_{e_l}^{\ 2} = 16$				
4	.002	.099	.329	.128	.128	.128	.128	.128	.128
4	.01	.127	.599	.347	.347	.347	.347	.347	.347
4	.02	.158	.825	.591	.591	.591	.591	.591	.591
4	.05	.357	.986	.939	.939	.939	.939	.939	.939
4	.08	.564	.996	.994	.994	.994	.994	.994	.994
					$n_j = 24;\ \sigma_{e_l}^{\ 2} = 64$				
4	.002	.107	.530	.268	.268	.268	.268	.268	.268
4	.01	.127	.952	.825	.825	.825	.825	.825	.825
4	.02	.174	.993	.981	.981	.981	.981	.981	.981
4	.05	.325	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	.08	.564	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). ${}_kP_j$ = degree of disproportionate subgroup sample sizes. For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Table 12

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes with Equal ${}_kP_j$s and Heteroscedasticity Exists (N = 144, k = 3 and 4)

k	f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
					$n_j = 48;\ \sigma_{e_i}^2 = 4$				
3	.002	.072	.161	.081	.081	.081	.081	.081	.081
3	.01	.153	.351	.232	.232	.232	.232	.232	.232
3	.02	.275	.560	.431	.431	.431	.431	.431	.431
3	.05	.654	.875	.811	.811	.811	.811	.811	.811
3	.08	.877	.984	.968	.968	.968	.968	.968	.968
					$n_j = 48;\ \sigma_{e_i}^2 = 16$				
3	.002	.098	.239	.150	.150	.150	.150	.150	.150
3	.01	.141	.716	.609	.609	.609	.609	.609	.609
3	.02	.244	.943	.898	.898	.898	.898	.898	.898
3	.05	.668	.998	.997	.997	.997	.997	.997	.997
3	.08	.933	1.000	1.000	1.000	1.000	1.000	1.000	1.000
					$n_j = 48;\ \sigma_{e_i}^2 = 64$				
3	.002	.102	.585	.473	.473	.473	.473	.473	.473
3	.01	.150	.996	.991	.991	.991	.991	.991	.991
3	.02	.229	1.000	1.000	1.000	1.000	1.000	1.000	1.000

3	.05	.655	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	.08	.951	1.000	1.000	1.000	1.000	1.000	1.000	1.000
					$n_j = 36;\ \sigma_{e_l}^2 = 4$				
4	.002	.076	.200	.085	.085	.085	.085	.085	.085
4	.01	.134	.385	.213	.213	.213	.213	.213	.213
4	.02	.221	.566	.391	.390	.390	.390	.390	.390
4	.05	.567	.895	.772	.772	.772	.772	.772	.772
4	.08	.801	.973	.929	.929	.929	.929	.929	.929
					$n_j = 36;\ \sigma_{e_l}^2 = 16$				
4	.002	.106	.287	.135	.135	.135	.135	.135	.135
4	.01	.161	.673	.495	.495	.495	.495	.495	.495
4	.02	.223	.911	.820	.820	.820	.820	.820	.820
4	.05	.545	.999	.995	.995	.995	.995	.995	.995
4	.08	.832	1.000	1.000	1.000	1.000	1.000	1.000	1.000
					$n_j = 36;\ \sigma_{e_l}^2 = 64$				
4	.002	.114	.540	.358	.358	.358	.358	.358	.358
4	.01	.156	.977	.955	.955	.955	.955	.955	.955
4	.02	.214	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	.05	.559	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	.08	.879	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). ${}_kP_j$ = degree of disproportionate subgroup sample sizes. For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Table 13

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes with Equal ${}_kP_j$s and Heteroscedasticity Exists (N = 192, k = 3 and 4)

k	f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
					$n_j = 64;\ \sigma_{e_l}^2 = 4$				
3	.002	.085	.182	.113	.113	.113	.113	.113	.113
3	.01	.187	.401	.317	.317	.317	.317	.317	.317
3	.02	.378	.666	.588	.588	.588	.588	.588	.588
3	.05	.794	.951	.934	.934	.934	.934	.934	.934
3	.08	.961	.995	.994	.994	.994	.994	.994	.994
					$n_j = 64;\ \sigma_{e_l}^2 = 16$				
3	.002	.097	.301	.202	.202	.202	.202	.202	.202
3	.01	.189	.796	.728	.728	.728	.728	.728	.728
3	.02	.337	.966	.954	.954	.954	.954	.954	.954
3	.05	.865	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	.08	.991	1.000	1.000	1.000	1.000	1.000	1.000	1.000
					$n_j = 64;\ \sigma_{e_l}^2 = 64$				
3	.002	.107	.705	.631	.631	.631	.631	.631	.631
3	.01	.179	1.000	.997	.997	.997	.997	.997	.997
3	.02	.326	1.000	1.000	1.000	1.000	1.000	1.000	1.000

3	.05	.877	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	.08	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				$n_j = 48;\ \sigma_{e_l}^{\ 2} = 4$					
4	.002	.072	.178	.091	.091	.091	.091	.091	.091
4	.01	.156	.409	.273	.273	.273	.273	.273	.273
4	.02	.295	.658	.490	.490	.490	.490	.490	.490
4	.05	.746	.944	.895	.895	.895	.895	.895	.895
4	.08	.939	.998	.990	.990	.990	.990	.990	.990
				$n_j = 48;\ \sigma_{e_l}^{\ 2} = 16$					
4	.002	.114	.267	.152	.152	.152	.152	.152	.152
4	.01	.164	.732	.619	.620	.619	.619	.619	.619
4	.02	.286	.951	.916	.916	.916	.916	.916	.916
4	.05	.757	1.000	.999	.999	.999	.999	.999	.999
4	.08	.969	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				$n_j = 48;\ \sigma_{e_l}^{\ 2} = 64$					
4	.002	.110	.613	.474	.474	.474	.474	.474	.474
4	.01	.188	.995	.986	.986	.986	.986	.986	.986
4	.02	.274	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	.05	.787	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	.08	.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000

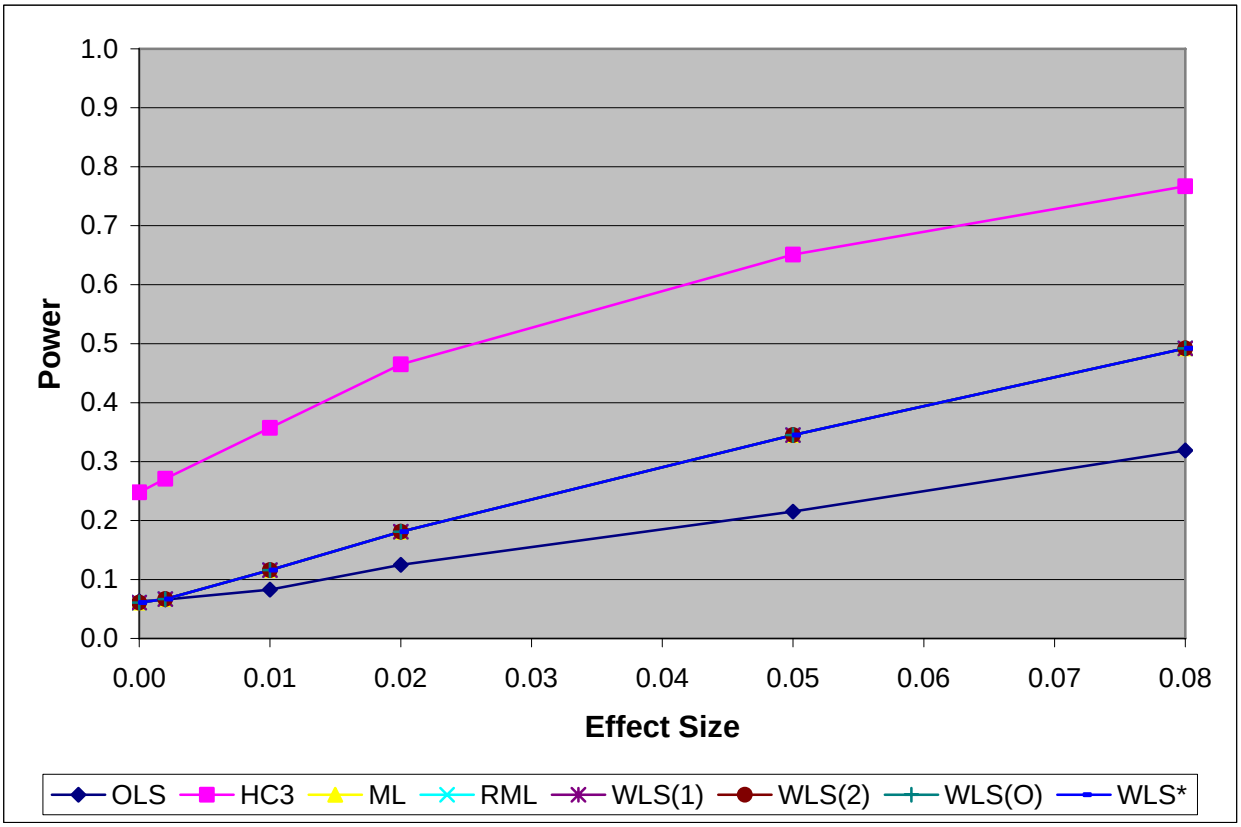
Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). ${}_kP_j$ = degree of disproportionate subgroup sample sizes. For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

There could be smooth curves *between* any two f^2s considered in this study. The lines were included to facilitate interpretation. In addition, because F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ performed identically when ${}_kP_j$s were equal, there power curves coincide.

Figure 11 shows the power of the eight tests when $k = 3$ and $\sigma_{e_j}^2 = 4, 1, 1$. In Figure 11A, $N = 48$ (see also Table 10). Although F_{HC3} clearly has the greatest power among the tests, as noted above, it had the most inflated Type I error rate, resulting in a marked upward shift in its power function. Note that the remaining tests have power levels that increase monotonically but their power curves begin much closer to α than that of F_{HC3}. In Figure 11B, $N = 192$ (see also Table 13). All power curves are much steeper. Note that the power of F_{HC3} is much more closely aligned with the other tests than when $N = 48$ and the minimum of its curve is closer to α, but is still inflated (i.e., .1), resulting in a specious power advantage. Although the Type I error rate for F_{OLS} does not differ markedly from α, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had higher power levels. Finally, in Figure 11C, $N = 336$. In it, the power of F_{HC3} virtually mirrors that of F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$. Importantly, the Type I error rate for F_{HC3} was .068. Therefore, as N increases, it appears that F_{HC3}, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ may be asymptotically equivalent when ${}_kP_j$s are equal.

Figure 12 shows the power of the eight tests when $k = 3$ and $\sigma_{e_j}^2 = 16, 1, 1$. In Figure 12A, $N = 48$ (see also Table 10). Note that, due to increased heteroscedasticity, the power of F_{OLS} increased slightly compared to Figure 11A. However, the power levels of F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$, again, are still considerably higher than that of F_{OLS} without sacrificing control of Type I error rates. Due to its inability to control Type I error rates, F_{HC3} was the most powerful test. These trends can also be seen in Figure 12B where $N = 192$ (see also Table 13). (Text continues on p. 99.)

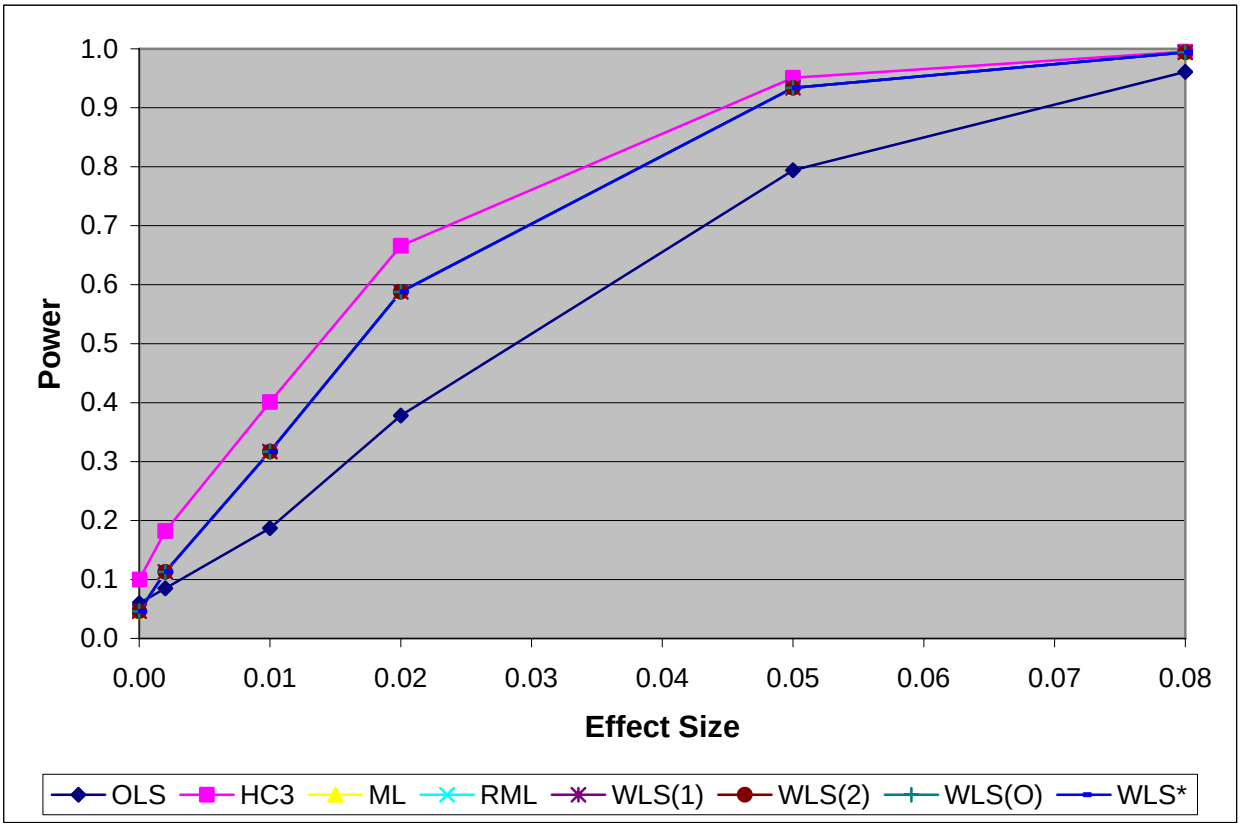
Panel A



Note. Power for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

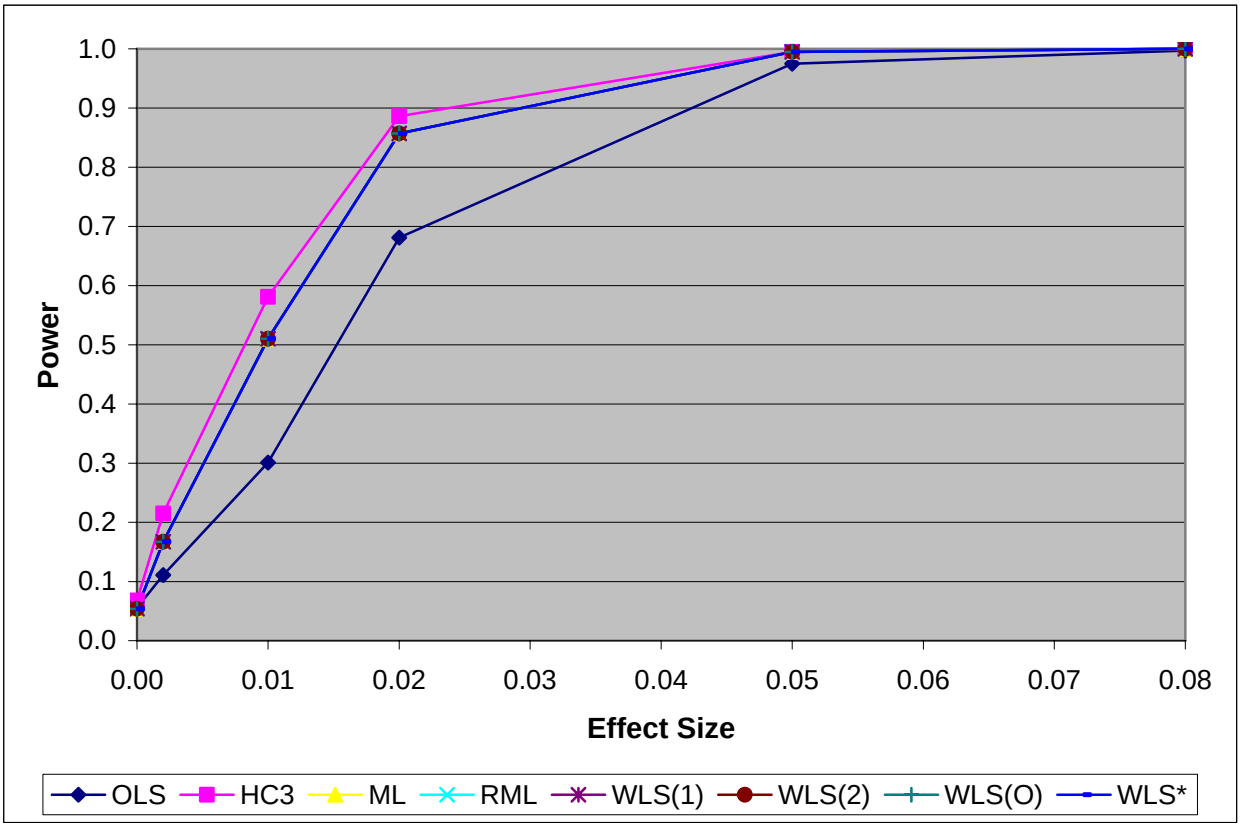
Figure 11. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, equal subgroup sample sizes, $\sigma_{e_j}^2$ s = 4, 1, 1, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B



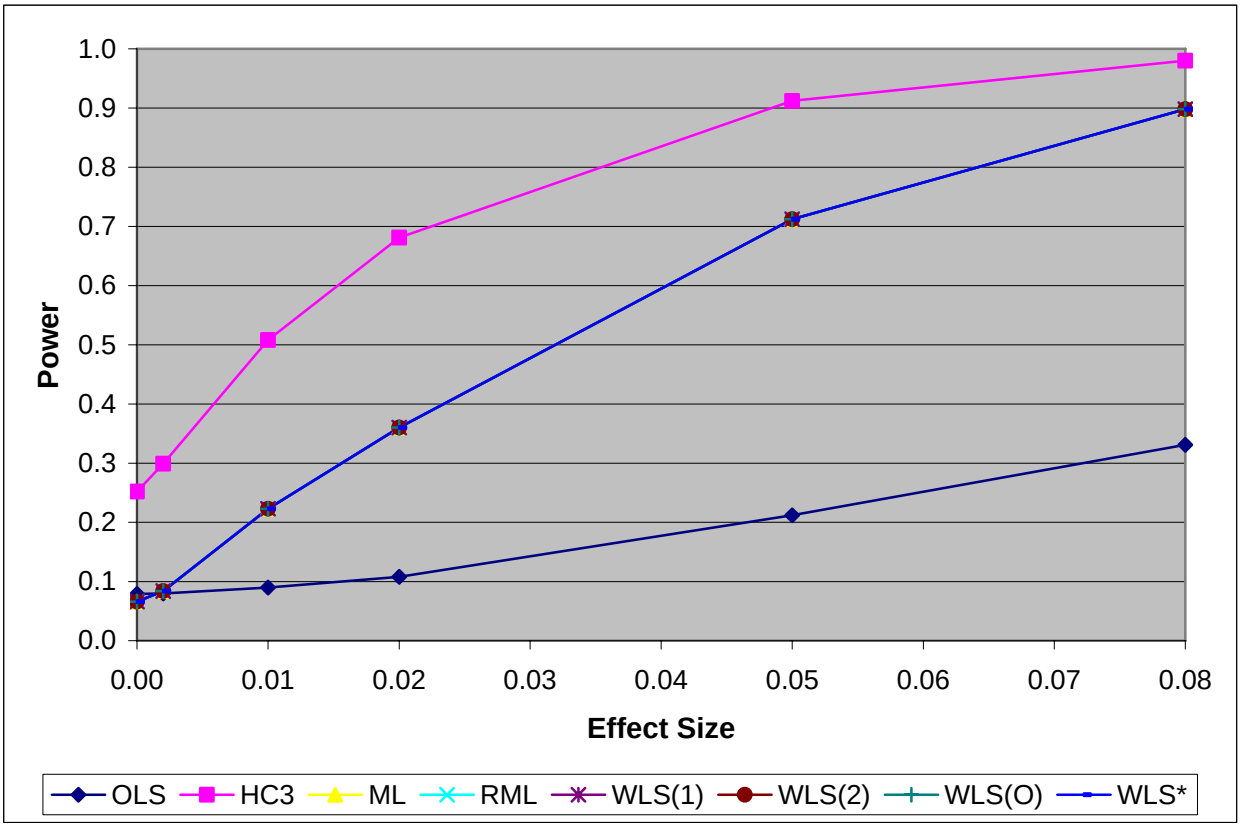
Note. Power for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

Panel C



Note. Power for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

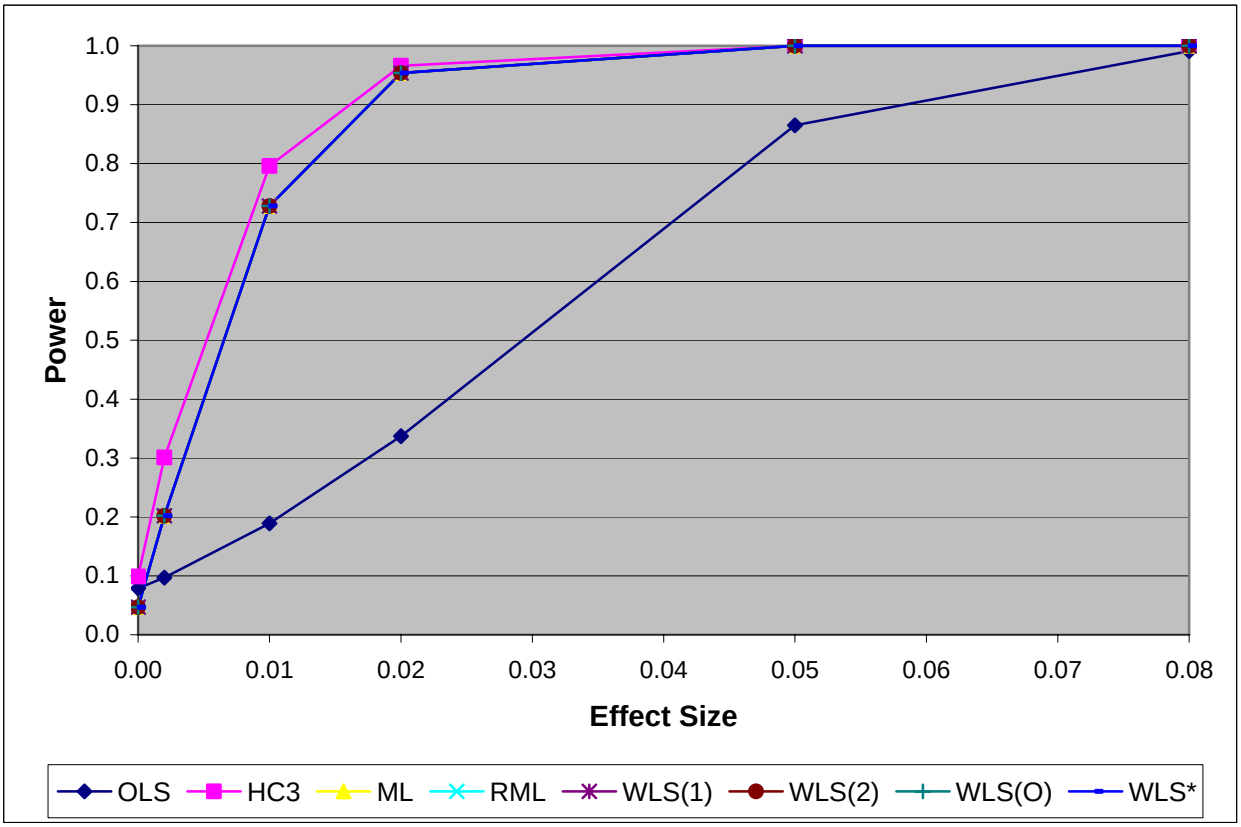
Panel A



Note. Power for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

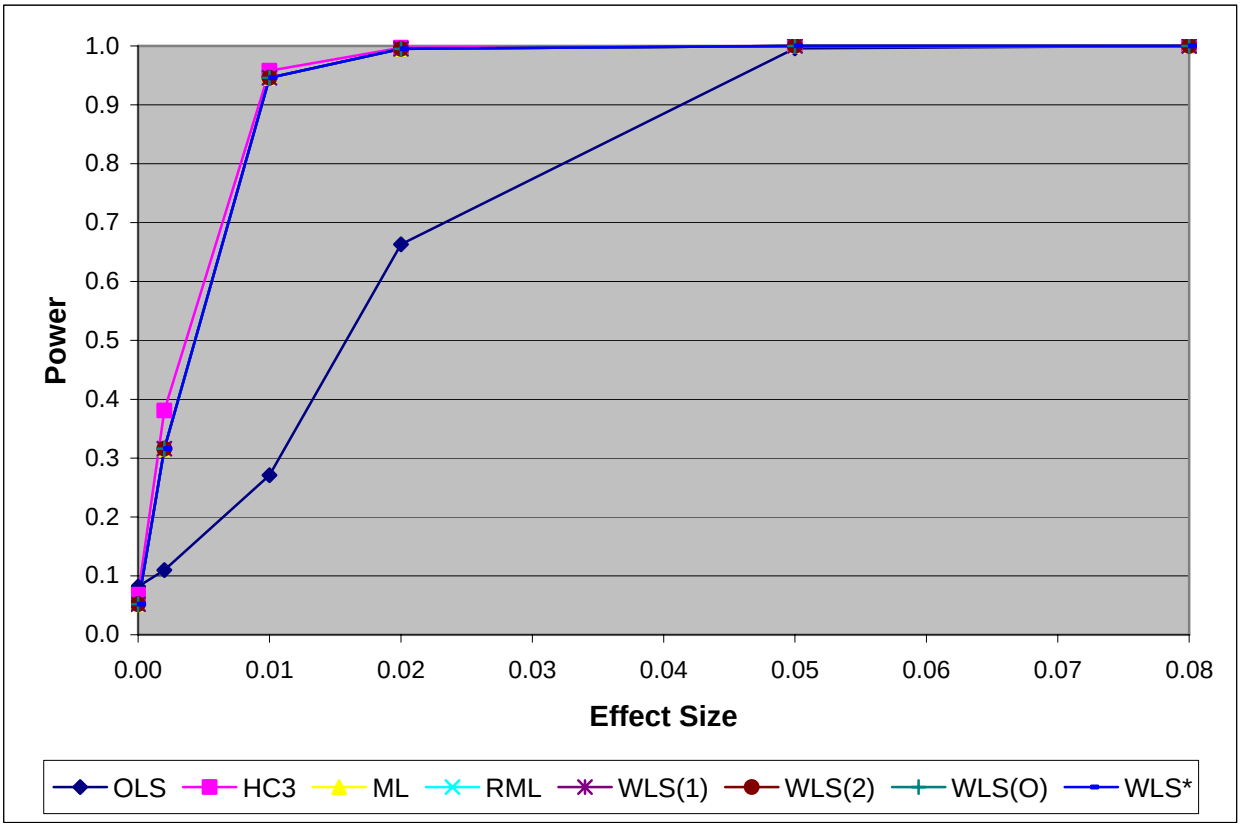
Figure 12. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, equal subgroup sample sizes, $\sigma_{e_j}^2$ s = 16, 1, 1, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B



Note. Power for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

Panel C



Note. Power for F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) coincide.

Note that the power functions of F_{HC3}, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ appear to coincide more quickly when heteroscedasticity increases along with N. This can be seen by comparing (a) Figure 11B with 12B, and (b) Figure 11C with 12C.

Table 14 shows the average power of the tests as a function of k and f^2. On average, for F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$, increasing k tends to lower power. On average, for F_{HC3}, increasing k tends to increase power. For F_{OLS}, because increasing k tends to result in increases in Type error rates, average power increases which is noticeable at smaller f^2s.

Heteroscedasticity With Unequal ${}_kP_j$s

Table 15 presents the average power of the eight tests as a function of ${}_kP_j$s and pairing. In it, F_{OLS} had less power with direct pairing than indirect pairing and it was lowest when direct pairing was combined with very unequal ${}_kP_j$s (.3419). The opposite was true for the remaining seven tests. Specifically, they had greater power with direct pairing than indirect pairing and it was greatest when direct pairing was combined with very unequal ${}_kP_j$s. F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had very similar average power levels. F_{HC3} had the highest average power.

Because inspecting averages across various conditions may not reveal unique and interesting trends, for a subset of representative conditions, I present the power of the eight tests when $k = 3$ in (a) Tables 16 – 19 for direct pairing, and (b) Tables 20 – 23 for indirect pairing.

Direct pairing. For F_{OLS}, the results were consistent with previous research. Namely, power levels were generally low for F_{OLS}. For example, in Table 16 ($N = 48$), F_{OLS} had power equal to .023 to detect an $f^2 = .002$ when $\sigma_{e_j}^2 = 4, 1, 1$ and n_js = 24, 12, 12. Recall that its Type I error rates were very conservative with direct pairing, particularly with very unequal ${}_kP_j$s. For example, consider a similar condition in the same table, but with n_js = 32, 8, 8. The power of

F_{OLS} decreased to .007. For all the alternative methods, power was generally much greater than that of F_{OLS}. Although F_{HC3} always had the greatest power, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had power levels that became very similar to that of F_{HC3} when f^2 increased, ${}_kP_j$s were moderately unequal, and as N increased. This can be seen by inspecting Table 17 when N = 96, Table 18 when N = 144, and Table 19 when N =192.

Figures 13 – 16 depict the power curves for the eight tests across various conditions with direct pairing when k = 3. In Figure 13A, N = 48, ${}_kP_j$s are moderately unequal, and $\sigma_{e_j}^2$ = 4, 1, 1 (see also Table 16). Consistent with the previous results, F_{HC3} had the greatest power. F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ performed similarly with F_{ML} having the greatest power among them. F_{OLS}, with its conservative Type I error rates, had the lowest power. For this condition but as N increased to 192 (Figure 13B) and 336 (Figure 13C), F_{HC3}, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had nearly identical power levels. Note however that F_{HC3} had inflated Type I error rates and F_{OLS} had conservative Type I error rates.

Figure 14 is like that of Figure 13, however, with increased heteroscedasticity (i.e., $\sigma_{e_j}^2$ = 16, 1, 1). Comparing Figure 14A to Figure 13A, it is evident that increased heteroscedasticity resulted in (a) increased power for F_{HC3}, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$, and (b) decreased power for F_{OLS}. In addition, although power increased for all the methods as N increased to 192 (Figure 14B) and 336 (Figure 14C), F_{HC3} had inflated Type I error rates and F_{OLS} had conservative Type I error rates.

Figure 15 depicts the power of the tests when ${}_kP_j$s were very unequal and $\sigma_{e_j}^2$ = 4, 1, 1. (Text continues on p. 127.)

Table 14

Average Empirical Power (at α = .05) as a Function of k and f^2 When Testing for the Equality of Regression Slopes With Equal Subgroup Sample Sizes and Heteroscedasticity Exists

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
				$k = 3$				
.002	.0930	.3856	.2880	.2880	.2880	.2880	.2880	.2880
.01	.1870	.7263	.6380	.6380	.6380	.6380	.6380	.6380
.02	.3620	.8586	.7931	.7931	.7931	.7931	.7931	.7931
.05	.7206	.9590	.9227	.9227	.9227	.9227	.9227	.9227
.08	.8450	.9836	.9616	.9616	.9616	.9616	.9616	.9616
				$k = 4$				
.002	.1040	.3999	.2425	.2425	.2425	.2425	.2425	.2425
.01	.1758	.7237	.5880	.5881	.5880	.5880	.5880	.5880
.02	.3098	.8543	.7427	.7427	.7426	.7426	.7426	.7426
.05	.6705	.9607	.9019	.9019	.9019	.9019	.9019	.9019
.08	.8025	.9840	.9472	.9472	.9472	.9472	.9472	.9472

Note. An *F* statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). k = number of

groups. f^2 = modified effect size (Aguinis et al., 2005). For all tests, the means were based on 21 conditions.

Table 15

Average Empirical Power (at α = .05) as a Function of ${}_kP_j$ and Pairing when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists

${}_kP_j$s	Pairing	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
Moderately Unequal	Direct	.3722	.8204	.7429	.7403	.7417	.7407	.7376	.7286
Moderately Unequal	Indirect	.4679	.7809	.6912	.6888	.6900	.6888	.6858	.6757
Very Unequal	Direct	.3419	.8531	.7733	.7669	.7704	.7669	.7577	.7631
Very Unequal	Indirect	.5179	.7853	.6714	.6626	.6674	.6626	.6507	.6549

Note. An F statistic was used to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). ${}_kP_j$ = degree of disproportionate subgroup sample sizes. For all tests, averages were based on 210 conditions except for F_{WLS*} when ${}_kP_j$s were Very Unequal which was based on 195 conditions.

Table 16

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists with Direct Pairing (N = 48, k = 3)

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
$n_j = 24, 12, 12;\ \sigma_{e_1}^2 = 4$								
.002	.023	.281	.070	.067	.070	.067	.059	.050
.01	.034	.378	.116	.114	.115	.114	.110	.099
.02	.049	.480	.192	.181	.187	.181	.172	.159
.05	.145	.726	.415	.410	.413	.410	.392	.368
.08	.251	.855	.608	.588	.597	.588	.567	.538
$n_j = 32, 8, 8;\ \sigma_{e_1}^2 = 4$								
.002	.007	.391	.093	.081	.085	.081	.062	.035
.01	.014	.518	.196	.171	.186	.171	.135	.070
.02	.026	.621	.294	.260	.280	.259	.210	.107
.05	.110	.830	.555	.517	.537	.517	.439	.261
.08	.215	.926	.718	.672	.697	.672	.606	.412
$n_j = 24, 12, 12;\ \sigma_{e_1}^2 = 16$								
.002	.018	.389	.104	.095	.098	.095	.091	.086
.01	.024	.591	.289	.277	.280	.277	.263	.244
.02	.034	.765	.459	.435	.446	.435	.412	.392

.05	.110	.946	.831	.810	.816	.810	.806	.771
.08	.215	.987	.951	.942	.943	.942	.931	.927
				$n_j = 32, 8, 8;\ \sigma_{e_l}^2 = 16$				
.002	.003	.446	.148	.132	.141	.132	.103	.055
.01	.006	.710	.345	.310	.338	.310	.252	.160
.02	.011	.850	.584	.543	.563	.543	.475	.310
.05	.065	.976	.879	.861	.872	.861	.820	.681
.08	.158	.995	.969	.958	.965	.958	.940	.882
				$n_j = 24, 12, 12;\ \sigma_{e_l}^2 = 64$				
.002	.020	.554	.196	.188	.192	.188	.175	.159
.01	.028	.932	.715	.689	.702	.689	.668	.642
.02	.059	.987	.938	.932	.936	.932	.922	.902
.05	.111	.998	.997	.997	.997	.997	.996	.995
.08	.192	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				$n_j = 32, 8, 8;\ \sigma_{e_l}^2 = 64$				
.002	.003	.637	.289	.253	.275	.253	.190	.108
.01	.004	.943	.795	.752	.776	.752	.676	.516
.02	.016	.987	.954	.940	.945	.940	.912	.813
.05	.060	1.000	.998	.997	.998	.997	.992	.981
.08	.152	1.000	1.000	1.000	1.000	1.000	.999	.995

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_{\mathrm{O}}w_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Table 17

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists with Direct Pairing (N = 96, k = 3)

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
$n_j = 48, 24, 24;\ \sigma_{e_1}^2 = 4$								
.002	.026	.198	.078	.077	.077	.077	.074	.070
.01	.060	.385	.254	.247	.252	.247	.239	.236
.02	.123	.576	.365	.362	.364	.362	.361	.350
.05	.385	.879	.739	.727	.734	.727	.721	.711
.08	.628	.963	.896	.895	.895	.895	.892	.886
$n_j = 64, 16, 16;\ \sigma_{e_1}^2 = 4$								
.002	.009	.281	.115	.105	.109	.105	.094	.087
.01	.035	.494	.255	.241	.250	.241	.220	.201
.02	.082	.707	.462	.442	.456	.442	.417	.395
.05	.352	.919	.819	.806	.813	.806	.785	.758
.08	.625	.981	.945	.932	.941	.932	.925	.907
$n_j = 48, 24, 24;\ \sigma_{e_1}^2 = 16$								
.002	.024	.306	.156	.154	.155	.154	.150	.145
.01	.045	.718	.515	.510	.510	.510	.501	.498
.02	.100	.925	.835	.832	.833	.832	.826	.821

.05	.338	.997	.992	.992	.992	.992	.992	.991
.08	.618	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				n_j = 64, 16, 16; $\sigma_{e_l}^2$ =16				
.002	.001	.381	.180	.163	.168	.163	.146	.134
.01	.011	.796	.615	.584	.597	.584	.558	.528
.02	.042	.945	.866	.855	.861	.855	.837	.815
.05	.289	1.000	.998	.998	.998	.998	.995	.994
.08	.612	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				n_j = 48, 24, 24; $\sigma_{e_l}^2$ = 64				
.002	.019	.593	.398	.384	.389	.384	.376	.368
.01	.046	.991	.977	.973	.977	.973	.968	.966
.02	.087	1.000	.998	.998	.998	.998	.998	.998
.05	.320	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.635	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				n_j = 64, 16, 16; $\sigma_{e_l}^2$ = 64				
.002	.002	.707	.477	.448	.465	.448	.422	.368
.01	.009	.995	.973	.972	.973	.972	.966	.955
.02	.034	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.05	.262	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.607	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Table 18

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists with Direct Pairing (N = 144, k = 3)

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
$n_j = 72, 36, 36;\ \sigma_{e_1}^2 = 4$								
.002	.033	.183	.109	.110	.109	.110	.107	.107
.01	.089	.461	.304	.301	.301	.301	.297	.294
.02	.215	.640	.534	.531	.534	.531	.527	.520
.05	.628	.939	.900	.898	.900	.898	.895	.895
.08	.881	.992	.988	.988	.988	.988	.987	.987
$n_j = 96, 24, 24;\ \sigma_{e_1}^2 = 4$								
.002	.007	.227	.109	.101	.105	.101	.098	.091
.01	.063	.519	.368	.357	.363	.357	.344	.332
.02	.165	.757	.638	.622	.630	.622	.612	.587
.05	.620	.980	.944	.935	.940	.935	.931	.925
.08	.869	.996	.993	.992	.992	.992	.992	.991
$n_j = 72, 36, 36;\ \sigma_{e_1}^2 = 16$								
.002	.025	.326	.192	.187	.189	.187	.183	.181
.01	.052	.811	.685	.683	.683	.683	.680	.674
.02	.145	.976	.944	.943	.943	.943	.941	.940

.05	.638	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.902	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n_j = 96, 24, 24;\ \sigma_{e_l}^2 = 16$								
.002	.003	.394	.242	.235	.235	.235	.233	.209
.01	.024	.885	.788	.784	.785	.784	.765	.747
.02	.102	.985	.975	.973	.974	.973	.965	.959
.05	.604	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.917	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n_j = 72, 36, 36;\ \sigma_{e_l}^2 = 64$								
.002	.022	.693	.573	.568	.572	.568	.560	.549
.01	.063	.998	.996	.996	.996	.996	.995	.994
.02	.151	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.05	.587	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.944	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n_j = 96, 24, 24;\ \sigma_{e_l}^2 = 64$								
.002	.003	.796	.658	.651	.657	.651	.635	.620
.01	.023	1.000	.999	.999	.999	.999	.999	.998
.02	.088	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.05	.596	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.927	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Table 19

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists with Direct Pairing (N = 192, k = 3)

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
$n_j = 96, 48, 48;\ \sigma_{e_1}^2 = 4$								
.002	.024	.178	.101	.100	.101	.100	.099	.094
.01	.117	.488	.384	.381	.383	.381	.376	.371
.02	.326	.760	.689	.686	.688	.685	.681	.679
.05	.816	.979	.965	.965	.965	.965	.964	.964
.08	.963	.998	.997	.997	.997	.997	.997	.997
$n_j = 128, 32, 32;\ \sigma_{e_1}^2 = 4$								
.002	.013	.210	.126	.124	.124	.124	.120	.118
.01	.098	.590	.465	.457	.458	.457	.440	.423
.02	.274	.851	.760	.752	.760	.752	.740	.722
.05	.804	.993	.982	.981	.981	.981	.981	.980
.08	.961	1.000	.998	.998	.998	.998	.998	.998
$n_j = 96, 48, 48;\ \sigma_{e_1}^2 = 16$								
.002	.023	.337	.245	.244	.244	.244	.241	.240
.01	.103	.895	.850	.849	.850	.849	.849	.848
.02	.260	.990	.982	.982	.982	.982	.982	.982

.05	.848	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.993	1.000	1.000	1.000	1.000	1.000	1.000	1.000
n_j = 128, 32, 32; $\sigma_{e_l}^2$ = 16								
.002	.007	.423	.285	.278	.283	.278	.268	.256
.01	.050	.940	.895	.889	.892	.888	.879	.875
.02	.197	.996	.994	.994	.994	.994	.994	.994
.05	.850	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
n_j = 96, 48, 48; $\sigma_{e_l}^2$ = 64								
.002	.013	.788	.696	.695	.696	.695	.692	.688
.01	.071	.999	.999	.999	.999	.999	.999	.999
.02	.217	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.05	.871	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000
n_j = 128, 132, 32; $\sigma_{e_l}^2$ = 64								
.002	.003	.878	.803	.796	.800	.796	.790	.780
.01	.042	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.02	.169	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.05	.846	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Panel A

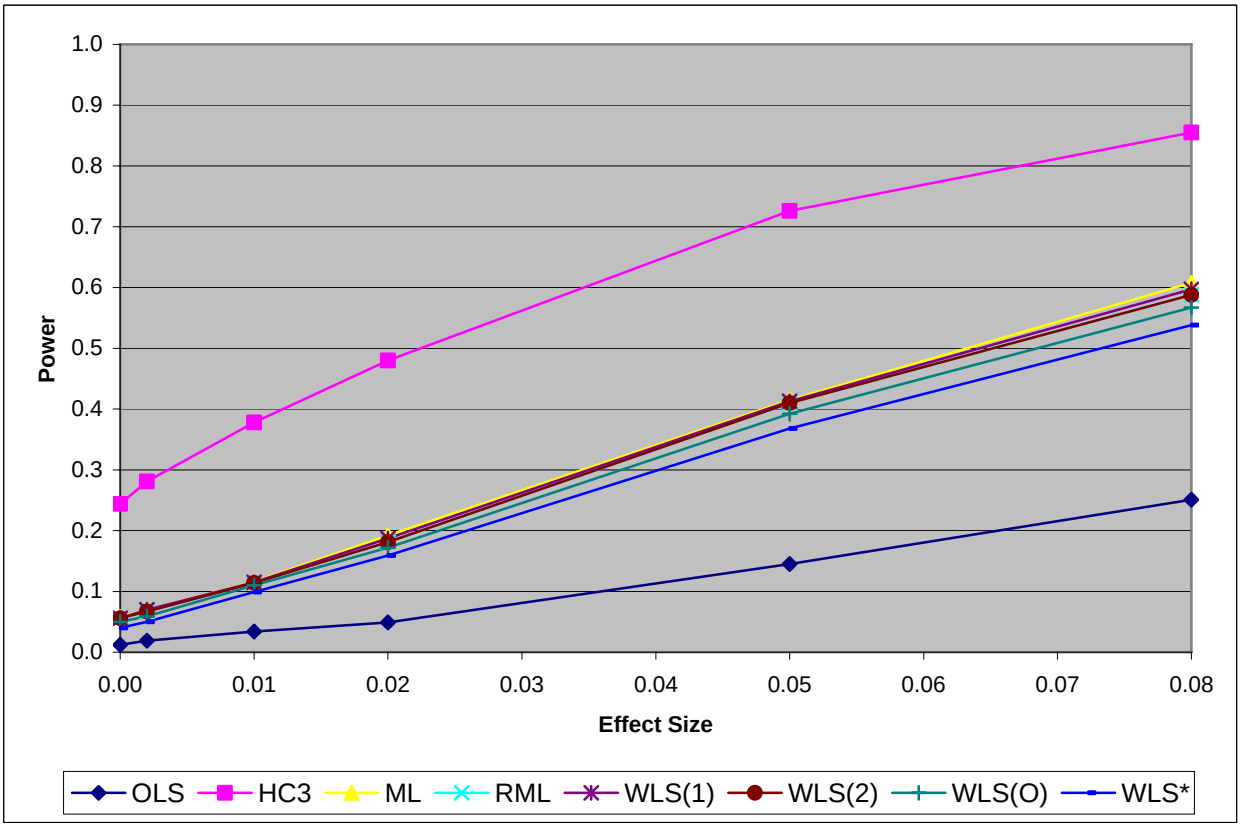
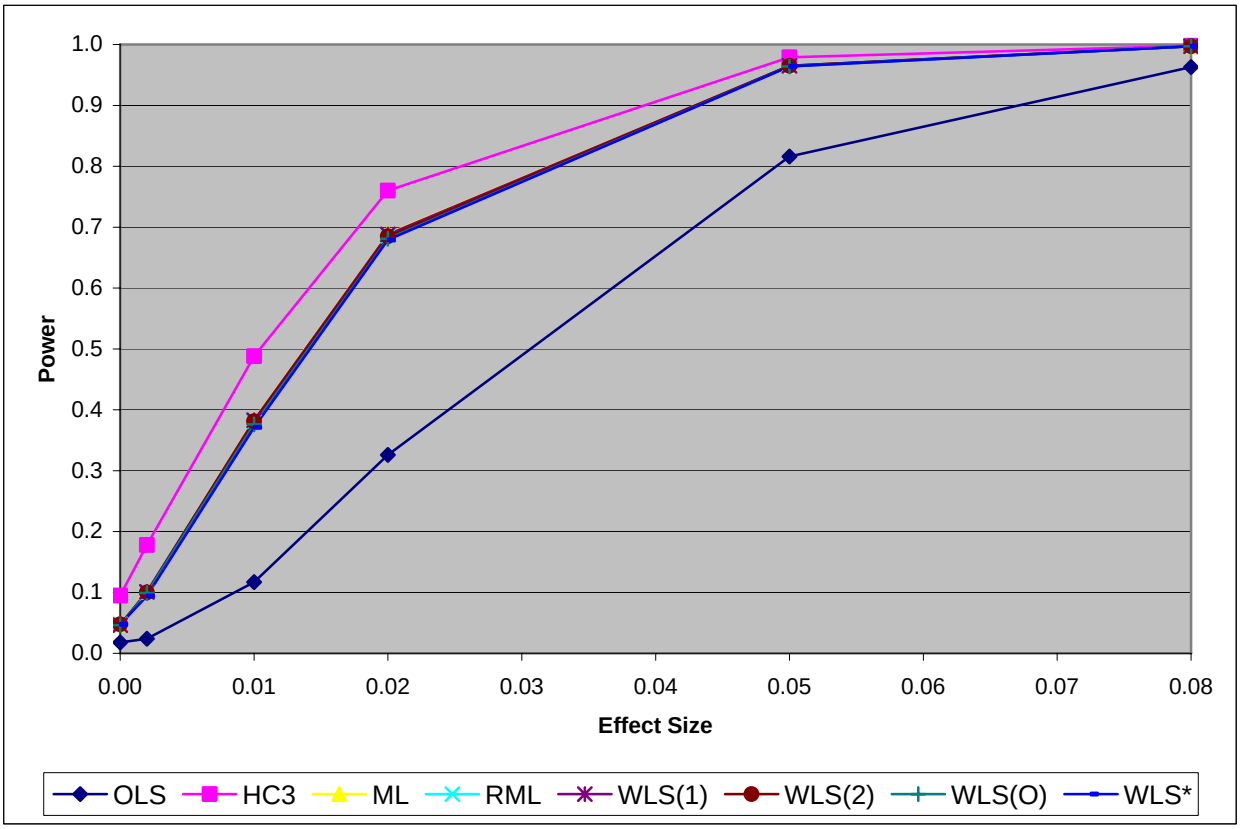
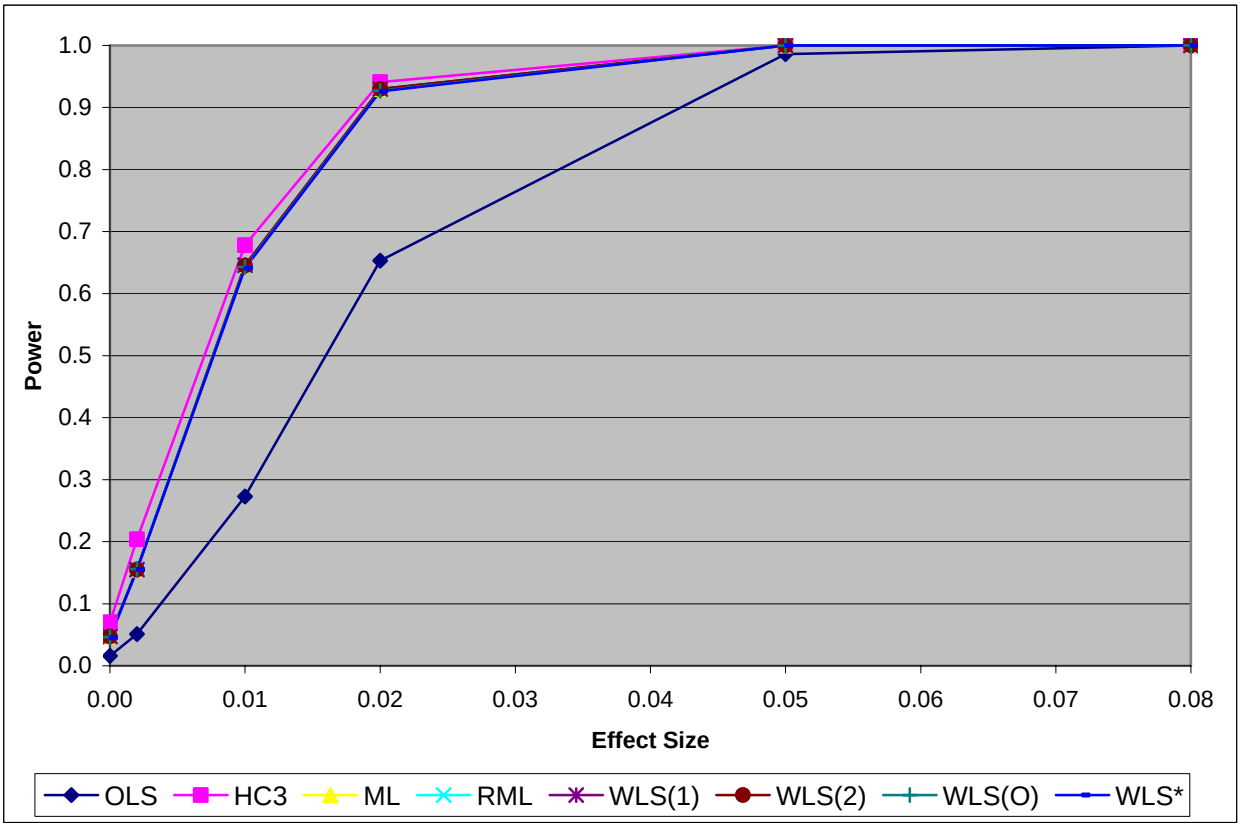


Figure 13. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, direct pairing, $\sigma_{e_j}{}^2$ s = 4, 1, 1, moderately unequal proportions, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

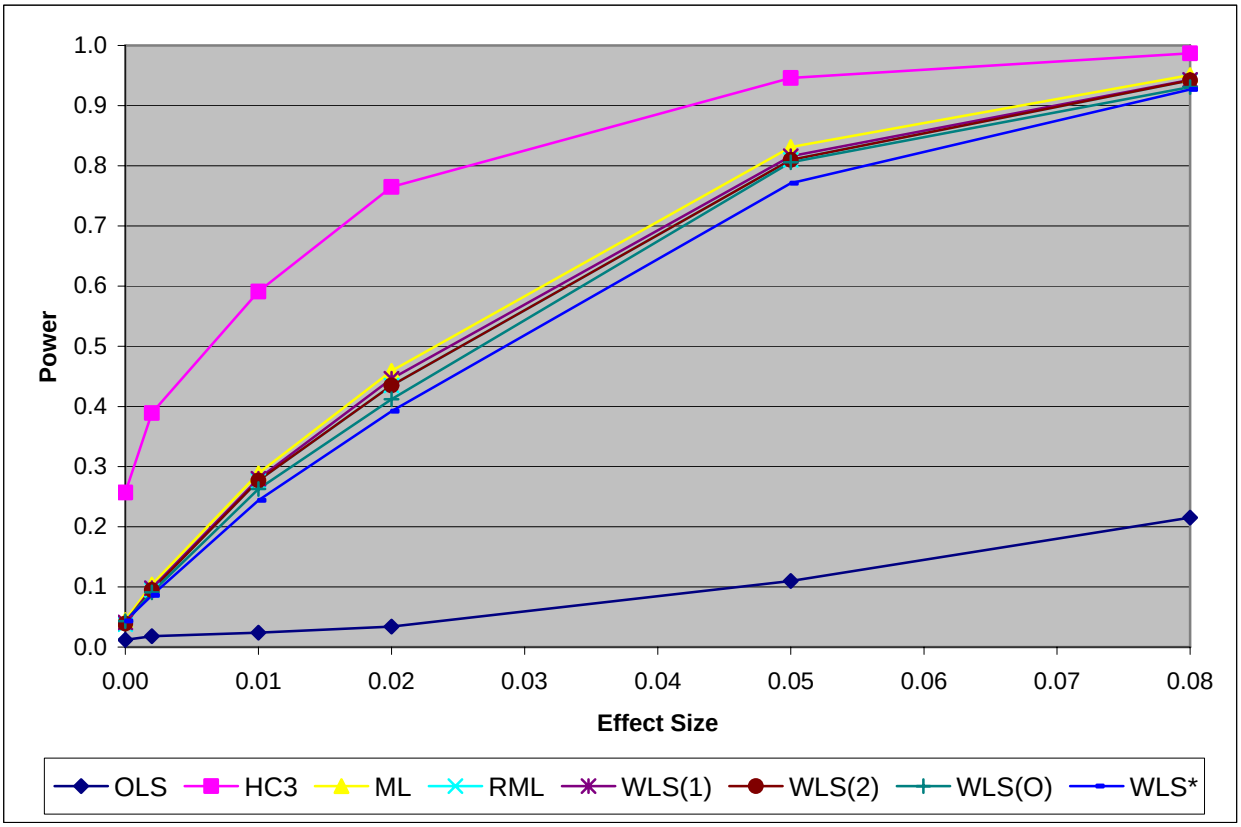
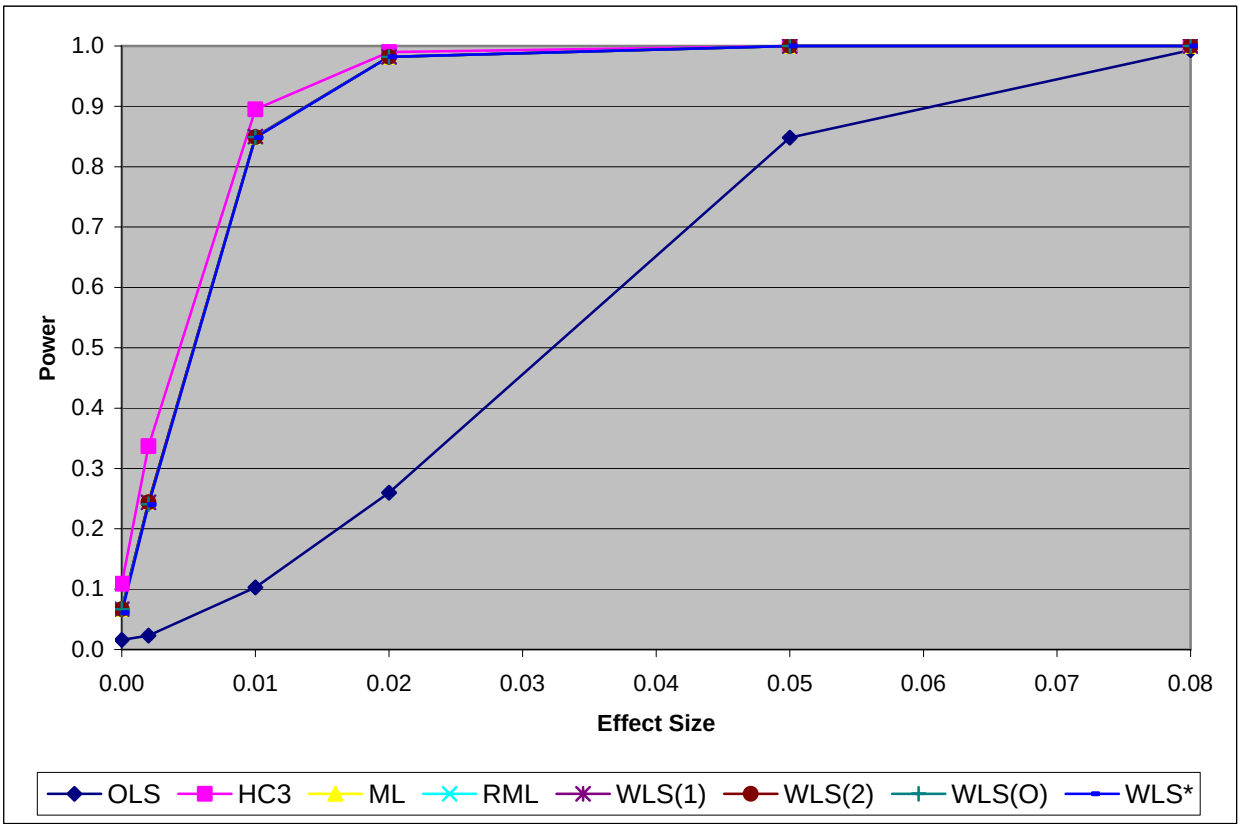
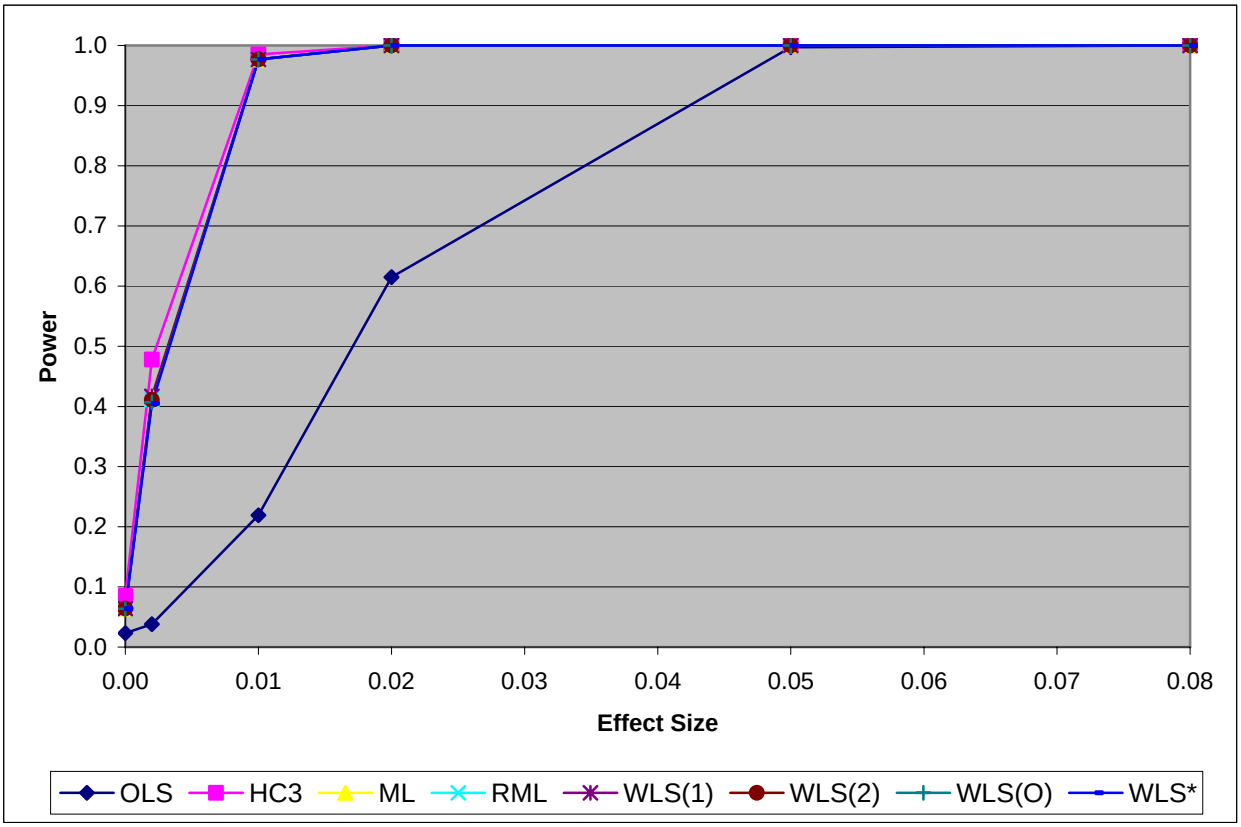


Figure 14. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, direct pairing, $\sigma_{e_j}^2$ s = 16, 1, 1, moderately unequal proportions, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

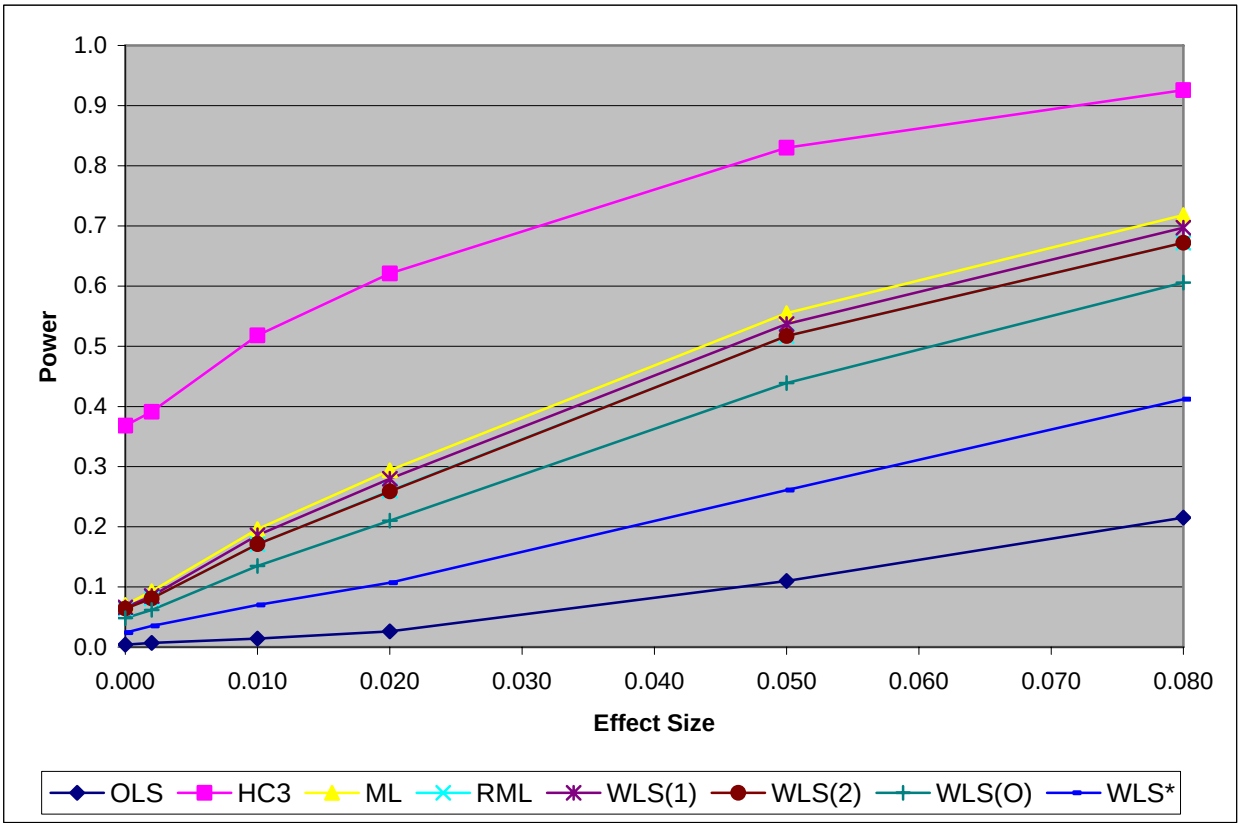
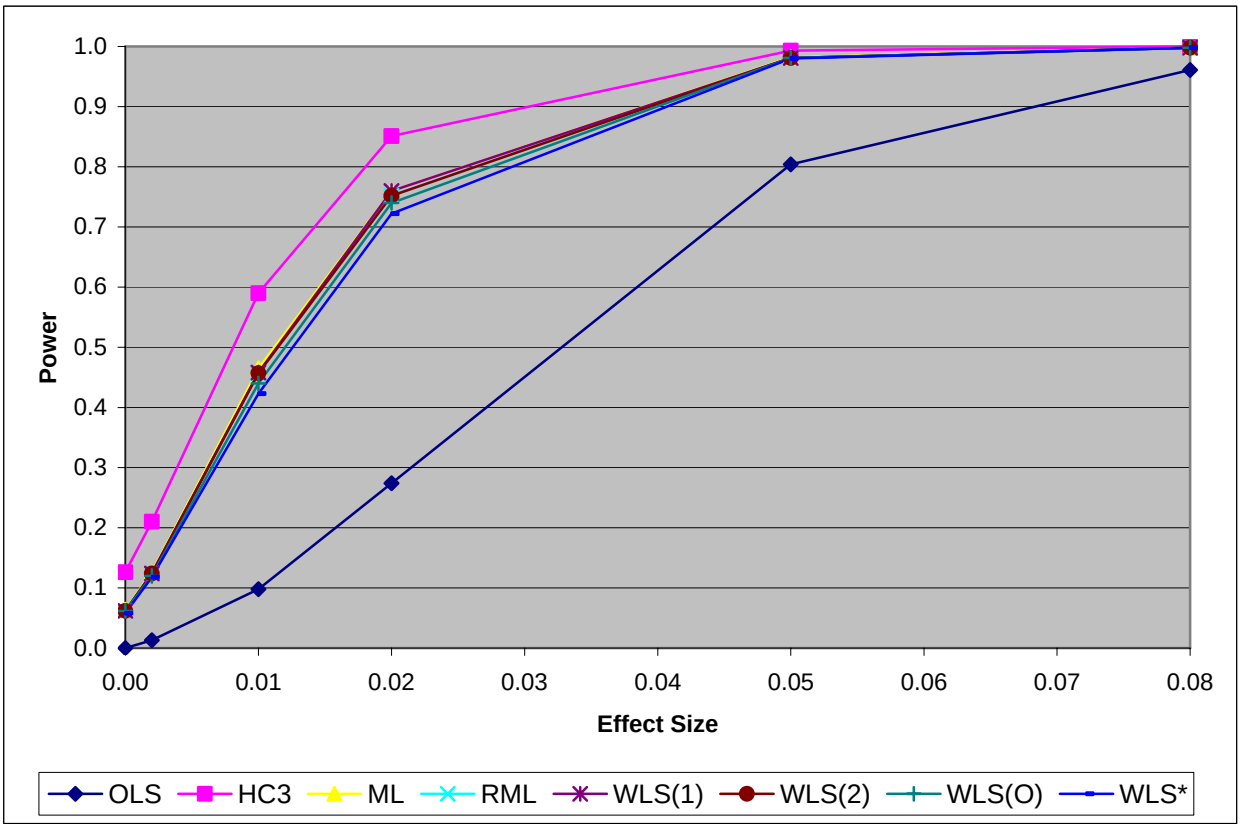
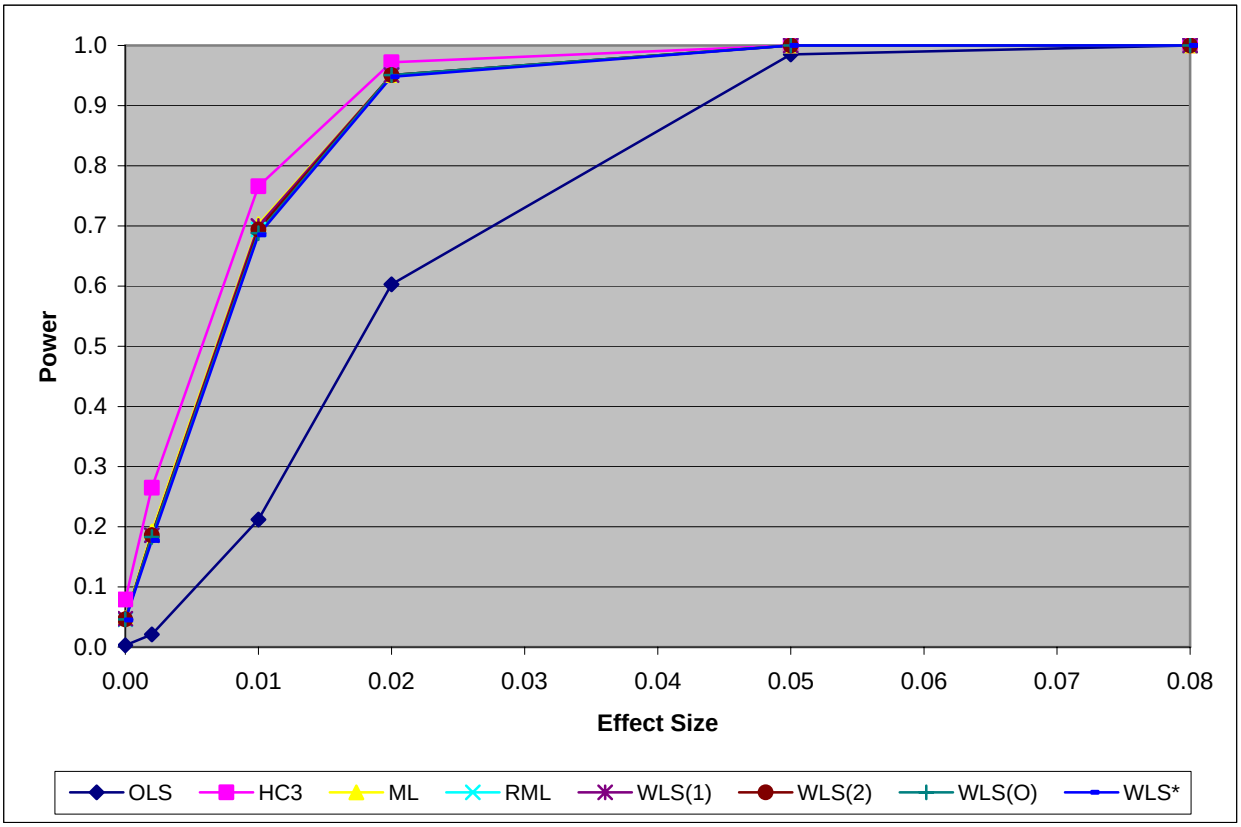


Figure 15. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, direct pairing, $\sigma_{e_j}^2$ s = 4, 1, 1, very unequal proportions, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

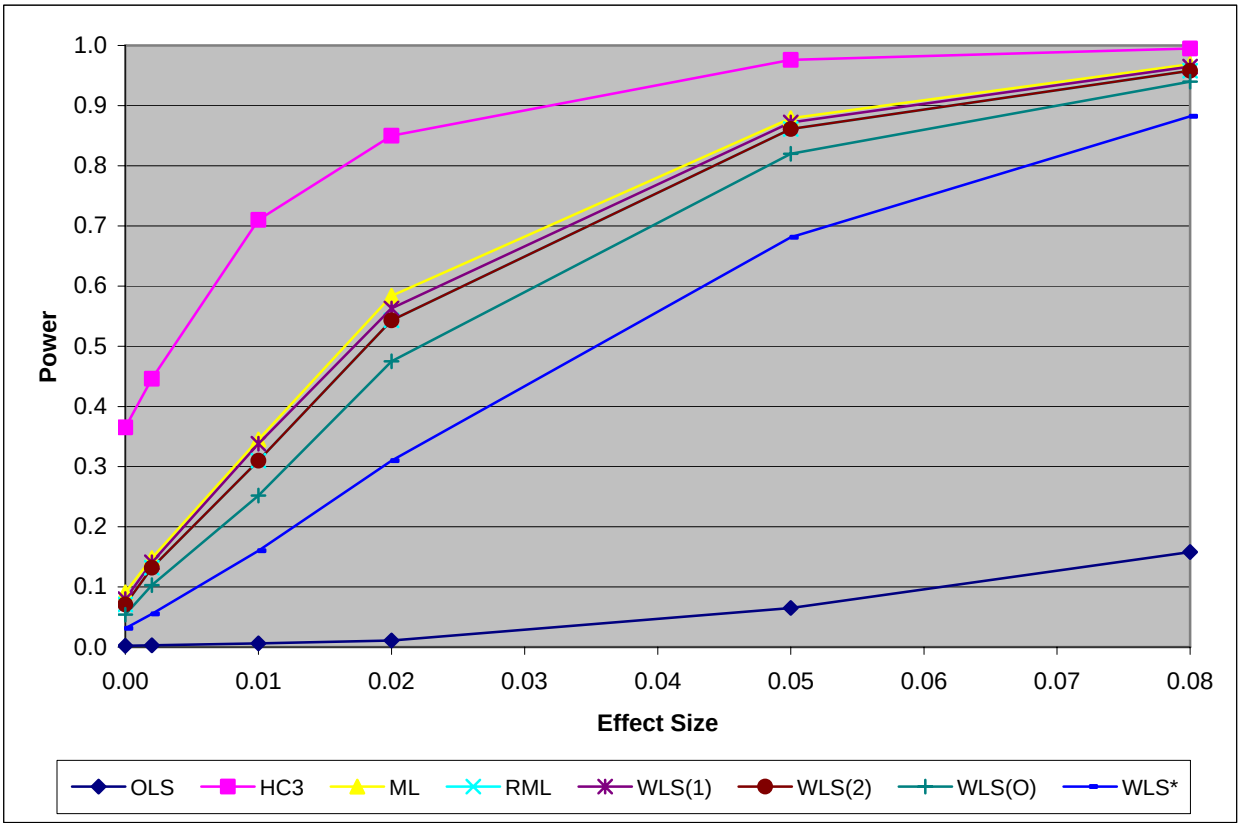
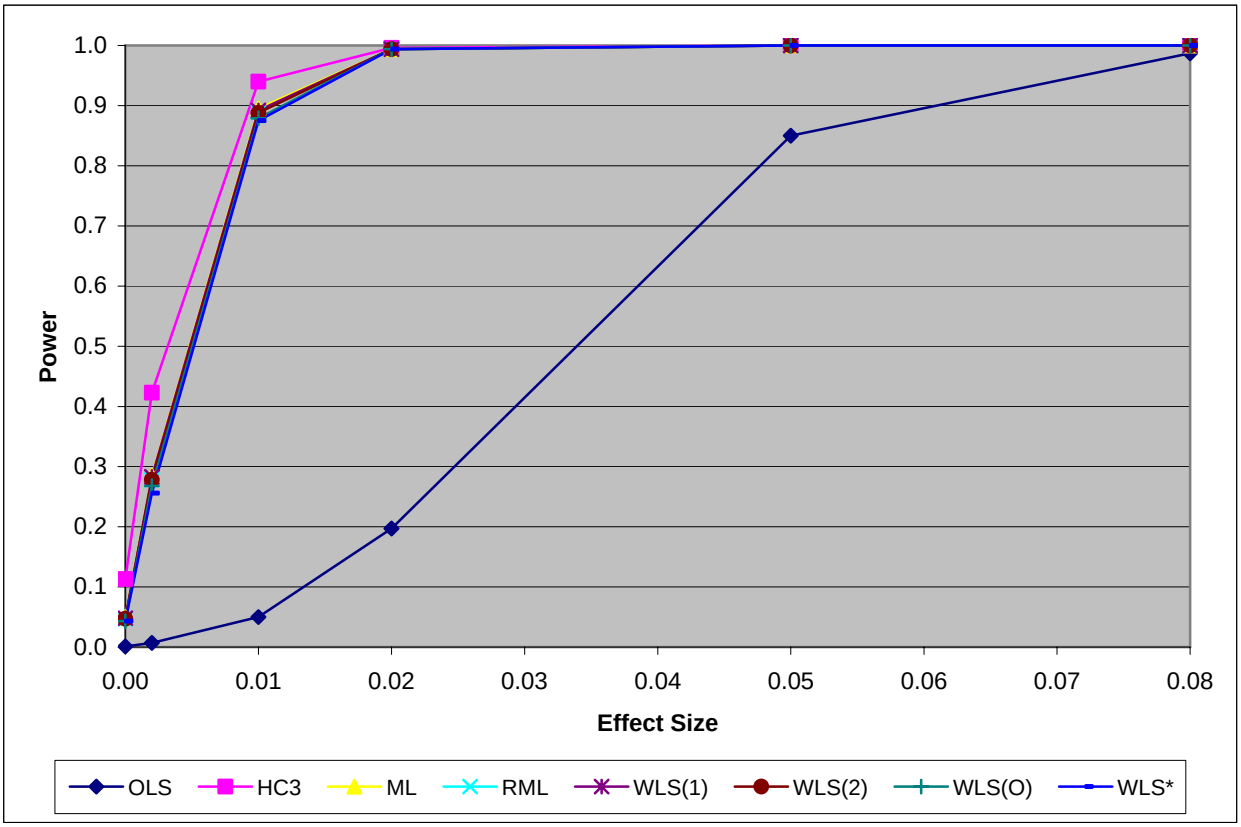
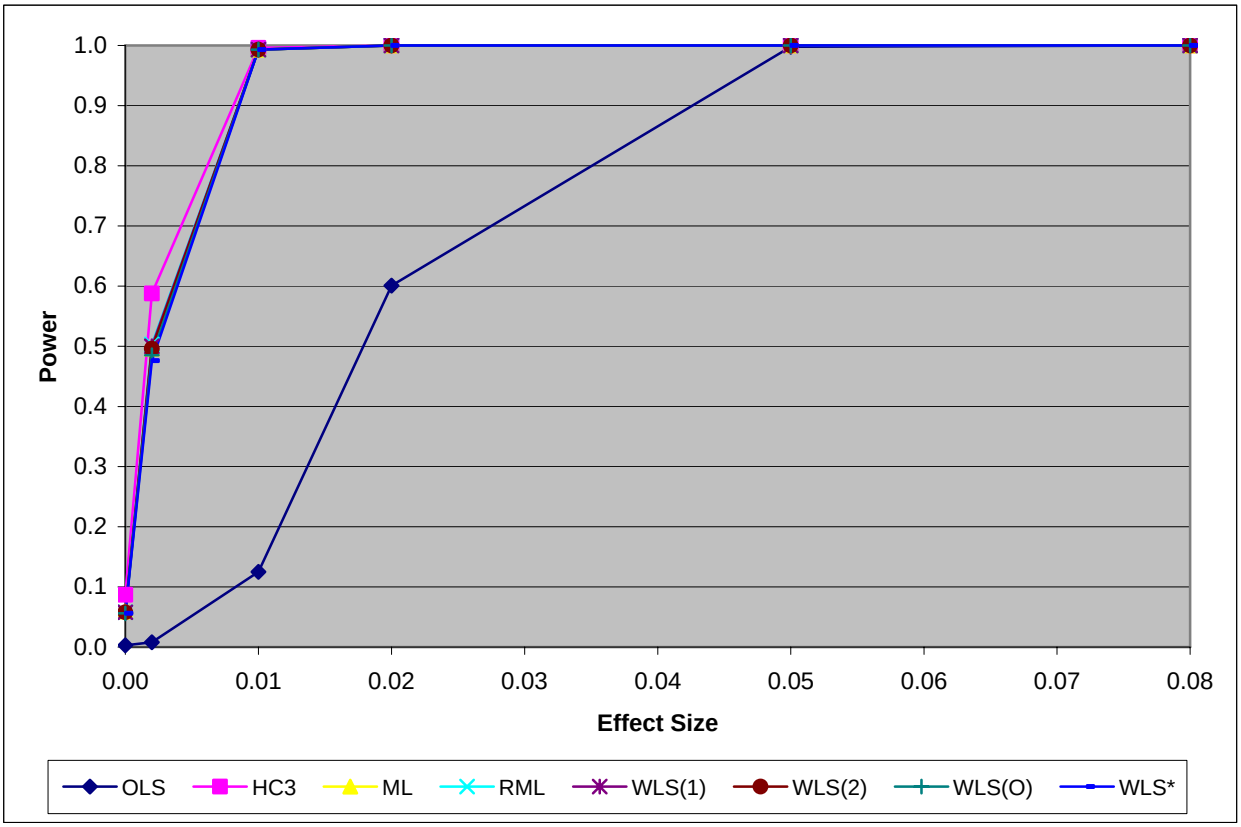


Figure 16. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, direct pairing, $\sigma_{e_j}^2$ s = 16, 1, 1, very unequal proportions, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

In Figure 15A, when $N = 48$, it is apparent that F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ do not have similar power levels as they did with moderately unequal ${}_kP_j$s (see Figure 13A). Note that $F_{WLS(O)}$ and $F_{WLS(*)}$ had power levels lower than that of F_{ML}, F_{RML}, $F_{WLS(1)}$, and $F_{WLS(2)}$. However, as N increased to 192 (Figure 15B) and 336 (Figure 15C), F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had very similar power levels again with $F_{WLS(O)}$ and $F_{WLS(*)}$ having Type I error rates closer to the nominal level than the other methods. The Type I error rates continued to be inflated for F_{HC3} and conservative for F_{OLS}.

Figure 16 is like that of Figure 15, however, with increased heteroscedasticity (i.e., $\sigma_{e_j}^2 =$ 16, 1, 1). It seems that the very dissimilar power levels of F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ that were evident when ${}_kP_j$s were very unequal but with mild heteroscedasticity (i.e., $N =$ 48, $\sigma_{e_j}^2 = 4, 1, 1$ in Figure 15A) are less noticeable in Figure 16A. In addition, the power of F_{OLS} decreased. Note that when N increased to 192 (Figure 16B) and 336 (Figure 16C), F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had very similar power levels again with $F_{WLS(O)}$ and $F_{WLS(*)}$ having Type I error rates closer to the nominal level than the other methods. The Type I error rates continued to be inflated for F_{HC3} and conservative for F_{OLS}.

Indirect pairing. For F_{OLS}., the results were consistent with previous research. Namely, power levels were somewhat greater for F_{OLS}. For example, in Table 20 ($N = 48$), F_{OLS} had power equal to .111 to detect an $f^2 = .002$ when $\sigma_{e_j}^2 = 1, 1, 4$, and n_js = 24, 12, 12. Recall that its Type I error rates were very liberal with indirect pairing, particularly with very unequal ${}_kP_j$s. For example, consider a similar condition in the same table, but with n_js = 32, 8, 8. The power of F_{OLS} increased to .161. For all the alternative methods, power was not always greater than that of F_{OLS}. Because of its inflated Type I error rates, the trends are more complex and will be described using figures below. Although F_{HC3} had the greatest power in most instances with

indirect pairing, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ had power levels that became very similar to that of F_{HC3} when f^2 increased, ${}_kP_j$s were moderately unequal, and as N increased. This can be seen by inspecting Table 21 when $N = 96$, Table 22 when $N = 144$, and Table 23 when $N = 192$.

Figures 17 – 23 depict the power curves for the eight tests across various conditions with indirect pairing. In Figure 17A, $N = 48$, ${}_kP_j$s are moderately unequal, and $\sigma_{e_j}^2 = 1, 1, 4$ (see also Table 20). A very interesting finding can be discerned from this plot that is less apparent in the above-mentioned tables. More precisely, although F_{OLS} had inflated Type I error rates, giving it an illusory power advantage, note that as f^2 increases, this power advantage diminishes. Its power function crosses that of all the other methods until it is again the lowest in rank order. All the other methods performed similar to previous figures with $F_{WLS(O)}$ and $F_{WLS(*)}$ having Type I error rates near the nominal level. These trends also occurred as N increased to 192 (Figure 17B) and 336 (Figure 17C). Note however, in Figure 17C, the Type I error rate for F_{HC3} decreased to .062 and the Type I error rate for F_{OLS} was .082.

The same trends are evident in Figure 18 which is like that of Figure 17, but with increased heteroscedasticity (i.e., $\sigma_{e_j}^2 = 1, 1, 16$). Note that the Type I error rates of F_{OLS} became more inflated (i.e., > .1). Regardless of $N = 48$ (Figure 18A), 192 (Figure 18B), or 336 (Figure 18C), the power function of F_{OLS} always crossed that of the other methods. This was also true for F_{HC3} because the leftmost side of its power function decreased as N increased.

Figure 19 depicts the power of the tests when ${}_kP_j$s were very unequal and $\sigma_{e_j}^2 = 1, 1, 4$. Consistent with previous figures, with small Ns (i.e., 48) when ${}_kP_j$s were very unequal and with moderate levels of heteroscedasticity, the power levels for F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and $F_{WLS(*)}$ differed noticeably (see Figure 19A). Noteworthy, the Type I error rate for F_{OLS} was

greatly inflated (i.e., .152) as was that for F_{HC3} (.432). Again, the power function for F_{OLS} crossed that of the other methods. As N increased to 192 (Figure 19B) and 336 (Figure 19C), F_{OLS} had greater power than some of the methods at small f^2s.

Figure 20 is like that of Figure 19, however, with increased heteroscedasticity (i.e., $\sigma_{e_j}^2$ = 1, 1, 16). The trends are similar across N = 48 (Figure 20A), 192 (Figure 20B), and 336 (Figure 20C). Notably, an inspection of Figure 20A shows that the Type I error rate for F_{OLS} was greatly inflated (i.e., .268) as was that for F_{HC3} (.495). Increasing N to 336 (see Figure 20C) results in a Type I error rate for F_{OLS} and F_{HC3} equal to .218 and .112, respectively. Again, F_{OLS} had greater power than some of the methods at small f^2s.

Comment. Similar relations occurred when k = 4, with the tests ordered in the same manner in terms of power. Additional tables and figures can be obtained from the author. (Text continues on p. 154.)

Table 20

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists with Indirect Pairing (N = 48, k = 3)

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
$n_j = 24, 12, 12;\ \sigma_{e_3}^2 = 4$								
.002	.111	.335	.081	.076	.079	.076	.068	.059
.01	.132	.398	.124	.118	.122	.118	.104	.095
.02	.161	.477	.192	.182	.187	.182	.172	.148
.05	.235	.652	.342	.327	.335	.327	.311	.287
.08	.349	.792	.493	.483	.490	.483	.473	.448
$n_j = 32, 8, 8;\ \sigma_{e_3}^2 = 4$								
.002	.161	.468	.125	.098	.117	.100	.072	.029
.01	.184	.532	.162	.140	.158	.140	.104	.044
.02	.226	.602	.230	.194	.215	.194	.152	.070
.05	.311	.762	.380	.332	.360	.332	.266	.138
.08	.409	.848	.528	.459	.482	.459	.382	.230
$n_j = 24, 12, 12;\ \sigma_{e_3}^2 = 16$								
.002	.160	.353	.109	.106	.108	.106	.098	.081
.01	.167	.531	.226	.217	.224	.217	.199	.179
.02	.189	.659	.360	.357	.360	.357	.342	.318

.05	.292	.907	.699	.695	.699	.695	.683	.665
.08	.370	.961	.851	.846	.847	.846	.840	.829
$n_j = 32, 8, 8;\ \sigma_{e_3}^2 = 16$								
.002	.274	.520	.142	.120	.135	.120	.086	.035
.01	.296	.634	.229	.206	.220	.206	.165	.082
.02	.315	.730	.354	.313	.328	.313	.235	.137
.05	.404	.877	.616	.573	.602	.573	.496	.313
.08	.501	.946	.778	.740	.761	.740	.671	.472
$n_j = 24, 12, 12;\ \sigma_{e_3}^2 = 64$								
.002	.174	.465	.181	.168	.171	.168	.156	.142
.01	.189	.823	.546	.532	.538	.532	.514	.490
.02	.212	.946	.822	.817	.819	.817	.806	.783
.05	.294	.997	.987	.987	.987	.987	.984	.980
.08	.388	.998	.998	.998	.998	.998	.997	.996
$n_j = 32, 8, 8;\ \sigma_{e_3}^2 = 64$								
.002	.310	.605	.192	.166	.181	.166	.115	.052
.01	.324	.802	.469	.431	.446	.431	.348	.191
.02	.362	.921	.727	.664	.706	.664	.595	.389
.05	.437	.987	.927	.905	.912	.905	.878	.758
.08	.547	.998	.976	.968	.975	.968	.948	.885

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_{\mathrm{O}}w_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Table 21

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists with Indirect Pairing (N = 96, k = 3)

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
$n_j = 48, 24, 24;\ \sigma_{e_3}^2 = 4$								
.002	.105	.211	.083	.078	.080	.078	.074	.071
.01	.158	.352	.186	.178	.183	.178	.174	.166
.02	.236	.521	.304	.296	.299	.296	.289	.286
.05	.489	.797	.634	.630	.632	.630	.625	.617
.08	.706	.914	.858	.857	.858	.857	.851	.848
$n_j = 64, 16, 16;\ \sigma_{e_3}^2 = 4$								
.002	.166	.293	.101	.089	.096	.089	.078	.061
.01	.205	.405	.185	.172	.182	.172	.158	.138
.02	.295	.534	.292	.270	.283	.270	.260	.239
.05	.518	.806	.608	.586	.598	.586	.552	.514
.08	.725	.925	.806	.785	.796	.785	.772	.742
$n_j = 48, 24, 24;\ \sigma_{e_3}^2 = 16$								
.002	.154	.286	.120	.118	.119	.118	.113	.109
.01	.182	.540	.381	.374	.377	.374	.367	.361
.02	.262	.808	.668	.665	.665	.665	.661	.651

.05	.481	.981	.962	.959	.961	.959	.955	.952
.08	.731	1.000	.994	.992	.992	.992	.992	.991
				$n_j = 64, 16, 16;\ \sigma_{e_3}^2 = 16$				
.002	.258	.338	.146	.130	.138	.130	.113	.092
.01	.302	.567	.338	.321	.328	.321	.282	.264
.02	.353	.766	.528	.510	.515	.510	.478	.461
.05	.581	.967	.866	.851	.861	.851	.840	.831
.08	.795	.992	.975	.972	.975	.972	.967	.960
				$n_j = 48, 24, 24;\ \sigma_{e_3}^2 = 64$				
.002	.179	.482	.288	.281	.282	.280	.276	.272
.01	.193	.938	.873	.872	.873	.872	.868	.861
.02	.254	.996	.989	.989	.989	.989	.988	.988
.05	.483	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.758	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				$n_j = 64, 16, 16;\ \sigma_{e_3}^2 = 64$				
.002	.289	.499	.263	.237	.251	.237	.213	.187
.01	.324	.892	.755	.732	.746	.732	.712	.691
.02	.382	.983	.941	.937	.941	.937	.928	.912
.05	.602	1.000	.997	.996	.996	.996	.996	.995
.08	.831	1.000	1.000	1.000	1.000	1.000	1.000	.998

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_{\mathrm{O}}w_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_1}^2$ denotes population error variance in Group 1; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Table 22

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists with Indirect Pairing (N = 144, k = 3)

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
$n_j = 72, 36, 36;\ \sigma_{e_3}^2 = 4$								
.002	.112	.172	.099	.096	.097	.096	.094	.083
.01	.195	.356	.229	.226	.227	.226	.224	.218
.02	.318	.571	.437	.435	.435	.435	.431	.425
.05	.667	.885	.830	.826	.828	.826	.825	.822
.08	.882	.978	.956	.956	.956	.956	.956	.954
$n_j = 96, 24, 24;\ \sigma_{e_3}^2 = 4$								
.002	.164	.211	.092	.085	.091	.085	.080	.072
.01	.258	.421	.256	.242	.250	.242	.229	.214
.02	.379	.622	.452	.437	.446	.437	.399	.381
.05	.689	.885	.780	.768	.775	.768	.757	.742
.08	.875	.964	.923	.922	.923	.922	.916	.910
$n_j = 72, 36, 36;\ \sigma_{e_3}^2 = 16$								
.002	.155	.258	.152	.151	.152	.151	.150	.147
.01	.227	.679	.529	.524	.527	.524	.521	.515
.02	.331	.920	.874	.871	.871	.871	.866	.866

.05	.720	1.000	.996	.996	.996	.996	.996	.996
.08	.955	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				n_j = 96, 24, 24; $\sigma_{e_3}^2$ = 16				
.002	.246	.272	.139	.124	.134	.124	.116	.112
.01	.312	.639	.466	.450	.461	.450	.427	.409
.02	.439	.876	.766	.748	.760	.748	.730	.718
.05	.789	.990	.975	.973	.973	.973	.972	.969
.08	.936	.998	.998	.997	.998	.997	.997	.996
				n_j = 72, 36, 36; $\sigma_{e_3}^2$ = 64				
.002	.164	.537	.396	.395	.396	.395	.393	.389
.01	.218	.982	.973	.973	.973	.973	.971	.971
.02	.327	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.05	.764	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.970	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				n_j = 96, 24, 24; $\sigma_{e_3}^2$ = 64				
.002	.281	.492	.322	.315	.320	.314	.300	.274
.01	.341	.962	.907	.896	.900	.896	.892	.879
.02	.448	.995	.994	.993	.994	.993	.992	.991
.05	.812	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.951	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_3}^2$ denotes population error variance in Group 3; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Table 23

Empirical Power (at α = .05) when Testing for the Equality of Regression Slopes when Heteroscedasticity Exists with Indirect Pairing (N = 192, k = 3)

f^2	F_{OLS}	F_{HC3}	F_{ML}	F_{RML}	$F_{WLS(1)}$	$F_{WLS(2)}$	$F_{WLS(O)}$	F_{WLS*}
$n_j = 96, 48, 48;\ \sigma_{e_3}^2 = 4$								
.002	.116	.165	.100	.097	.099	.097	.095	.095
.01	.220	.409	.301	.296	.297	.296	.294	.291
.02	.447	.644	.589	.580	.585	.580	.579	.576
.05	.841	.952	.930	.928	.930	.928	.926	.925
.08	.974	1.000	.997	.997	.997	.997	.995	.995
$n_j = 128, 32, 32;\ \sigma_{e_3}^2 = 4$								
.002	.180	.215	.110	.105	.108	.105	.096	.086
.01	.329	.468	.327	.318	.323	.318	.302	.291
.02	.479	.669	.533	.522	.527	.522	.511	.492
.05	.840	.944	.895	.888	.892	.888	.882	.872
.08	.954	.988	.980	.980	.980	.980	.977	.973
$n_j = 96, 48, 48;\ \sigma_{e_3}^2 = 16$								
.002	.165	.277	.187	.186	.187	.186	.184	.178
.01	.261	.779	.704	.701	.704	.701	.689	.687
.02	.432	.968	.933	.931	.933	.931	.929	.929

.05	.895	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.993	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				n_j = 128, 32, 32; $\sigma_{e_3}^2$ = 16				
.002	.261	.316	.194	.182	.187	.182	.173	.164
.01	.361	.737	.597	.589	.593	.589	.578	.565
.02	.515	.929	.878	.868	.873	.868	.863	.856
.05	.896	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.985	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				n_j = 96, 48, 48; $\sigma_{e_3}^2$ = 64				
.002	.187	.654	.552	.548	.550	.548	.544	.187
.01	.242	.998	.997	.997	.997	.997	.997	.242
.02	.405	1.000	1.000	1.000	1.000	1.000	1.000	.405
.05	.932	1.000	1.000	1.000	1.000	1.000	1.000	.932
.08	.998	1.000	1.000	1.000	1.000	1.000	1.000	.998
				n_j = 128, 32, 32; $\sigma_{e_3}^2$ = 64				
.002	.286	.592	.441	.431	.435	.431	.413	.401
.01	.384	.985	.971	.967	.970	.967	.966	.964
.02	.542	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.05	.938	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.08	.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Rejection rates based on 1,000 replications per condition using an F to test for the equality of regression slopes with ordinary least squares (F_{OLS}), HC3 (F_{HC3}), maximum likelihood (F_{ML}), restricted maximum likelihood (F_{RML}), weighted least squares with ${}_1w_j$ ($F_{WLS(1)}$), weighted least squares with ${}_2w_j$ ($F_{WLS(2)}$), weighted least squares with ${}_Ow_j$ ($F_{WLS(O)}$), and weighted least squares with w_j^* (F_{WLS*}). For all conditions, population standard deviations on x equal 1.5. N denotes total sample size and, in each of the k groups, n_j denotes the subgroup sample size. $\sigma_{e_3}^2$ denotes population error variance in Group 3; error variance in each of the remaining groups equals 1.0. When $k = 3$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3. When $k = 4$, the population correlation coefficient and standard deviation on y equal .6 and 1.25, respectively, in Group 3 and 4. When $\sigma_{e_1}^2 = 4$, $\sigma_{e_1}^2 = 16$, and $\sigma_{e_1}^2 = 64$, respectively, the population correlation coefficient in Group 1 equals .351123, .184288, and .09334. For these values and a given effect size (f^2), the slope in Group 2 was computed.

Panel A

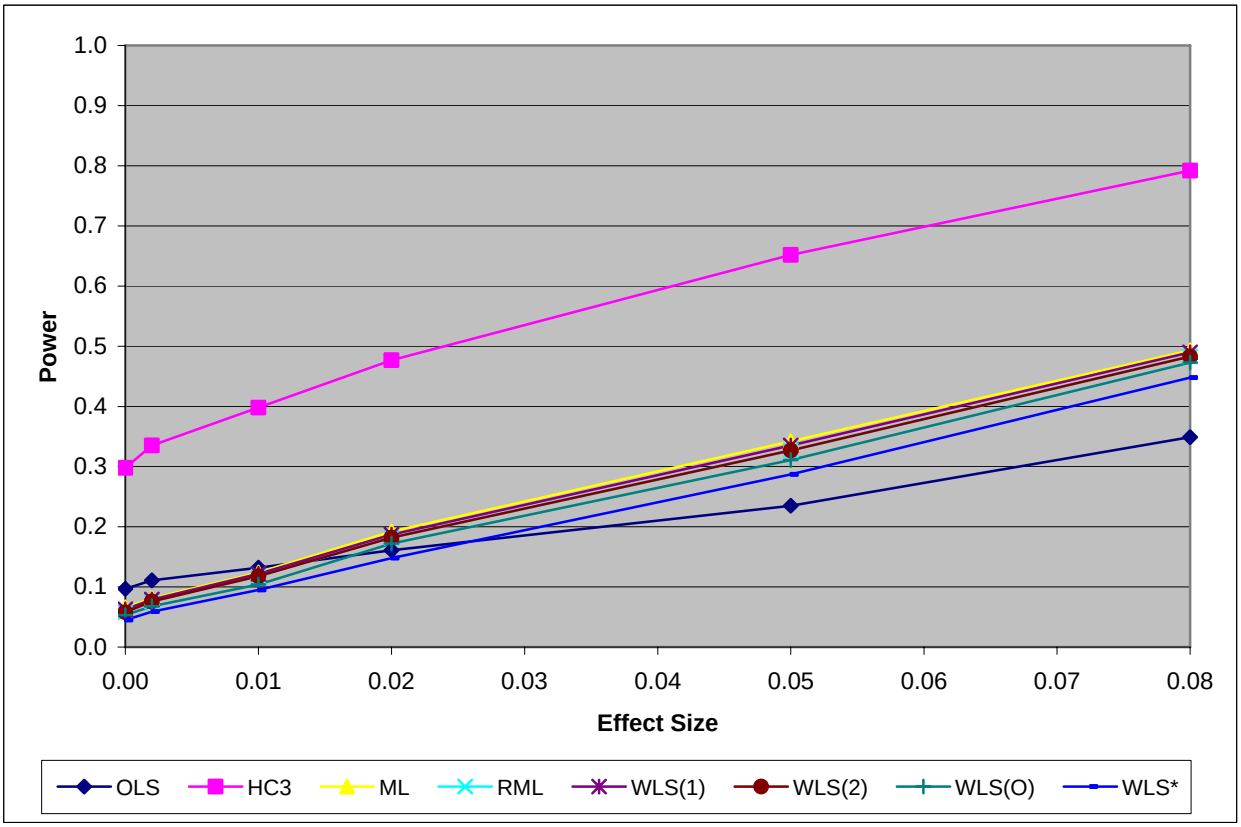
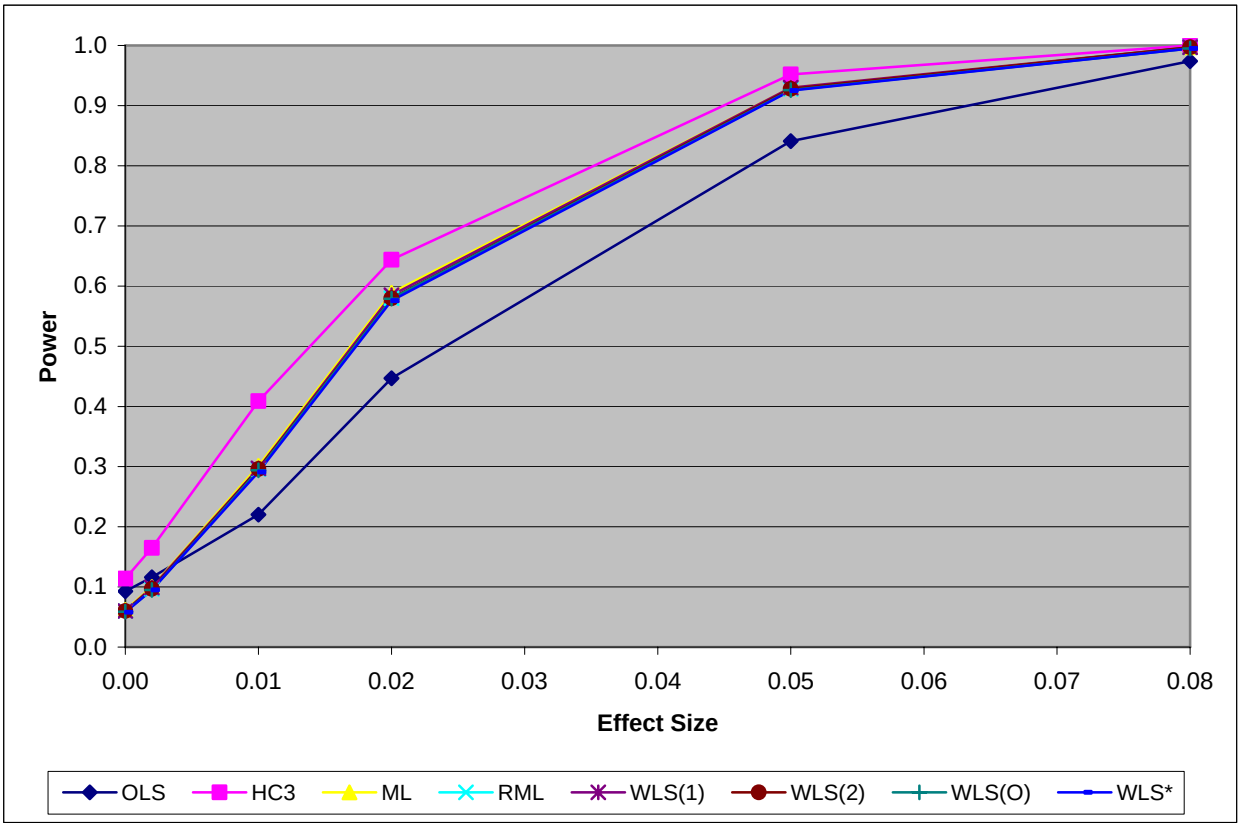
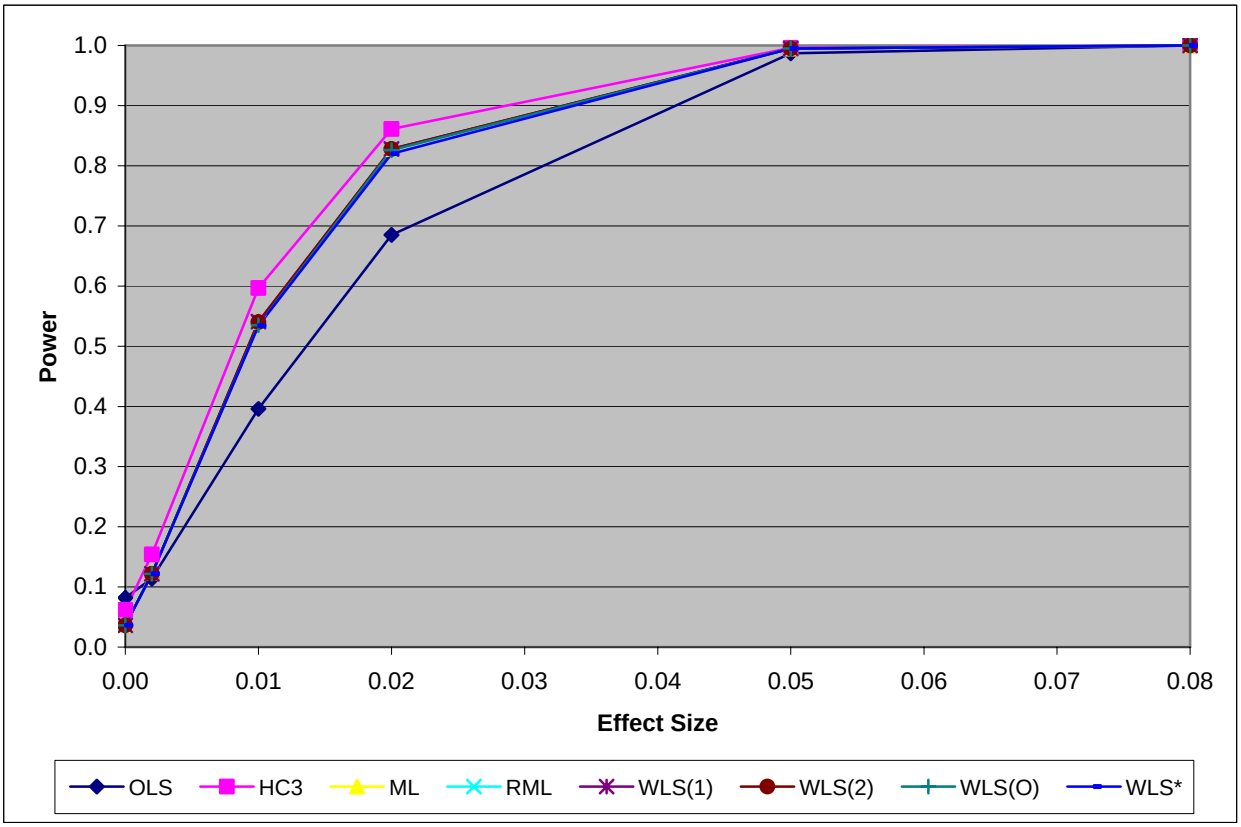


Figure 17. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, indirect pairing, $\sigma_{e_j}^2$ s = 1, 1, 4, moderately unequal proportions, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

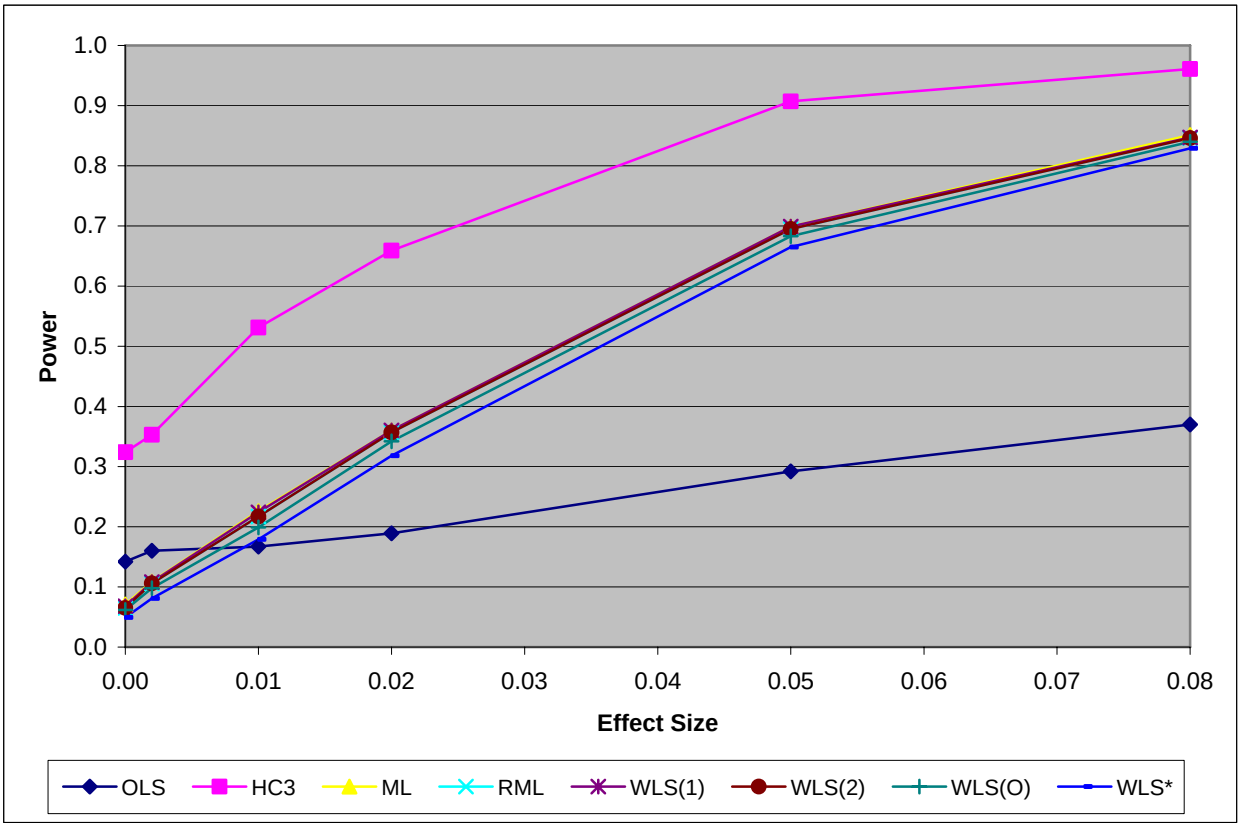
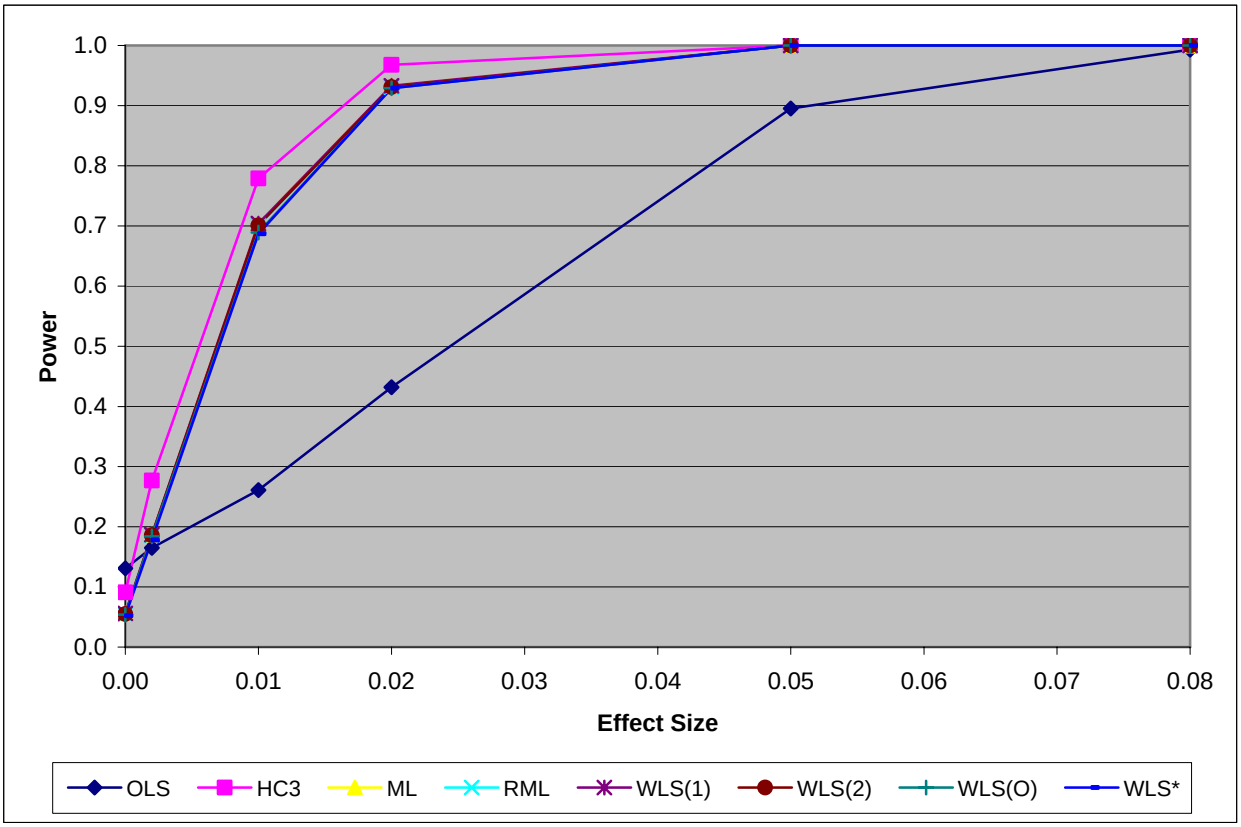
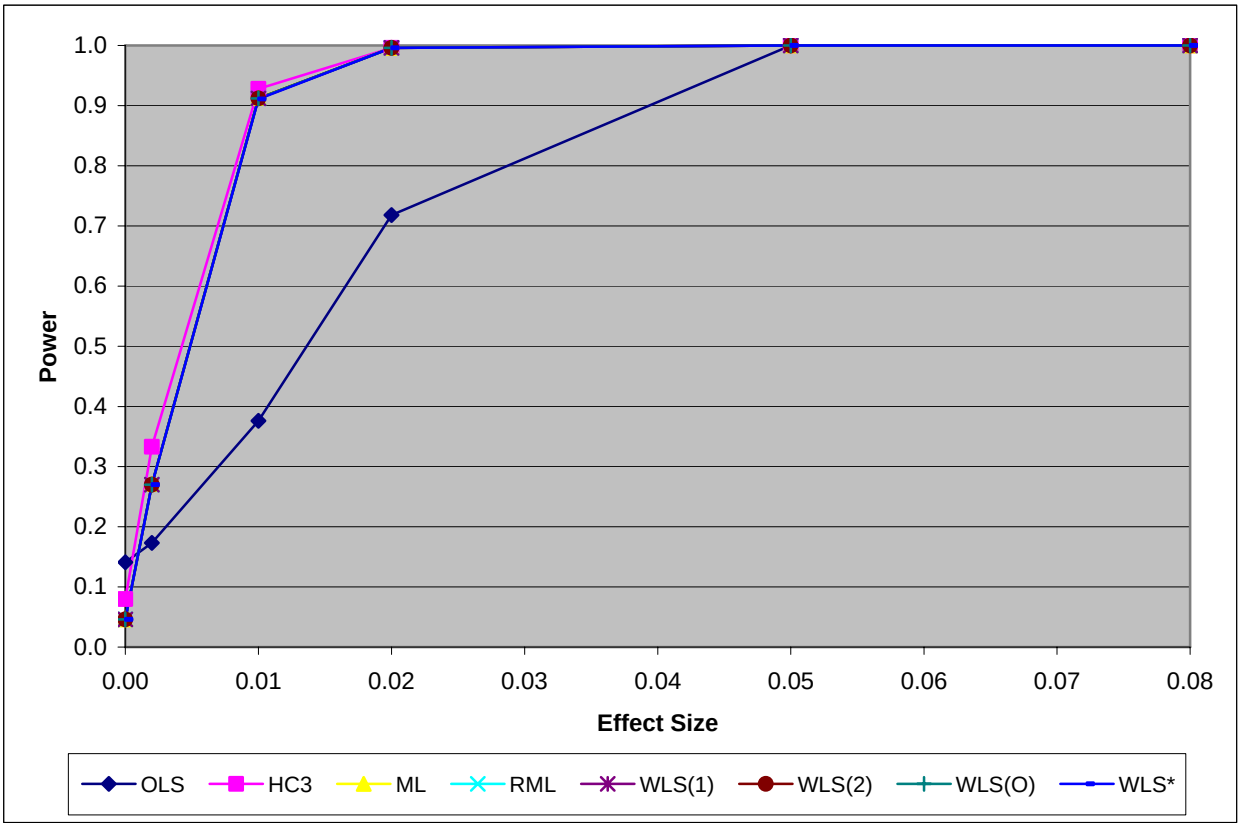


Figure 18. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, indirect pairing, $\sigma_{e_j}{}^2$ s = 1, 1, 16, moderately unequal proportions, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

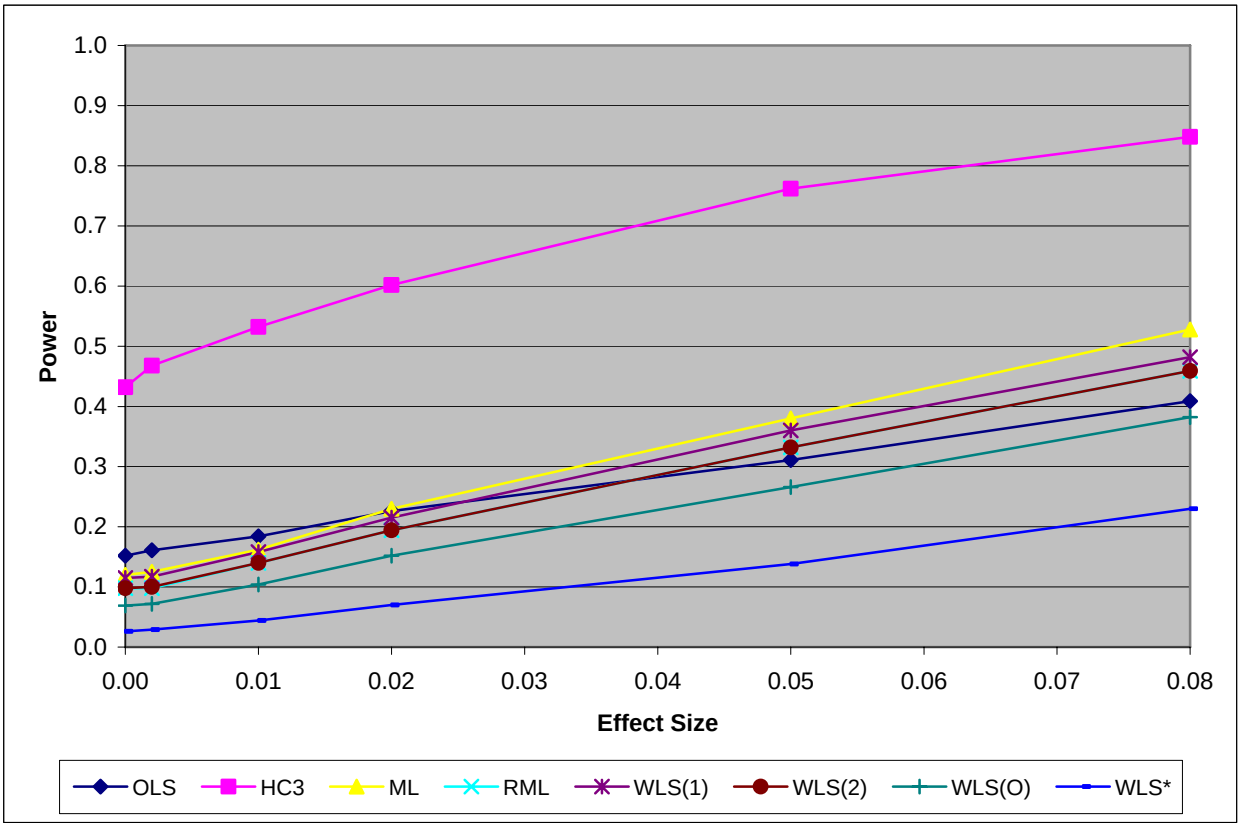
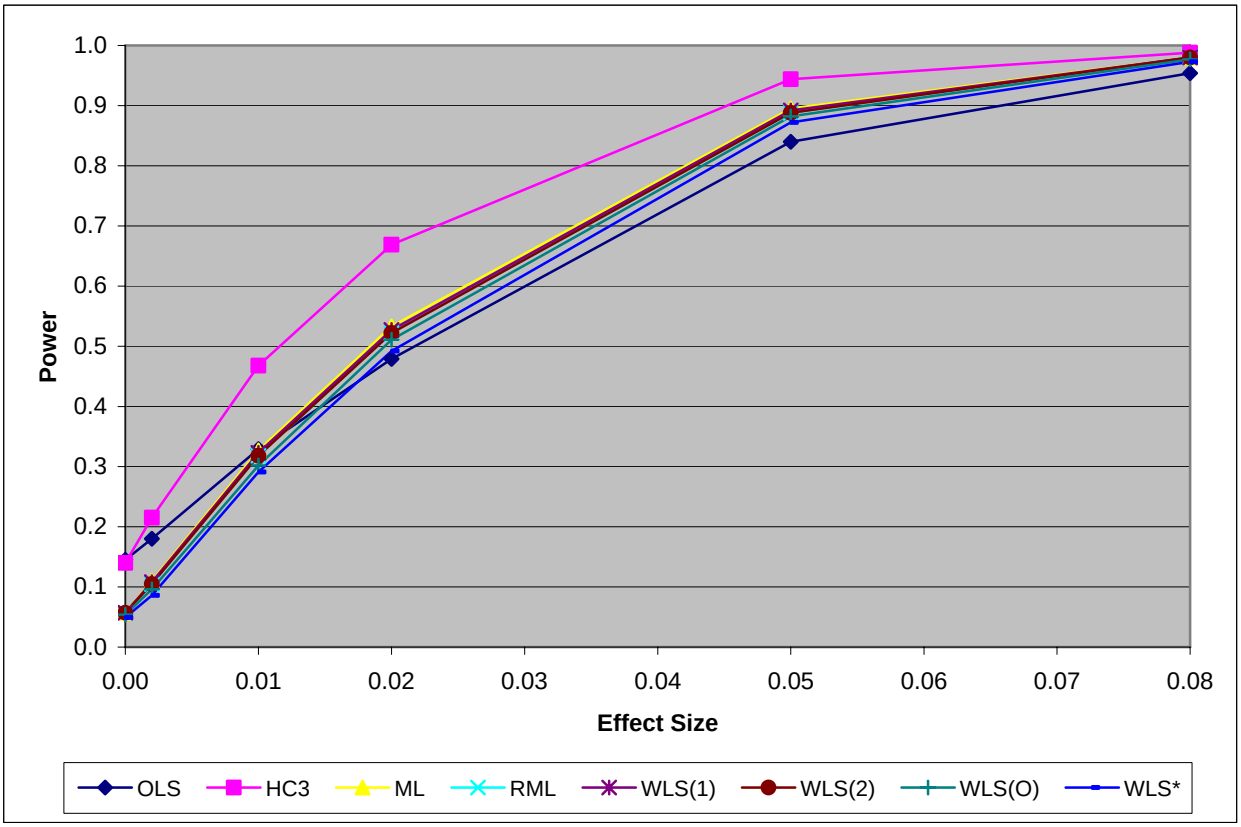
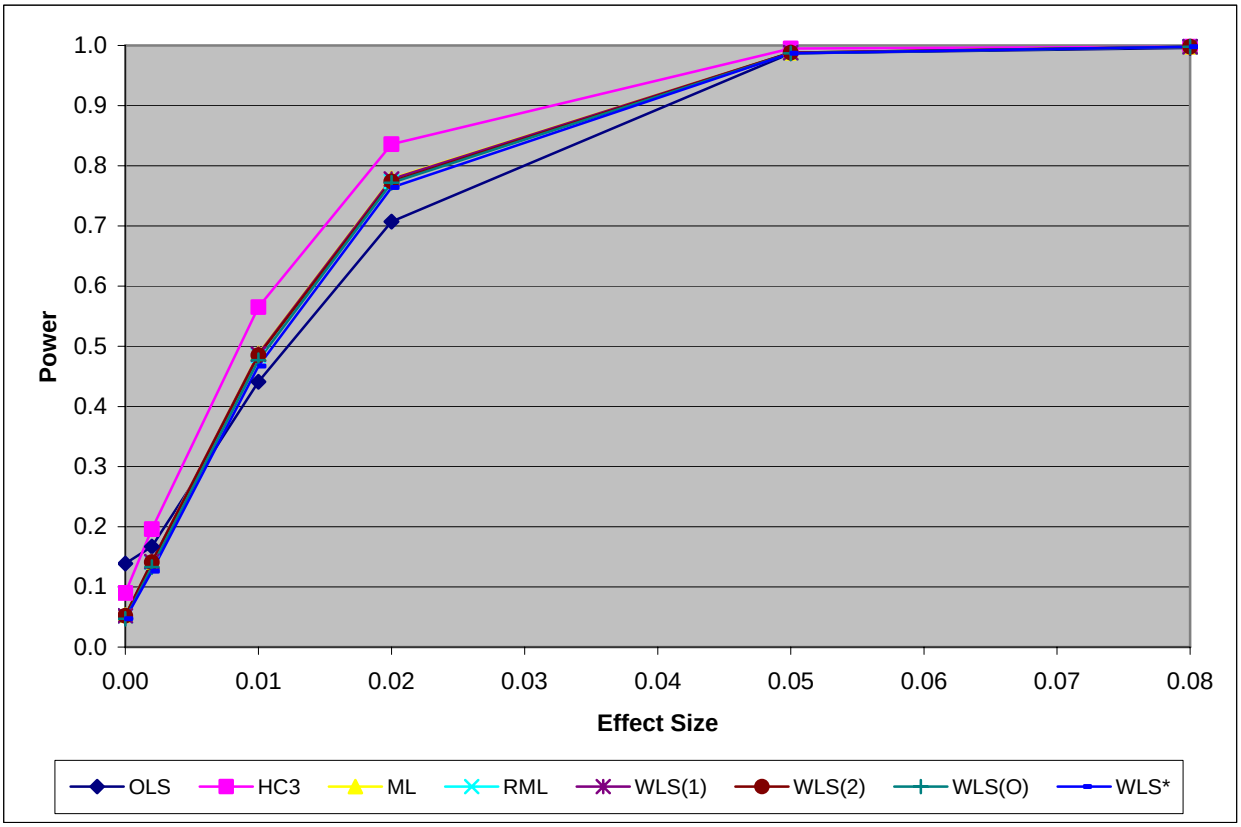


Figure 19. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, indirect pairing, $\sigma_{e_j}^2$ s = 1, 1, 4, very unequal proportions, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel A

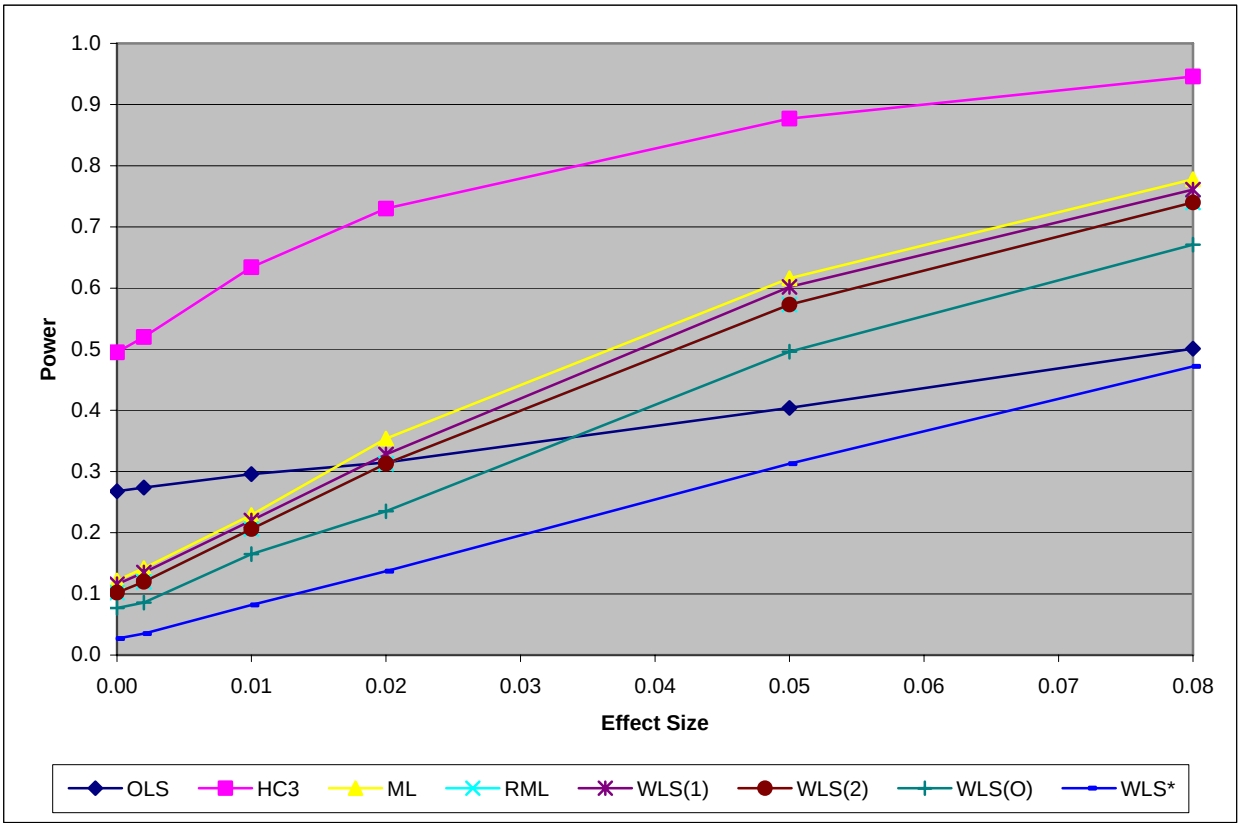
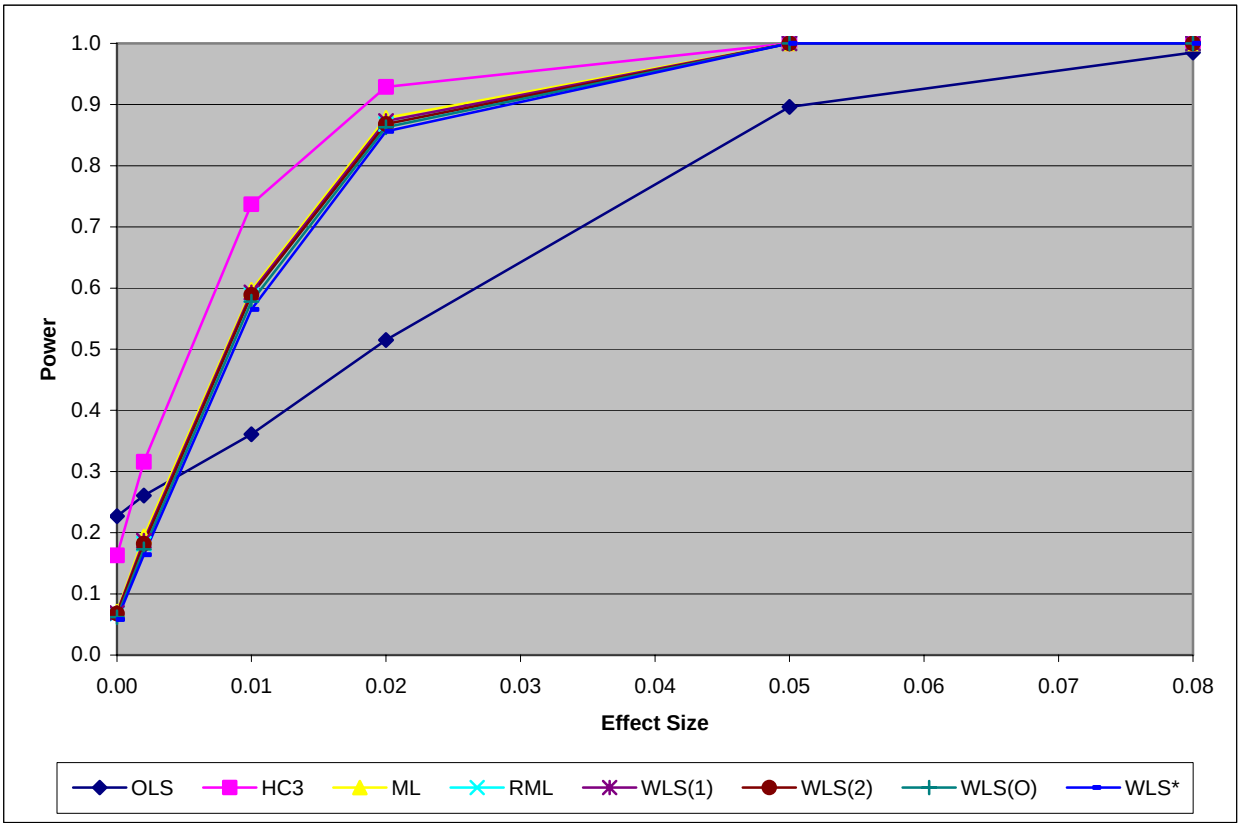
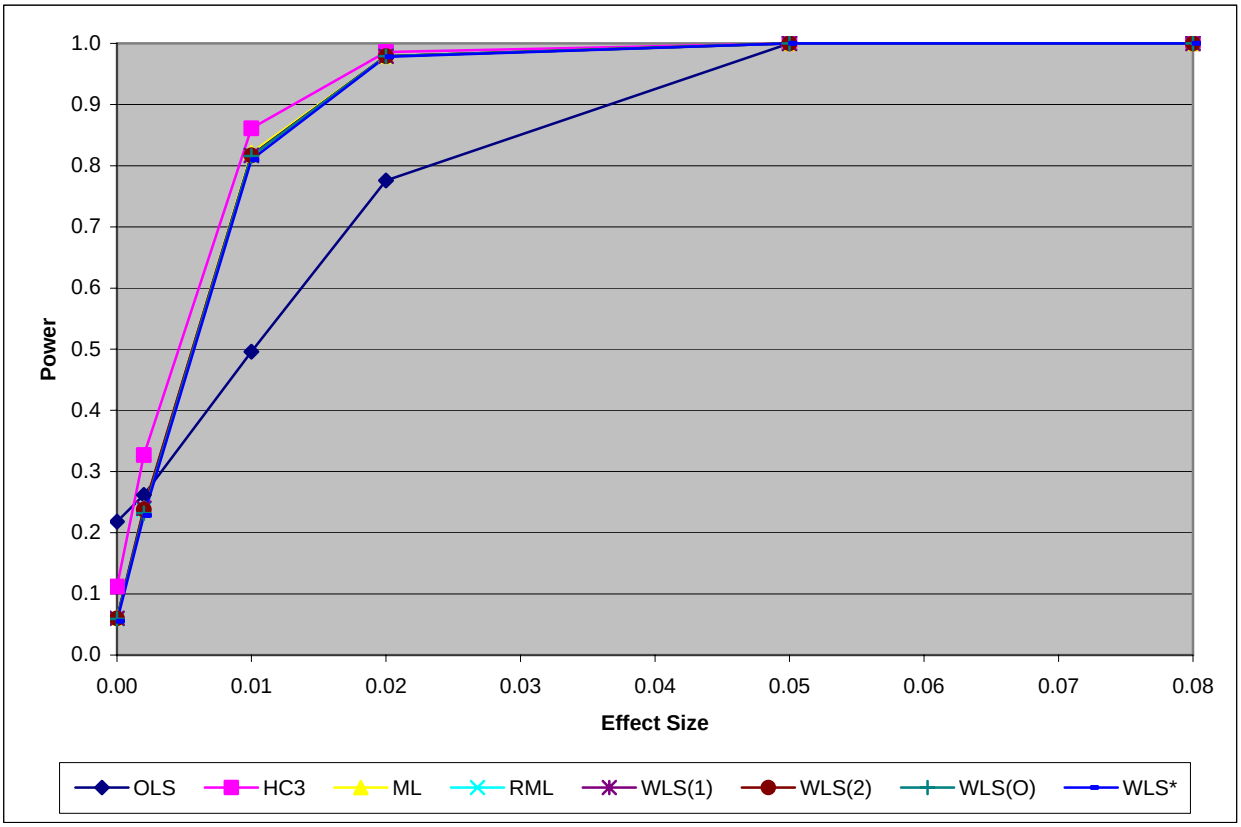


Figure 20. Statistical power as a function of effect size for F_{OLS} (OLS), F_{HC3} (HC3), F_{ML} (ML), F_{RML} (RML), $F_{WLS(1)}$ (WLS(1)), $F_{WLS(2)}$ (WLS(2)), $F_{WLS(O)}$ (WLS(O)), and F_{WLS*} (WLS*) with three groups, indirect pairing, $\sigma_{e_j}^2$ s = 1, 1, 16, very unequal proportions, and (A) $N = 48$, (B) $N = 192$, and (C) $N = 336$.

Panel B


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

Panel C


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Power
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Effect Size
OLS
HC3
ML
RML
WLS(1)
WLS(2)
WLS(O)
WLS*

CHAPTER FOUR: DISCUSSION

Contributions of the Present Study

The present study makes a number of unique contributions to extant research on testing for the equality of regression slopes.

First, although statistical approximations exist which have been shown to perform well when the homoscedasticity assumption is violated (DeShon & Alexander, 1996), this study considered seven alternatives to F_{OLS} that have never been previously compared. That is, in contrast to the statistical approximations, this study compared those methods that (a) are simpler to compute, (b) are available in standard statistical software, and (c) permit post hoc analyses of a statistically significant interaction.

Second, to the author's knowledge, the performance of an HCCM when testing for the equality of regression slopes has never been investigated. Although using an HC3 has been described by Long and Ervin (2000) as an effective alternative to $\text{cov}(\hat{\boldsymbol{\beta}})$ even when the form of heteroscedasticity is unknown, they considered the following model: $y_i = 1 + 1x_{1i} + 1x_{2i} + 1x_{3i} + 0x_{4i} + \tau\varepsilon_i$ (with various forms of heteroscedasticity for the error term) which did not vary heteroscedasticity as a function of a categorical predictor. The present study did.

Third, this study described a proposed extension (i.e., w_j^*) of Overton's (2001) WLS method for testing the equality of regression slopes and compared its performance to alternative approaches for estimating weights. More precisely, Overton (2001) stated that the WLS method should not be applied to $k > 2$ because Type I error rates were inflated. However, to the contrary, this study showed that WLS can be effectively applied to ameliorate the biasing effects of heteroscedasticity even for $k = 4$, and using w_j^* controlled Type I error rates at the nominal level better than the alternative approaches for estimating weights.

Fourth, the results of the present study have very important implications for researchers in psychology and other behavioral sciences. Because, in such areas, low power is argued as being a key reason for the failure to detect hypothesized effects (Cohen, 1988) including interactions (Aguinis & Stone-Romero, 1997; McClelland & Judd, 1993; Stone-Romero & Liakhovitski, 2002; Zedeck, 1971) and OLS-based parameter estimates are known to be inefficient when heteroscedasticity exists, thus, affecting statistical inferences (i.e., hypothesis tests, confidence intervals, joint confidence bands; Cook & Weisberg, 1999; Draper & Smith, 1966; Greene, 2003; Neter et al., 1996; Rencher, 2000), it would be beneficial for researchers to utilize those methods that can help address these issues. That is, when heteroscedasticity is suspected, a researcher should utilize more appropriate methods, specifically, those which provide the greatest statistical power without sacrificing control of Type I error rates. A number of more appropriate methods were discussed above (e.g., *A* and *J* approximations). In addition to these, the present study identified other methods that outperform F_{OLS} in (a) controlling Type I error rates, and (b) providing increased statistical power. In short, when heteroscedasticity exists, the use of WLS (specifically, w_j^*) when testing for the equality of regression slopes is recommended, particularly because this study showed that it can increase the likelihood of detecting hypothesized interactions and still control Type I error rates at the nominal level.

As noted above, OLS is a special case of WLS regression. Both are special cases of generalized least squares (Fox, 1997; Greene, 2003; Neter et al. 1996; Rencher, 2000). Although WLS regression is often recommended as a remedy for heteroscedasticity in other fields (e.g., agriculture, econometrics, engineering, statistics), it has generally received less attention and application in psychology. However, there are exceptions (Overton, 2001; Steel & Kammeyer-

Mueller, 2002). In the following paragraphs, I consider some positions against WLS regression and further delineate reasons that warrant its use.

Consider the following statement: "In WLS the measures of standardized effect size, such as R^2 . . . do not have a straightforward meaning as they do in OLS" (Cohen et al., 2003, p. 147). Because it is argued that effect size measures typically employed in psychology and other behavioral sciences (Cohen, 1988) cannot be easily interpreted in WLS, some view it as generally less desirable than OLS regression. However, with respect to effect size, Willett and Singer (1988) presented a pseudo-R^2 for WLS regression. In an example of its application, the pseudo-R^2 was nearly equal to the OLS-based R^2. These researchers argue that the pseudo-R^2 is not likely to differ much from the OLS-based R^2 (Willett & Singer, 1988). They suggest that researchers should "refocus attention on other aspects of the analysis, particularly the increased precision of the estimates of β" (Willett & Singer, 1988, p. 238). That is, using OLS in the presence of heteroscedasticity results in inaccurate and often inflated standard errors for the estimated regression coefficients, precluding the detection of hypothesized effects. However, as discussed above, WLS regression results in accurate, smaller standard errors. More recently, Aguinis and colleagues (2005) have shown that the standard effect size for interactions among categorical and continuous predictors (Aiken & West, 1991; Cohen, 1988) is inappropriate when heteroscedasticity exists and they have derived a more appropriate measure which was used in the present study. Taken together, the issue of effect size as a reason not to use WLS regression seems tenuous, particularly in light of the next matter.

Consider the following statement: "Because of the imprecision in estimating weights, OLS regression will often perform nearly as well as (or sometimes even better than) WLS regression when the sample size is small . . . suggest[ing] that OLS regression will be preferable

to WLS regression except in cases where the sample size is large or there is a very serious problem of nonconstant variance" (Cohen et al., 2003, p.147). Note that there were no references to bolster this claim. Based on the findings of the present simulation, the results of Overton's (2001) study, and literature in econometrics (Greene, 2003) and statistics (Carroll & Ruppert, 1988; Mak, 1992; Rencher, 2000), this statement cannot be supported when testing for the equality of regression slopes. Namely, WLS was virtually always more powerful than OLS even with $N = 48$, 96, and 144 and when the number of estimated regression coefficients was 6 and 8. In addition, because the instances where OLS outperformed WLS regression in this study were with *indirect pairing*, OLS had inflated Type I error rates. Therefore, the power advantage was illusory, i.e., at the expense of too many Type I error rates. However, recall that this advantage diminished as f^2 increased, to the point that OLS had the lowest power of every method in this investigation. Notably, the superior performance of the WLS methods was evident even with mild levels of heteroscedasticity. Moreover, with slight deviations from non-zero f^2s, the WLS methods had more rapid increases in power than OLS regression. Therefore, it appears that WLS regression would be a very desirable alternative when testing for the equality of regression slopes when heteroscedasticity exists, particularly because the WLS methods were quite sensitive to detecting non-zero f^2s with considerable power gains in the range of f^2s typical in applied psychology (e.g., .002 and .009, Aguinis et al., 2005). In short, WLS regression (specifically, w_j^*) can be used in studies where heteroscedasticity exists, potentially detecting hypothesized effects that would have otherwise gone undetected.

In the following sections, the overall effects of the manipulated variables on the performance of the various methods are discussed.

Effects of Manipulated Variables on F_{OLS}

The effects of the manipulated variables on F_{OLS} were consistent with previous research findings in virtually every respect. When ${}_kP_j$s were unequal, with *direct pairing*, Type I error rates were conservative and became *increasingly conservative* when (a) ${}_kP_j$s were very unequal, or (b) heteroscedasticity increased. When extreme heteroscedasticity was combined with very unequal ${}_kP_j$s, Type I error rates were dramatically conservative. In contrast, with *indirect pairing*, Type I error rates were inflated and became *extremely inflated* when (a) ${}_kP_j$s were very unequal, or (b) heteroscedasticity increased. When extreme heteroscedasticity was combined with very unequal ${}_kP_j$s, Type I error rates were inflated markedly. This has implications for power. More precisely, depending on the type of pairing, the minimum of the power function (i.e., β_0) will either (a) shift downwards away from α (direct pairing), overall, causing power to decrease, or (b) shift upwards away from α (indirect pairing), overall, causing power to increase illegitimately, particularly at smaller f^2s. According to Aguinis and Pierce (1998), the former is most likely in organizational settings where the majority group (e.g., White) have the larger n_j, but the smaller validity coefficient (and therefore the larger $\sigma_{e_j}{}^2$) than that of the minority group (e.g., African-American) (Hattrup & Schmitt, 1990). Note that, in the latter, because the minimum of the power function is shifted upwards, F_{OLS} has an "apparent power advantage" (DeShon & Alexander, 1996, p. 265) acquired at the expense of inflated Type I error rates (Alexander & DeShon, 1994; Overton, 2001). With indirect pairing, a very interesting and unique finding is that, ceteris paribus, as f^2 increases, the power function of F_{OLS} crosses that of all the other methods (even F_{HC3} in some instances) and eventually has lower power than *every* method compared in this study.

When ${}_kP_j$s were equal, the results of the study also produced unique findings. Namely, when heteroscedasticity exists, Dretzke et al. (1982) suggested that F_{OLS} was robust when ${}_kP_j$s were equal. In their study, note that $k = 2$ and σ_{x_j} was *equal* across groups. However, DeShon and Alexander (1996) later showed that F_{OLS} was not robust when ${}_kP_j$s were equal; Type I error rates became conservative. In their study, note that $k = 2$ and σ_{x_j} was *unequal* across groups. That is, across the two groups, although σ_{x_j} varied, the ratios of σ_{x_j} to σ_{y_j} were constant (viz., .7071). Therefore, in their study, the test for the equality of correlation coefficients and regression slopes was equivalent. In the present investigation, when ${}_kP_j$s were equal, I found further evidence demonstrating that F_{OLS} was not robust. Specifically, Type I error rates became inflated! The reason for this was that k was manipulated as well as heteroscedasticity. That is, although σ_{x_j} was equal across groups (like Dretzke et al., 1982), for a fixed degree of heteroscedasticity, as k increased, Type I error rates increased. Further, for a fixed k, as heteroscedasticity increased, Type I error rates increased. Note that this does not negate the very valuable findings that Dretzke et al. (1982) reported. Because they did not manipulate k nor did they manipulate more than two levels of heteroscedasticity (i.e., $\sigma_{e_j}{}^2 = 3.75,\ 0.19\overline{4}$; $\sigma_{e_j}{}^2 = 0.\overline{4}$, $0.19\overline{4}$), the present study found that F_{OLS} was not robust even when ${}_kP_j$s were equal. Therefore, consistent with DeShon and Alexander (1996), F_{OLS} is not immune from the biasing effects of heteroscedasticity when ${}_kP_j$s are equal. Type I error rates can be (a) conservative when, across groups, σ_{x_j} is unequal, but the ratios of σ_{x_j} to σ_{y_j} are equal (DeShon & Alexander, 1996), or (b) inflated when, across groups, σ_{x_j} is equal, but the ratios of σ_{x_j} to σ_{y_j} are unequal (the present study).

The very serious implication of the foregoing is that when the homoscedasticity assumption is violated, the actual Type I error rates of F_{OLS} will be biased, affecting power. In any given study, if researchers do not assess whether the homoscedasticity assumption was violated, then how can sample-based conclusions be trusted? That is, with heteroscedasticity, a substantive conclusion can change as Aguinis et al. (1999) illustrated.

Effects of Manipulated Variables on F_{HC3}

When testing for the equality of regression slopes and heteroscedasticity exists, F_{HC3} was clearly the most powerful test. Even when *N*s and f^2s were small, its power was many times greater than that of all the other methods. Concurrently, it had the most inflated Type I error rates. Therefore, with power functions which have a minimum at a value far greater than α, it had an illusory power advantage over all tests in nearly every condition. However, there were some exceptions. For example, there were some conditions where the Type I error rate of F_{OLS} was more inflated than that of all the other methods including F_{HC3}. But, as f^2 increased, the illusory power advantage of F_{OLS} diminished until its power function was below that of all the other methods including F_{HC3}. In general, for F_{HC3}, its power advantage became more pronounced as heteroscedasticity increased. In addition, it was more powerful with direct pairing than indirect pairing especially with very unequal ${}_kP_j$s. In contrast to all the other methods, increasing *k* tended to result in greater power.

Of the manipulated variables, *N* greatly affected the Type I error rates of F_{HC3}. Specifically, the inflated Type I error rates decreased as *N* increased. Although as f^2 increased, power *increased*, ceteris paribus, the minimum of its power function shifted *downwards* as *N* increased. Therefore, ceteris paribus, computing the average power of this test as a function of *N* is nonsensical because the leftmost segment of the power function (i.e., at smaller f^2s, including

zero) is shifting downwards as *N* increases. It deserves stressing that HCCMs like HC3 provide asymptotically correct statistical inferences (White, 1980). As Greene (2003) noted with respect to HC0, the estimator upon which HC3 is based, "The asymptotic properties of the estimator are unambiguous, but its usefulness in small samples is open to question" (p. 220). Because Long and Ervin (2000) suggested it performs well even with small *N*s and when the form of heteroscedasticity is unknown, the HC3 was included in the present study. It is likely that at very large *N*s, outside the range considered in the present study, F_{HC3} will have more desirable properties. However, based on this investigation and considering the *N*s typically used in psychology and other behavioral sciences, F_{HC3} cannot be recommended for testing the equality of regression slopes. If it were used, a researcher would have a very high likelihood of rejecting the null hypothesis of equal slopes even if the null were true in the population. With a nominal $\alpha = .05$, the actual Type I error rates were as high as .587. If aspects of a psychological theory were tested using F_{HC3}, any inferences drawn from such a test could misdirect future research. Furthermore, if subsequent tests were also conducted using F_{HC3}, then this would provide additional specious support for the theory. Clearly, this would have serious implications for theoreticians and the practitioners who attempt to apply a theory (say, in an organizational setting) that is supported by compounded Type I error rates. This does not diminish the overall value of using HC3 because its asymptotic properties are inarguable, but its application in the social and behavioral sciences would be limited to very large samples such as those obtained for large-scale survey research.

Effects of Manipulated Variables on F_{ML} and F_{RML}

The effects of the manipulated variables on F_{ML} and F_{RML} were very similar. When using ML and RML for estimation and heteroscedasticity is a function of the categorical predictor (i.e.,

z), the tests generally performed much better compared to F_{OLS} and F_{HC3}. Although the Type I error rates for F_{ML} and F_{RML} were inflated (especially at small *N*s), F_{RML} tended to control them slightly better than F_{ML}. Although as f^2 (or heteroscedasticity) increased power increased, the gains in power were larger when there was extreme heteroscedasticity with direct pairing. In addition, F_{ML} was slightly more powerful than F_{RML}. In the present context, the power functions for F_{ML} and F_{RML} can be viewed as being near parallel to one another where the power function for F_{ML} is higher than that of F_{RML}.

Because RML adjusts the estimated variances for the number of fixed effects, as *k* increases, the power functions of F_{ML} and F_{RML} will likely diverge. Specifically, as *k* increases, the power of F_{RML} will increase less rapidly than F_{ML}. Overall, compared to F_{ML}, F_{RML} is more preferable because its Type I error rates were less inflated and provided similar levels of power. Note that the performance of F_{RML} was virtually identical to that of $F_{WLS(2)}$. This may not be true in all situations. Recall that estimation of parameters based on likelihood functions requires numerical methods and depending on the algorithm used by a statistical software package, these may not perform exactly the same. However, because their performance was so similar either could be used. It deserves stressing however that F_{RML} was less able to control Type I error rates compared to other methods to be discussed next.

Effects of Manipulated Variables on $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*}

The effects of the manipulated variables on $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*} were very similar. In terms of ability to control Type I error rates, the methods can be ranked from *least able* to *most able* as follows: $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*}. That is, across the various conditions, the Type I error rates for F_{WLS*} were closest to the nominal level. The others, particularly $F_{WLS(1)}$, tended to have more inflated Type I error rates. This was most noticeable at

smaller *N*s or as *k* increased. Not surprisingly, in terms of statistical power, the methods can be ranked from *least powerful* to *most powerful* as follows: F_{WLS*}, $F_{WLS(O)}$, $F_{WLS(2)}$, and $F_{WLS(1)}$. In other words, their power functions are generally parallel to one another and in the rank order noted above. Note that inspection of the power tables and figures will show that their relative differences were small. Furthermore, as *N*, heteroscedasticity, or f^2 increased, the power functions for all four methods converged. However, when the ${}_kP_j$s were very unequal and *N*s were small, the power function for F_{WLS*} was generally lower than that of the other three WLS methods. It deserves mentioning that F_{WLS*} provided greater control over Type I error rates and thus its power function was not shifted upwards away from α due to the biasing effects of heteroscedasticity like that of the other methods.

In terms of the type of pairing, this had an affect on the WLS methods. Ceteris paribus, power was greater with direct pairing than indirect pairing. However, again, across conditions, regardless of the type of pairing, F_{WLS*} and $F_{WLS(O)}$ controlled Type I error rates at the nominal level better than the other WLS methods as well as F_{OLS}, F_{HC3}, F_{ML}, and F_{RML}.

In general, *k* had similar effects on all the WLS methods. As *k* increased, power decreased. This is because increasing *k* results in more parameters to estimate and, therefore, fewer *df*. Note also that, for the WLS methods, increasing *k* results in fewer *df* for estimating weights. For example, for a fixed $N = 100$ and equal ${}_kP_j$s, when $k = 2$, $n_1 = n_2 = 50$. If *k* increased to 4, then $n_1 = n_2 = n_3 = n_4 = 25$. Therefore, for a fixed *N*, researchers should be certain that *k* does not increase to the point where insufficient *df* are available for estimating weights. In other words, consistent with recommendations for a well-designed and executed experiment, a research study with more participants in a condition relative to the number of parameters to estimate will provide better estimates of weights. Furthermore, this is especially important for

F_{WLS*} because it appears to be more affected by k than the other methods because w_j^* changes depending on the model. In contrast, ${}_1w_j$, ${}_2w_j$, and ${}_Ow_j$ remain the same regardless of the complexity of the model.

Overall, F_{RML}, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*} demonstrated better performance than F_{OLS}, F_{HC3}, F_{ML}, and $F_{WLS(1)}$,. However, because the performance of statistical tests should be evaluated not only in terms of statistical power but also its ability to provide accurate Type I error rates, some practical recommendations are next discussed, including data-analytic ones.

Some Practical Recommendations and Considerations

Consistent with previous research (Aguinis, 2004; Aguinis et al., 1999; Alexander & DeShon, 1994; Box, 1954; DeShon & Alexander, 1996; Luh & Guo, 2002; Overton, 2001; Wilcox, 1997) and the findings from the present study, when heteroscedasticity exists, F_{OLS} should not be used to test for the equality of regression slopes. Therefore, from a data analysis perspective, after fitting any regression model, residual diagnostics should be performed to assess whether any assumptions were violated, including homoscedasticity (Cook & Weisberg, 1999; Fox, 1997; Neter et al., 1996). If heteroscedasticity is detected, researchers should use an alternative method. For example, statistical approximations are available (DeShon & Alexander, 1996) and Aguinis (2004) provides a computer program that performs these.

For alternatives which can be employed in most common statistical software, one of the following can be used for $k \leq 4$: F_{RML}, $F_{WLS(2)}$, $F_{WLS(O)}$, or F_{WLS*}. However, because $F_{WLS(2)}$ is effectively identical to F_{RML}, the recommendations simplify to F_{RML}, $F_{WLS(O)}$, or F_{WLS*}. Note, however, because F_{WLS*} provides comparable power levels without sacrificing control of Type I error rates (e.g., at small Ns), it may be slightly preferable.

For WLS methods in general, it is important that there are enough df_{E_j} to estimate weights. Although F_{WLS*} still performed well even with $df_{E_j} = 2$ in the smallest group (i.e., when $N = 48$ and n_js = 32, 8, 8), this method could not be used in other instances (i.e., $N = 48$ and n_js = 24, 8, 8, 8). Therefore, n_js should be > q, preferably with at least $df_{E_j} = 10$ in all groups to be consistent with the recommendation by Bement and Williams (1969).

Noteworthy, when ${}_kP_j$s are equal, it makes little difference whether a researcher uses F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, or F_{WLS*} because their performance was identical with Type I error rates near the nominal level. Note that the performance of F_{ML} and F_{RML} may not be the same across different software packages depending on the numerical algorithm used to find the parameter estimates that maximize the likelihood function.[1] When sample sizes are large, differences are normally minimal.

It deserves stressing that residual diagnostics be conducted *after* fitting a model. This is consistent with ruling out one of the threats to statistical conclusion validity described by Shadish et al. (2002)—violated assumptions of statistical tests. That is, if a regression equation is computed and hypothesized effects are detected, a researcher should evaluate whether the effect was an artifact of heteroscedasticity. If diagnostics suggest that the homoscedasticity assumption was tenable, then the researcher can be more confident that the result was not due to inflated Type I error rates attributed to heteroscedasticity. Similarly, if a regression equation is computed and hypothesized effects are *not* detected, a researcher should assess whether the *failure* to detect such an effect was an artifact of heteroscedasticity. If diagnostics suggest that homoscedasticity was violated, then use an alternative method such as those mentioned above. If diagnostics suggest that the homoscedasticity assumption was tenable, then the researcher can be more confident that the result was not due to low power attributed to heteroscedasticity. In either case,

after fitting a model, an analysis of the residuals can help (a) detect potential problems in the model, and (b) avoid erroneous inferences stemming from violating the assumptions of statistical tests.

In the sections that follow, I discuss a few standard graphical methods in addition to a variety of tests that can be used to detect heteroscedasticity. Then, the issue of functional form is briefly discussed.

Graphical Methods for Detecting Heteroscedasticity

A number of graphical methods exist for detecting heteroscedasticity (Cook, 1994; Cook & Weisberg, 1999; Fox, 1997; Neter et al., 1996). All of the approaches use the residuals (i.e., e_is) or variants of them. As noted above, the N errors (i.e., ε_is) are unknown and each is assumed to follow a distribution with a zero expectation and (in OLS) common variance of σ^2 (i.e., homoscedasticity). Although we do not know the actual ε_is, they can be estimated by the e_is. In the two general types of plots to be described below, it should be noted that either the e_is or variants of them (e.g., studentized e_is) can be employed. Numerous texts describe the variants in detail (Fox, 1997; Neter et al., 1996). Furthermore, the plots can be generated as part of the usual output in common statistical software (e.g. SAS, S-PLUS, SPSS).

Fox (1997) describes the two typical plots succinctly. "Because the regression surface is . . . [p-dimensional] . . . and embedded in a space of . . . [q] . . . dimensions, it is generally impractical to assess the assumption of constant error variance by direct graphical examination of the data when . . . [p] . . . is larger than 1 or 2. Nevertheless, it is common for error variance to increase as the expectation of Y grows larger, or there may be a systematic relationship between error variance and a particular X [like that of the present study]. The former can often be

detected by plotting residuals against fitted values, and the latter by plotting residuals against each X" (Fox, 1997, p. 301).

Using a scatterplot for the first, it is common to plot the e_is on the vertical axis and the fitted values ($\hat{y}_i$s) on the horizontal axis. Such a plot can provide evidence of whether the ε_is depend on the $\hat{y}_i$s . If the scatterplot of points have no systematic pattern (i.e., not wedge-shaped, not curved, etc.), they will have an approximate rectangular form. This suggests that the ε_is do not depend on the $\hat{y}_i$s . That is, homoscedasticity was not violated.

For the second, using scatterplots (for continuous predictors) or boxplots (for categorical predictors), the e_is can be plotted on the vertical axis and a predictor on the horizontal axis. For example, if the e_is were plotted against a continuous predictor that was used in the regression analysis (e.g., x), this could provide evidence of whether the ε_is increase or decrease with the values of the predictor, a violation of the homoscedasticity assumption. Similarly, to address the form of heteroscedasticity in this study, the e_is could be plotted against the categorical predictor used in the regression analysis (e.g., z). If the boxplots show that the e_is have markedly different spread across the levels of the predictor, than there is evidence of heteroscedasticity. Furthermore, linear combinations of predictors can be used on the horizontal axis. No matter which type is used, these plots should not suggest markedly different variability across the values (or levels) of a predictor. Otherwise, this is an indicant of heteroscedasticity.

Recognize that plots of e_is are not without their problems (Cook, 1994). As will be discussed below, heteroscedasticity could signal that an incorrect functional form was specified or that an important predictor was omitted (Fox, 1997; Neter et al., 1996). However, the use of graphical methods is an important tool for understanding the data being analyzed. Stated another way, how can a researcher know whether an assumption was violated? The answer seems

apparent. Researchers should inspect the residual plots and conduct analyses such as those described below. Notably, this is consistent with Cohen's (1994) recommendation that "even before we, as psychologists, seek to generalize from our data, we must seek to understand and improve them. A major breakthrough to the approach to data, emphasizing 'detective work' rather than 'sanctification' was heralded by John Tukey . . . [by applying] . . . simple, flexible, informal, and largely graphic techniques . . . for understanding the set of data in hand" (p. 1001). Considering the major advances in technology including processor speed and computer graphics, such analyses often require a minimal amount of time and effort in standard statistical software.

Tests for Heteroscedasticity

To further assist researchers in detecting heteroscedasticity, I briefly describe a general test, a number of specific tests, and a heuristic method.

Breusch and Pagan (1979) and, independently, Cook and Weisberg (1983) developed a score test and it assumes that the σ_i^2 s are independently and normally distributed. It can be used to detect various forms of heteroscedasticity. For example, it can be used to test whether the σ_i^2 s are related to (a) one predictor (e.g., a categorical variable like ethnicity), (b) a combination of predictors (e.g., continuous and/or categorical), or (c) the $\hat{y}_i$s. The test statistic is simple to compute and is based on two OLS regression analyses. In the first, the *SSE* from the regression equation of interest (e.g., Equation 1) is required. In a second regression, the squared e_is from the first analysis (i.e., Equation 1) are regressed on the predictors believed to be the cause of the heteroscedasticity (i.e., in the present study, *z*). The regression sum of squares (*SSR*) is required from this analysis. The test statistic is $(SSR / 2) \div (SSE / N)^2$ and is asymptotically distributed as χ^2 with *df* equal to the number of variables used to predict the squared e_is. If statistically significant at some predetermined α, this suggests that heteroscedasticity is a function of the

predictors used to predict the squared e_is. An example with one predictor is given in Neter et al., (1996, p. 115). Note that this test and two others are discussed by Greene (2003, pp. 222-225).

A specific test which can be used to detect the form of heteroscedasticity discussed in the present study is that described by Bartlett (1937). This test assumes the underlying distributions are normal. Although the test statistic is complex to compute, Aguinis et al. (1999) provide a computer program that performs the necessary calculations. This test is sensitive to non-normality (Box, 1953; DeShon & Alexander, 1996; Kirk, 1995; Levene, 1960) as is the Breusch and Pagan (1979) test (Greene, 2003). Stated differently, a statistically significant test statistic could signal non-normality, heteroscedasticity, or both. Perhaps due to this lack of diagnosticity, various researchers have proposed robust methods.

A robust method was described by Levene (1960). Although typically used in ANOVA, it can be easily adapted to detect the form of heteroscedasticity described here. Namely, it is a one-way ANOVA performed on the absolute value of the e_is where the categorical predictor serves as the "factor." If the F is statistically significant, at say, .10, then there may be evidence of heteroscedasticity. Brown and Forsythe (1974) extended Levene's (1960) method by using the absolute value of the e_is about their respective group medians (as well as using trimmed means). Conover, Johnson, and Johnson (1981) recommended the method by Brown and Forsythe (1974), along with two other more complex approaches, because of their robustness and power. More recently, Sarkar, Kim, and Basu (1999) recommended a robust method that modifies Levene's (1960) test using weighted likelihood estimates, arguing that it outperforms the approach by Brown and Forsythe (1974). Note that the tests by Levene (1960) and Brown and Forsythe (1974) are available in such common statistics software as S-PLUS and SPSS.

Finally, DeShon and Alexander (1996) provide a useful heuristic for when to invoke a data-analytic alternative. Namely, when the variance of the e_is in one group is approximately 1.5 times greater than that of another group, the conventional F_{OLS} should not be used. Because this heuristic was derived based on a variety of simulations when $k = 2$, it is unclear whether this generalizes to $k > 2$. For example, if $k = 4$ and the ratio of residual variances across groups is 1:1:1:1.6, does the heuristic apply with equal force? Nevertheless, the heuristic provides a practical standard.

Functional Form

Failure to specify the correct functional form can introduce heteroscedasticity into the model (Fox, 1997; Long & Ervin, 2000; Neter et al., 1996). For example, a researcher may specify the following functional form of how two predictors are related to y: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i$. This model results in a 2-dimensional regression surface (e.g., a sheet of paper) in 3-dimensional space which could be tilted (i.e., main effects), curved (i.e., interaction) or both (i.e., main effects and interaction). If the functional form was incorrectly specified, the e_is would have a nonconstant spread about the regression surface (i.e., heteroscedasticity) (Neter et al., 1996).

To continue with this example, after substituting sample-based estimates, the actual regression equation might take the following form: $\hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_{1i}^2 - \hat{\beta}_2 x_{2i}^2 - \hat{\beta}_3 x_{1i} x_{2i}$. Namely, x_1 and x_2 are related to y such that the actual regression surface is concave (e.g., like an umbrella) (Neter et al., 1996, p. 299). In this case, the e_is might be evenly spread about the regression surface. Consider the following illustration. A researcher might hypothesize that team stress (x_1) and the number of team members (x_2, with values from 3 to 14) are related to team performance (y) in an interactive curvilinear manner similar to the just-noted regression equation. That is, the

marginal relation between stress/arousal and team performance is an inverted U-shape (see Yerkes & Dodson, 1908). The marginal relation between the number of team members and team performance is also an inverted U-shape. Namely, as the number of team members increases, team performance, generally, increases. However, beyond a particular number of team members, team performance tends to decline. Perhaps having too many team members (with very low stress and very high stress) or too few team members (with very low stress and very high stress) might result in the lowest levels of performance (e.g., like the bottom points of an umbrella). However, performance is maximized at moderate levels of stress and approximately eight team members (i.e., the maximum, the top of the umbrella). Naturally, the regression surface would not perfectly model the relations, but based on theory and previous research there may be instances where the functional form is posited to be complex.

Complex functional forms have been used in industrial and organizational psychology and related fields (see e.g., Atkins & Wood, 2002) and have been discussed by various researchers (Edwards & Parry, 1993; Ganzach, 1997; MacCallum & Mar, 1995; Shadish et al., 2002). In short, specifying an incorrect functional form can lead to heteroscedasticity, a threat to statistical conclusion validity, which could result in erroneous inferences.

Summary. Because F_{OLS} performs poorly when the homoscedasticity assumption is violated, one of the practical concerns is: How would a researcher know if it was violated? In this section, I provided basic recommendations describing how a researcher should proceed if heteroscedasticity exists. In addition, I (a) described basic graphical methods for detecting heteroscedasticity, (b) described a variety of tests for heteroscedasticity, and (c) underscored the importance of hypothesizing and modeling the correct functional form.

Limitations and Future Research

Although the present study provided unique contributions to the literature, this study had a number of limitations.

In this study, only 1,000 replications were used for each condition. This affects the standard errors associated with the estimated probabilities. If more replications were used, this would have provided more stable estimates of the Type I error rates and statistical power. Future research should utilize a larger number of replications per condition to obtain more stable estimates (e.g., 10,000 replications in Alexander & DeShon, 1994; 50,000 replications in DeShon & Alexander, 1996). Although increasing the number of replications would have resulted in more precise estimates of the probabilities, this would not materially change the study's findings because the trends would be approximately the same.

In this study, reliability of the variables was not manipulated. Therefore, this study provides no indication of how reliability affects F_{HC3}, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*}. Because unreliability affects the power of F_{OLS} (Bohrnstedt & Marwell, 1978; Dunlap & Kemery, 1988), reliability is an important factor to be considered in future studies. For example, it may be the case that even with very unreliable measures, certain methods are able to control Type I error rates and provide levels of power greater than that of F_{OLS}. If so, this would provide researchers with useful methods to test for the equality of regression slopes.

In this study, range restriction was not manipulated. Considering that range restriction is a topic of great importance in such areas as personnel selection (Gatewood & Feild, 2001; Guion, 1998) and range restriction on x is known to attenuate the power of F_{OLS} particularly with large f^2s (Aguinis & Stone-Romero, 1997), future research should investigate how it affects

F_{HC3}, F_{ML}, F_{RML}, $F_{WLS(1)}$, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*}. Such research may discover that certain methods are less influenced by the effects of range restriction.

In this study, σ_{x_j} was fixed across groups. However, because DeShon and Alexander (1996) have shown that F_{OLS} is still adversely affected by heteroscedasticity when $k = 2$ when σ_{x_j} differs across groups even with equal ${}_2P_j$s and the present study provides further evidence of its poor performance when σ_{x_j} is equal across groups, a future study should vary σ_{x_j} across groups as k increases to evaluate the performance of the alternative methods.

In this study, statistical approximations were not included in the simulation. Specifically, because the *A* and *J* approximations have been shown to perform well (DeShon & Alexander, 1996), these could have been compared with the methods used in this study. However, because the purpose of the study was to focus on those methods that are simpler to compute and permit post hoc analyses, they were not included. To minimize the number of empirical rejection rates to be computed per condition, a future study could compare the *A* and *J* approximations to F_{RML}, $F_{WLS(2)}$, $F_{WLS(O)}$, and F_{WLS*}.

In this study, normally distributed variables were used. Considering that variables measured in psychology and related disciplines are not likely to follow a normal distribution (Micceri, 1989; Wilcox, 2005), this would be a very pragmatic issue to investigate. For example, previous research has shown that the *A* approximation tends to outperform the *J* and *F** approximations at controlling Type I error rates across various patterns of non-normality and that the χ^2 test for the equality of correlation coefficients performed even better than the approximations (DeShon & Alexander, 1996). Based on this, a future study could manipulate normality of *x* and *y* and compare various methods (e.g., *A*, χ^2, F_{RML}, $F_{WLS(2)}$, and F_{WLS*}) to identify which is able to control Type I error rates and provide the highest levels of power. Note

however to include the χ^2 test for the equality of correlation coefficients in such a study a researcher would have to ensure that the simulated conditions were equivalent to the test for the equality of regression slopes (see e.g., Alexander & DeShon, 1994; DeShon & Alexander, 1996).

In this study, only one categorical predictor (i.e., *z*) was considered. Future research should consider more complex models such as those involving two completely crossed categorical predictors with two levels, e.g., sex and treatment (i.e., experimental versus control group). In such a design, a researcher might evaluate whether slopes differ across (a) sex only, (b) treatment only, or (c) all four cells. In addition, this is a standard procedure *prior* to interpreting *adjusted means* in a traditional factorial analysis of covariance, although not necessary if the model is formulated to permit different slopes (Maxwell & Delaney, 2000; Rogosa, 1980; Rutherford, 1992). If heteroscedasticity exists (e.g., across sex only, across treatment only, or all four cells), certain methods may be far superior to F_{OLS}. Although some research has been conducted in this area (Jones, 1968; Liakhovitski, & Stone-Romero, 2000), future investigations may identify which methods perform optimally, say, assuming bivariate normality within cells when testing for the equality of correlation coefficients (e.g., Jones, 1968) or across a variety of other conditions. Such research could benefit various areas including personnel selection.

In this study, *k* did not exceed 4. Future research could evaluate whether the various methods maintain control of Type I error rates and provide power levels greater than F_{OLS} as *k* increases. Such research would be a very practical tool when using traditional analysis of covariance with $k > 4$. It is plausible that some WLS methods may be more adversely affected as *k* increases. Nevertheless, this study provided evidence that Overton's (2001) method could be generalized to $k > 2$ using w_j^* without inflating Type I error rates.

Closing Comments

As mentioned above, this study contributed to the literature on testing for the equality of regression slopes in important ways. Although there were limitations, the contributions were unique and this study provides other avenues of research.

The study's findings are consistent with previous research, showing that F_{OLS} performs poorly in the presence of heteroscedasticity (Aguinis, 2004; Aguinis et al., 1999; Alexander & DeShon, 1994; Box, 1954; DeShon & Alexander, 1996; Luh & Guo, 2002; Overton, 2001; Wilcox, 1997). In addition, F_{HC3} was clearly the most powerful test, but its Type I error rates were greatly inflated. The findings also showed that when heteroscedasticity exists and $k = 3$ or 4, Type I error rates can be controlled to some extent by F_{RML}, $F_{WLS(2)}$, and $F_{WLS(O)}$, but F_{WLS*} provided the most accurate Type I error rates across the study conditions. Furthermore, F_{WLS*} had power levels which approached that of the other methods as N increased, f^2 increased, or heteroscedasticity increased, but without sacrificing control of Type I error rates.

It deserves stressing that the purpose of this manuscript was *not* to suggest that F_{OLS} be supplanted when its assumptions are met. Rather, this research hoped to provide some evidence on the utility of alternative methods which can be used when an important assumption of F_{OLS} (i.e., homoscedasticity) is violated. If such assumptions as homoscedasticity are violated and a procedure is known to perform poorly (i.e., OLS), and more accurate, powerful methods exist (e.g., A approximation, WLS, mixed models), it would greatly benefit research and the cumulative advancement of knowledge if the more powerful alternatives were used.

To this end, the present investigation drew upon extant research from various domains, including econometrics, education, management, mathematics, psychology, and statistics. Because heteroscedasticity is a problem that can adversely affect any discipline that uses F_{OLS},

important, relevant findings in one area can be used to inform research and practice in others, including industrial and organizational psychology, social psychology, and allied fields. Overall, this can improve upon the methods researchers use (as well as how they are used), affecting the validity of substantive sample-based conclusions.

Considering that researchers frequently test for the equality of regression slopes in psychology and the social and behavioral sciences, in general (e.g., Cleary, 1968; Cronbach & Snow, 1977; Linn, 1978; Saad & Sackett, 2002; Smith & Sechrest, 1991), numerous fields can benefit by research that identifies those methods that not only control Type I error rates at the nominal level, but also provide high levels of power to detect hypothesized effects across various suboptimal conditions (e.g., heteroscedasticity). It is hoped that other studies similar to this one will be conducted so that the knowledge produced can be used to develop accurate, powerful methods for basic and applied researchers.

ENDNOTES

Chapter One

[1] An interaction (say, $x \times z$) is a symmetric concept. That is, the effect of x on y depends on z is equivalent to stating that the effect of z on y depends on x. Typically, however, one perspective is of theoretical interest. Consequently, the focal variable is sometimes assigned a distinct name, i.e., a *moderator* (e.g., Aguinis & Pierce, 1998; Saunders, 1956). In some instances, the continuous predictor is referred to as a moderator. In other instances, the categorical predictor is referred to as a moderator. For example, a child's age (x) can be said to "moderate" the relation between low-high exposure to media violence (z) and antisocial behavior (y). An equivalent and often-utilized perspective (e.g., in differential prediction of selection tests due to unequal regression slopes) is to state that race (z) "moderates" the relation between cognitive ability (x) and job performance (y). However, due to its asymmetrical application, this terminology is not used.

[2] A computer program is available online at http://carbon.cudenver.edu/~haguinis/mmr/ which performs the computations for the F^*, J, and A approximations based on user-supplied data (Aguinis, 2004).

[3] Note that OLS is a special case of WLS and that WLS is a special case of generalized least squares (GLS) (Greene, 2003; Neter et al., 1996; Rencher, 2000). GLS is "general" in the sense that it permits the modeling of heteroscedasticity *as well as* the modeling of different covariance/correlation structures among observations.

[4] When correcting for heteroscedasticity across groups, the estimated regression coefficients from OLS and WLS will be identical (Overton, 2001, p. 222). However, in other applications, differences may occur between OLS and WLS parameter estimates when correcting

for nonconstant variance. In such cases, it is recommended that WLS regression be conducted iteratively (Mak, 1992) until OLS and WLS estimates converge. Using this iteratively reweighted least squares approach, typically only one or two iterations are needed to arrive at similar estimated regression coefficients (Neter et al., 1996).

Chapter Two

[1] Note that σ_{x_j} was equal for all groups. This allowed for an investigation of whether the Type I error rates of the various methods (including F_{OLS}) were affected by k, heteroscedasticity, and ${}_kP_j$s, holding σ_{x_j} constant. Recall that Dretzke et al. (1982) also fixed σ_{x_j} to be equal across groups, but considered only $k = 2$. They concluded that F_{OLS} was robust when n_js are equal (see Table 3 in Dretzke et al., 1982). More recently, DeShon and Alexander (1996) performed a simulation with $k = 2$ where the test for the equality of regression slopes was equal to the test for the equality of correlation coefficients, allowing σ_{x_j} to *differ*. They showed that even with equal ${}_2P_j$s the Type I error rates of F_{OLS} were not robust; they became conservative when heteroscedasticity exists. By fixing σ_{x_j} across groups, and manipulating k, heteroscedasticity, and including various tests, the present simulation can be viewed as an extension of the study by Dretzke et al. (1982).

[2] For the jth group, values for n_j and $\sigma_{e_j}^2$ were those described above. $\sigma_{x_j} = 1.5$ for all groups. Under the null hypothesis, $\beta_{y_j.x_j} = 0.5$ for all groups. Under the alternative hypothesis, this was also true except that $\beta_{y_2.x_2}$ was allowed to differ so as to satisfy the requirement that f^2 equal a specified non-zero value (i.e., .002, .01, .02, .05, or .08). Equation 10 was used to solve for $\sigma_{y_j}^2$. Then, Equation 9 was used to solve for $\rho_{y_j x_j}$. To compute $\beta_{y_2.x_2}$, the `Solver` function in MS Excel 2003 was used. It can be used to minimize or maximize a formula by changing

user-specified cells. Alternatively, it can be used to set a formula to a specified value (e.g., f^2 = .05) by changing user-specified cells. More precisely, given the just-noted parameters and equalities, `Solver` was used to find the value for $\beta_{y_2.x_2}$ (referred to as the cell to be changed in MS Excel 2003) that would result in a specific f^2 (referred to as the target cell in MS Excel 2003). The target cell contained the formula by Aguinis et al. (2005, p. 105). Note that all default options were used in the function. However, the precision option was set to 1×10^{-17}. In some instances, the `Solver` function could not find a solution for $\beta_{y_2.x_2}$ that results in an *exact* f^2 equal to the required value, but the difference was miniscule (typically beyond the 12th decimal place) and therefore was retained. For example, for a given condition with parameters as specified above (i.e., n_j, $\sigma_{e_j}^2$, σ_{x_j}, σ_{y_j}, etc.), where f^2 should equal .05, the `Solver` function found the value for $\beta_{y_2.x_2}$ that results in an f^2 = .0499999999999997.

Note that the random number generator in S-PLUS combines a linear congruential and a Tausworthe generator. Multivariate normal data with specified covariance (or correlation) structures are generated using the Cholesky decomposition (Kennedy & Gentle, 1980). For comparability, the results of the data generation algorithm were checked against similar conditions considered by DeShon and Alexander (1996), Dretzke et al. (1982), and Overton (2001).

[3] To test for the equality of regression slopes using F_{HC3}, the `Anova` function in the `car` library was used. This library is described in Fox (2002) and can be downloaded from http://www.socsi.mcmaster.ca/jfox/Books/companion/.

Chapter Three

[1] Note that Type I error rates and statistical power could not be computed using F_{WLS*} for conditions where there were no *df* to estimate w_j^* (e.g., n_js = 24, 8, 8, 8). For Type I error rates, this resulted in 245 observations instead of 252. For statistical power, this resulted in 1,225 observations instead of 1,260.

[2] Based on the skewness statistics shown in Table 6 and an inspection of the marginal distributions of the eight *F* tests, the rejection rates for F_{HC3} showed considerable evidence of positive skew. Consequently, a ln transformation was applied to these rejection rates (where ln refers to the natural logarithm) (Cohen et al., 2003; Kirk, 1995).

Chapter Four

[1] In the software used in the present study, S-PLUS, the primary function for fitting linear mixed effects models is `lme`. That is, it fits linear models with random and/or fixed effects. Similarly, the `gls` function can be used but it does not allow for random effects (Pinheiro & Bates, 2000). These functions can fit models with errors that are correlated, nonconstant (i.e., heteroscedasticity), or both. In these, a hybrid approach is used for evaluating likelihood functions. A very popular optimization method is the expectation-maximization (EM) algorithm which has numerous applications (e.g., missing data analysis, linear mixed effects models, Bayesion methods, and factor analysis) (Dempster, Laird, & Rubin, 1977). The EM algorithm is generally easy to compute and has been shown to quickly progress towards the region where parameters are optimum. However, once near the optimum, the EM algorithm tends to converge slowly. In contrast, another optimization method known as the Newton-Raphson algorithm (Kennedy & Gentle, 1980) is more computer-intensive and tends to be unstable when far from the optimum. However, near the optimum, the Newton-Raphson algorithm converges quickly.

The hybrid approach consists of EM iterations (25 by default) to refine the initial estimates and approach the optimum, then switches to Newton-Raphson iterations for convergence (Pinheiro & Bates, 2000).

REFERENCES

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York: Guilford.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90,* 94-107.

Aguinis, H., Peterson, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods, 2,* 315-339.

Aguinis, H., & Pierce, C. A. (in press). Computation of effect size for moderating effects of categorical variables in multiple regression. *Applied Psychological Measurement*. [manuscript available online at http://carbon.cudenver.edu/~haguinis/pubs.html]

Aguinis, H., & Pierce, C. A. (1998). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods, 1,* 296-314.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 52,* 192-206.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin, 115,* 308-314.

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics, 19,* 91-101.

Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika, 35,* 88-96.

Atkins, P. W. B., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55,* 871-904.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society, A160,* 268-282.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research, 40,* 373-400.

Bement, T. R., & Williams, J. S. (1969). Variance of weighted regression estimators when sampling errors are independent and heteroscedastic. *Journal of the American Statistical Association, 64,* 1369-1382.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349-381). San Francisco: Jossey-Bass.

Bohrnstedt, G. W., & Marwell, G. (1978). The reliability of products of two random variables. In K. F. Schuessler (Ed.), *Sociological methodology* (pp. 254-273). San Francisco: Jossey-Bass.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40,* 318-335.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25,* 290-302.

Bradley, J. C. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica, 47,* 1287-1294.

Brown, M. B., & Forsythe, A. B. (1974). Robust test for the equality of variances. *Journal of the American Statistical Association, 69,* 364-367.

Carapeto, M., & Holt, W. (2003). Testing for heteroscedasticity in regression models. *Journal of Applied Statistics, 30,* 13-20.

Carroll, R. J., & Cline, D. B. H. (1988). An asymptotic theory for weighted least-squares with weights estimated by replication. *Biometrika, 75,* 35-43.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York: Chapman & Hall.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Chambers, J. M. (1998). *Programming with data: A guide to the S language*. New York: Springer.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115-124.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The Earth is round ($p < .05$). *The American Psychologist, 49,* 997-1003.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics, 23,* 351-361.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association, 89,* 177-189.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.

Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika, 70,* 1-10.

Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: Wiley.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York: Irvington.

Dean, A., & Voss, D. (1999). *Design and analysis of experiments*. New York: Springer-Verlag.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1-38.

DeShon, R. P. (2002). Generalizability theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 189-220). San Francisco: Jossey-Bass.

DeShon, R. P., & Alexander, R. A. (1994). A generalization of James's second-order approximation to the test for regression slope equality. *Educational and Psychological Measurement, 54,* 328-335.

DeShon, R. P., & Alexander, R. A. (1996). Alternative procedures for testing regression slope

homogeneity when group error variances are unequal. *Psychological Methods, 1,* 261-277.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis*. New York: Wiley.

Dretzke, B. J., Levin, J. R., & Serlin, R. C. (1982). Testing for regression homogeneity under variance heterogeneity. *Psychological Bulletin, 91,* 376-383.

Dunlap, W. P., & Kemery, E. R. (1988). Effects of predictor intercorrelations and reliabilities on moderated multiple regression. *Organizational behavior and human decision processes, 41,* 248-258.

Edwards, J. R., & Parry, M .E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36,* 1577-1613.

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics, 34,* 447-456.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Fox, J. (2002). *An R and S-PLUS companion to applied regression*. Thousand Oaks, CA: Sage.

Fry, J. C. (Ed.). (1994). *Biological data analysis*. New York: Oxford University Press.

Ganzach, Y. (1997). Misleading interaction and curvilinear terms. *Psychological Methods, 2,* 235-247.

Gatewood, R. D., & Feild, H. S. (2001). *Human resource selection* (5th ed.). Mason, OH: South-Western.

Gentle, J. E. (2002). *Elements of computational statistics*. New York: Springer.

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Delhi, India: Pearson Education.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Hattrup, K., & Schmitt, N. (1990). Prediction of trades apprentices' performance on job sample criteria. *Personnel Psychology, 43,* 453-466.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221-233). Berkeley, CA: University of California Press.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York: Wiley.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721-735.

James, G. S. (1951). The comparison of several groups of observations when the ratios of population variances is unknown. *Biometrika, 38,* 324-329.

Johnson, N. J. (1978). Modified *t* tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association, 73,* 536-544.

Jolicoeur, P. (1999). *Introduction to biometry*. New York: Kluwer Academic/Plenum.

Jones, M. B. (1968). Correlation as a dependent variable. *Psychological Bulletin, 70,* 69-72.

Kennedy, W. J., Jr., & Gentle, J. E. (1980). *Statistical computing*. New York: Marcel Dekker.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Lent, R. H., Aurbach, H.A., & Levin, L. S. (1971). Research design and validity assessment. *Personnel Psychology, 24,* 247-274.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics* (pp. 278-292). Stanford, CA: Stanford University Press.

Liakhovitski, D., & Stone-Romero, E. F. (2000, April). *The statistical power of moderated multiple regression for detecting joint dichotomous moderators*. Paper presented at the meeting of the Society for Industrial and Organizational Psychology. New Orleans, LA.

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology, 63,* 507-512.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute, Inc.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54,* 217-224.

Luh, W. M., & Guo, J. H. (2002). Using Johnson's transformation with approximate test statistics for the simple regression slope homogeneity. *Journal of Experimental Education, 71,* 69-81.

MacCallum, R. C., & Mar, C. M. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin, 118,* 405-421.

Mackinnon, J. G., & White, H. (1985). Some heteroscedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29,* 305-325.

Mak, T. K. (1992). Estimation of parameters in heteroscedastic linear models. *Journal of the Royal Statistical Society, Series B, 54,* 649-655.

Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Erlbaum.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376-390.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Moser, B. K., & Lin, Y. (1992). Equivalence of the corrected *F* test and the weighted least squares procedure. *The American Statistician, 46,* 122-124.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models* (3rd ed.). Chicago: Irwin.

Overton, R. C. (2001). Moderated multiple regression for interactions involving categorical variables: A statistical control for heterogeneous variance across two groups. *Psychological Methods, 6,* 218-233.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-Plus*. New York: Springer.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rencher, A. C. (2000). *Linear models in statistics*. New York: Wiley.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307-321.

Rutherford, A. (1992). Alternatives to traditional analysis of covariance. *British Journal of Mathematical and Statistical Psychology, 45,* 197-223.

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology, 87,* 667-674.

Salgado, J. F. (1998). Sample size in validity studies of personnel selection. *Journal of Occupational and Organizational Psychology, 71,* 161-164.

Sarkar, S., Kim, C., & Basu, A. (1999). Tests for homogeneity of variances using robust weighted likelihood estimates. *Biometrical Journal, 41,* 857-871.

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement, 16,* 209-222.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177.

Schott, J. R. (1997). *Matrix analysis for statistics*. New York: Wiley.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23,* 323-355.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Smith, B., & Sechrest, L. (1991). Treatment of aptitude × treatment interactions. *Journal of Consulting and Clinical Psychology, 59,* 233-244.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

S-PLUS. (2002). *S-PLUS 6.1* [Computer program]. Seattle, WA: Insightful.

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87,* 96-111.

Stone-Romero, E. F., & Liakhovitski, D. (2002). Strategies for detecting moderator variables: A review of conceptual and empirical issues. In G. R. Ferris & J. J. Martocchio (Eds.), *Research in personnel and human resources management* (Vol. 21, pp. 333-372). New York: JAI Press.

Vanable, P. A., Carey, M. P., Carey, K. B., & Maisto, S. A. (2002). Predictors of participation and attrition in a health promotion study involving psychiatric outpatients. *Journal of Consulting and Clinical Psychology, 70,* 362-368.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350-362.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38,* 330-336.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48,* 817-838.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilcox, R. R. (1997). Comparing the slopes of two independent regression lines when there is complete heteroscedasticity. *British Journal of Mathematical and Statistical Psychology, 50,* 309-317.

Wilcox, R. R. (2001). Comments on Long and Ervin. *The American Statistician, 55,* 374-375.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). New York: Elsevier Academic Press.

Willett, J. B., & Singer, J. D. (1988). Another cautionary note about R^2: Its use in weighted least-squares regression analysis. *The American Statistician, 42,* 236-238.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18,* 459-482.

Zedeck, S. (1971). Problems with the use of "moderator" variables. *Psychological Bulletin, 76,* 295-310.