12-1-2003

A Comparative Simulation of Type I Error and Power of Four Tests of Homogeneity of Effects For Random- and Fixed-Effects Models of Meta-Analysis

Lisa Therese Aaron
University of South Florida

A Comparative Simulation of Type I Error and Power of Four Tests of Homogeneity of Effects For Random- and Fixed-Effects Models of Meta-Analysis

by

Lisa Therese Aaron

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Interdisciplinary Studies
College of Education
University of South Florida

Major Professor: Jeffrey Kromrey, Ph.D
Robert Dedrick, Ph.D
John Ferron, Ph.D
Howard Johnston, Ph.D

Date of Approval:
December 1, 2003

Keywords: Meta-Analytic Q Tests, Homogeneity of Effects, Fixed-Effects Tests, Random-Effects Tests, Tau Squared

Dedication

This project is dedicated first and foremost to the One responsible for all hope, love, truth and persistence. When the interpersonal, physical, intellectual and logistical obstacles seemed most demoralizing, it was God alone who carried this work to completion and imbued it with a meaningful and redemptive purpose. Secondly, this effort is dedicated with love to Bruce. Thank you for freeing me to focus on the spaces in between. To Kellianne, my special facilitator, without whose help this work would not have been possible, may this effort be encouragement for the development of your potential and the realization of your goals - however out-of-reach they may seem. Lastly, I dedicate this work to the students in the Title I program who are responsible for inspiring my interest in an otherwise abstract subject. The application of the findings of this project to the evaluation and improvement of their educational experience is what will determine the ultimate value of this assignment.

Acknowledgments

Many friends and acquaintances have offered words of encouragement along the way. And I am grateful to them. But a few provided ongoing support and their active input for the completion of this work. First, I want to thank my parents who always believed in the value of education. Their direction instilled an early and persistent interest in the pursuit of higher education. My thanks to my Major Professor, Dr. Jeff Kromrey, and the other members of my committee for contributing many helpful suggestions in the conception of this project and its completion. Special thanks go to Dr. Howard Johnston for your friendship at a pivotal juncture in this process and for remaining on my committee despite the many disruptions of this endeavor. Thank you to Dr. Jack Vevea for lending crucial technical assistance central to the purpose of this project. Heartfelt thanks go to Dr. Tina Bacon who by way of taking a personal interest in a relative stranger's challenges helped me rein in my anxieties about the process challenges confronting me. My deepest gratitude goes to my brother, Larry, who never questioned the value of the protracted effort, nor my ability to accomplish the goal. Thank you for not only encouraging me and taking an active interest, but also pushing me towards the finish line.

Table of Contents

List of Tables

List of Figures

A Comparative Simulation of Type I Error and Power of Four Tests of Homogeneity of Effects for Random- and Fixed-Effects Models of Meta-analysis

Lisa Therese Aaron

ABSTRACT

In a Monte Carlo analysis of meta-analytic data, Type I and Type II error rates were compared for four homogeneity tests. The study controlled for violations of normality and homogeneity of variance.

This study was modeled after Harwell (1997) and Kromrey and Hogarty's (1998) experimental design. Specifically, it entailed a 2x3x3x3x3x3x2 factorial design. The study also controlled for between-studies variance, as suggested by Hedges and Vevea's (1998) study.

As with similar studies, this randomized factoral design was comprised of 5000 iterations for each of the following 7 independent variables: (1) number of studies within the meta-analysis (10 and 30); (2) primary study sample size (10, 40, 200); (3) score distribution skewness and kurtosis (0/0; 1/3; 2/6); (4) equal or random (around typical sample sizes, 1:1; 4:6; and 6:4) within-group sample sizes; (5) equal or unequal group variances (1:1; 2:1; and 4:1); (6) between-studies variance, τ^2 (0, .33, and 1); and (7) between-class effect size differences, δ_k (0 and .8).

The study incorporated 1,458 experimental conditions. Simulated data from each sample were analyzed using each of four significance test statistics including: a) the fixed-effects Q test of homogeneity; b) the random-effects modification of the Q test; c) the conditionally-random procedure; and d) permuted $Q_{between}$.

The results of this dissertation will inform researchers regarding the relative effectiveness of these statistical approaches, based on Type I and Type II error rates. This dissertation extends previous investigations of the Q test of homogeneity. Specifically, permuted Q provided the greatest frequency of effectiveness across extreme conditions of increasing heterogeneity of effects, unequal group variances and nonnormality. Small numbers of studies and increasing heterogeneity of effects presented the greatest challenges to power for all of the tests under investigation.

Chapter One

Introduction

Background

The purpose of meta-analysis is to discover if some treatment effect or some non-experimental factor consistently exerts influence over a broad, but similar, set of contexts or studies. Researchers hope to expose truth about a given population and whether an influence bears sufficient strength to produce an expected outcome, regardless of other competing forces. By examining the relationship across multiple contexts, samples and measures, one can determine the possible presence of a stable, more generalizable influence. According to Tukey (1969), the search for constant relationships between treatments and outcomes requires "seeking for irremovable complexities, rather than triv[i]al ones and choosing the numerical expression of our variables to make things as simple as possible, while grasping greedily at the unsimplicities that remain" (p. 86). Such an approach to meta-analysis will determine whether there is a single true effect across studies or if differences between studies result from other moderating influences.

Meta-analysis is a secondary analysis. Summary statistics from each of the primary studies included in the secondary sample of studies comprise the data. Prior to conducting a review of the literature, the meta-analyst establishes a set of criteria for directing the gathering of a collection of similar studies. Important factors of the studies (e.g., research design factors) to be collected are coded within each study. The coding used for marking the relevant aspects of each study also serves as the basis for building a model of the extent to which the sample's aggregated features reflect the characteristics of the population of interest. Applying one of several effect indices, effect sizes are computed for each study. In turn, sampling errors are estimated. Lastly, the corresponding Q test is calculated to provide information about the homogeneity/heterogeneity of effects across studies. Based on this result, the decision to pool effects is determined.

When conducting a meta-analysis, the crucial decision to pool effect sizes is confounded by an array of inconsistencies across studies, including differing measures, varying study designs, disparate

statistical tools and sampling error. Since Glass (1976) first originated the term "meta-analysis", the primary interest continues to be the identification of the amount of similarity among studies' effects. Erez, Bloom and Wells (1996) further explain that "The underlying assumption of meta-analysis is that combining information from independent, but similar, studies improves estimates of population parameters over those obtained from any single study" (p. 277). But the comprehensiveness of a sample of studies is becoming increasingly difficult to define, as more studies are being introduced through media other than academic journals.

The major issue when addressing the variability across effect sizes pertains to whether such variation occurs as a result of random variation in the true effect, moderators or sampling error (Bangert-Drowns, 1986). The test of homogeneity (also referred to as the Q test) is a tool used to determine the extent of this variability across studies by simultaneously evaluating the degree of within-study variability for each study within the collection. Bangert-Drowns further explains the test of homogeneity is an extension of Glass's concern with variability among studies, in that it also attends to the "variance associated with each effect size as a summary statistic" (p. 394).

Once the presence of homogeneity/heterogeneity is determined, the focus of the meta-analysis returns to summarization of the effect(s) through the process of approximate data pooling. Approximate data pooling is the statistical practice of combining effect sizes either to compute a common effect (in the case of homogeneity of effects) or an average effect (in the case of heterogeneity of effects). Bangert-Drowns (1986) cites Hedges (1982) and Rosenthal and Rubin (1982) as those who first explicated approximate data pooling. If the Q test results in a determination of multiple effects, the focus shifts to a description of moderators contributing to the varied effects. In such a case, linear regression is often used to model the various effects. Hedges (1982) credits Glass (1978) with the earliest application of regression in meta-analysis as he coded study characteristics as a vector of predictor variables, thereby regressing effect size estimates on these predictors to determine the relationship between the two.

Hedges (1982) devised the Q test of homogeneity of effect sizes, realizing conventional statistics were not applicable to meta-analysis. Unlike ANOVA, the Q statistic can withstand a violation of the assumption of homogeneity of variance without diminished sensitivity to treatment effect variances (Chang, 1993). Conventional statistics applied to meta-analysis are particularly subject to violations of the

assumptions of normality and homogeneity of variance, due to the multitude of measures and statistical techniques employed in the primary studies (Seltzer, 1991; Chang, 1993), resulting in inflated Type I and Type II error rates. Additionally, primary studies contribute their own sensitivity due to violations that are present, but not reported (Keselman et al., 1998). Problems inherent in violations of normality in primary studies are further compounded in meta-analysis.

Many assert that Hedges' (1982) Q is an effective and parsimonious tool for modeling the variability among standardized mean differences across studies (Harwell, 1997). The purpose of the statistic is to detect statistically significant differences, if any, between effect sizes across multiple studies. It tests the assertion that

$$H_o: \delta_1 = \delta_2 = \delta_3 = ... = \delta_k.$$

Hedges (1982) explains:

> The test of homogeneity of effect size (Hedges, 1982a) provides a method of empirically testing whether the variation in effect size estimates is greater than would be expected by chance alone. If the null hypothesis of homogeneity is not rejected, the reviewer is in a strong position vis-a-vis the argument that studies exhibit real variability, which is observed by coarse grouping (p. 246-7).

The true variability refers to the variance of the treatment across studies.

Hedges and Olkin (1985) supply several algorithms for the Q statistic. But the basic, large-sample derivation is

$$Q = \Sigma (d_i - d_+)^2/\sigma^2(d_i)$$

where d_+ is the weighted estimator of effect size, d_i is the population effect size estimate from the i^{th} study and $\sigma^2(d_i)$ is the estimated variance. Hedges and Olkin explain that "The test statistic Q is the sum of squares of the d_i about the weighted mean d_+, where the *i*th square is weighted by the reciprocal of the estimated variance of d_i" (p. 123). There is no differentiation of between-studies variance and sampling error across studies. This factor will become increasingly relevant to the present discussion.

Valid use of this analytic tool relies on probable *a priori* inferences about the relative characteristics of the sample to the population, as expressed by the model. Though methodologists advocate the use of random-effects tests once heterogeneity is found present, this procedure is not generally being employed (as evidenced by a survey to be described later). The determination of homogeneity resulting from the use of the Q test involves both theoretic and statistical implications. As with any statistic,

Hedges' fixed-effects Q provides a less valid analysis under certain conditions and other considerations still require investigation. For example, Harwell (1997) concludes that skewed distributions combined with unequal variances result in inflated Type I error rates for fixed-effects Q. Kromrey and Hogarty (1998) further corroborated these findings and cautioned against the use of Q under these conditions. Another consideration involves the influence of unequal sample sizes on Q test control of Type I and Type II error. Studies within a meta-analysis rarely have equal sample sizes, the primary factor permitting unbiased estimates of effects. Lastly, there is concern for the limited attention given to the influence of violations of normality (Wolf, 1990; Chang, 1993; Harwell, 1997; and Kromrey & Hogarty, 1998) and homogeneity of variance (Harwell, 1997; and Kromrey & Hogarty, 1998) on the control of meta-analytic Type I and Type II errors.

Two models, the fixed- and random-effects, differ in their characterization of the error involved in defining the relationship of the sample to the population. The fixed-effects model is appropriately applied to a collection of studies representing a single population (that is, when the entire population is available for analysis). When applied to a condition where the sample reflects a diverse array of effects, thereby representing a number of differing populations, the model underestimates the variance. The fixed-effects model depicts error only in terms of within-studies variance, whereas the random-effects model separates this error in terms of both within- and between-studies variance. The between- from within-study partitioning is expressed algorithmically through an added component of uncertainty, τ^2. In other words, the random-effects model expresses the effects as a random distribution, not limited to the sample of studies included in the meta-analysis, but including all other study effects not captured by that particular sample. The homogeneity test of effect size variance was developed to investigate whether significant differences are present in effects across studies. However, as several researchers (Erez, Bloom & Wells, 1996 and Abelson, 1997) have noted, the fixed-effects version of Q fails to distinguish between-study variance from sampling error. It is possible that the undifferentiated variance may contribute to the Q statistic's sensitivity to conditions of extreme parameter effects and small sample sizes. Chang (1993) found a pattern of significant discrepancies in which the fixed-effects Q test produced greater simulated power values than theoretical power values when either small sample sizes (and large k) or extreme parameter effects were present. In contrast, the random-effects model did not evidence significant

discrepancies between simulated and theoretical power values. However, Chang notes that population effects were normally distributed for the random-effects model and not similarly controlled for the fixed-effects model.

The theoretical simplicity and computational ease of the fixed-effects model, as well as traditional practice of researcher inferences guiding model selection has kept the majority of meta-analysts firmly entrenched in the consistent selection of fixed-effects. A survey of the *Review of Educational Research* for the past five years confirms the pervasiveness of this strategy (12 of 15 studies employed fixed-effects). The National Research Council (1992) recognized this practice, recommending increased use of the random-effects model to offset the excessive reliance on the fixed-effects model. For many meta-analysts, judgments about homogeneity of effects rely, in large part, on subjective interpretations about the sample's representation of the population in question. As homogeneity rarely characterizes treatment effects in education (Erez, Bloom & Wells, 1996; Abelson, 1997; and Harwell, 1997; Mulaik, Raju & Harshman, 1997), exclusive application of this model is inappropriate.

In addition, recent evidence suggests the fixed-effects Q test is not appropriate when significant heterogeneity of effects (Chang, 1993) and nonnormality are both present (Harwell, 1997; and Kromrey & Hogarty, 1998). The proliferation of studies, advances in the field of statistics and more sophisticated statistical software have all contributed to the interest in applying tests of homogeneity possessing greater robustness properties, as well as the growing body of evidence indicating that use of incongruent models impedes progress.

Whether applying the Q statistic alone (associated with the fixed-effects model) or in combination with the τ^2 (between-studies variance associated with the random-effects model), both statistics evidence higher degrees of sensitivity under certain conditions. For instance, Harwell (1997) found the fixed-effects Q test yielded increased Type I error rates when small sample sizes were paired with a larger number of studies. In contrast, because τ^2 incorporates an additional variance component, it tends to be a more conservative estimate. The random-effects standard error is typically larger (i.e., less precise) than that of fixed-effects models due to the added between-studies variance component, resulting in larger sampling error.

Three tests of homogeneity (the traditional fixed-effects Q, random-effects Q and fixed-effects permuted $Q_{between}$) and the conditionally-random procedure illustrate the manner in which homogeneity tests numerically elaborate the inferences expressed by each particular model. As their names imply, the traditional fixed-effects Q and permuted $Q_{between}$ are used to test the inferences expressed by the fixed-effects model. The random-effects Q tests the inferences extended by the random-effects model. The conditionally-random procedure applies the fixed-effects Q test as the decision-point to first determine the presence of homogeneity. The presence or absence of homogeneity then determines the model to be selected (Hedges & Vevea, 1998). As each model has its corresponding test(s), statistics are used which correspond to the selected model.

Prevalent and indiscriminate model selection results in repeated oversights regarding heterogeneity of effects, thereby obfuscating accurate interpretation of treatment effects in education. The typical application of fixed-effects has inadvertently sanctioned the practice of fitting data to the model. In so doing, the meta-analyst presents a faulty interpretation of an estimate of the mean population effect generated through an invalid analysis. The combined influences of sample size for any given study and the heterogeneity among the true effects determines the precision of each study's estimate of effect (Raudenbush, 1994). Therefore, only a probable estimate of model homogeneity/heterogeneity can render a probable parameter estimate of the mean treatment effect when Q is rejected (Friedman, 2000). Chang's (1993) study substantiates this conclusion as she found significant discrepancies between theoretical and simulated power estimates when the selected model did not correspond with actual homogeneity/heterogeneity conditions. Whether applying a fixed-effects test to a condition of heterogeneous effects or a random-effects test to a condition of homogeneous effects, power discrepancies result. As Chang points out, Type II error is of special concern in the determination of homogeneity of effects as it conveys the false conclusion that a treatment produces some uniform outcome when it actually generates any number of outcomes depending on the sample.

Therefore, to promote accurate interpretation of homogeneity of effects, the primary question under consideration is how do four tests of homogeneity of effects perform under varying degrees of skewness and kurtosis, random within-study sample sizes, differing within-studies variance and heterogeneous effects. Such an investigation should inform meta-analysts' future model selection. There

have been two basic decisional approaches used to select the meta-analytic models and corresponding statistics: 1) decide a priori (as indicated by Hedges and Vevea, 1998) the sort of generalization to be made about the samples or 2) conduct a test of homogeneity of effects (if heterogeneity exists, a random-effects model is adopted vs. if homogeneity is evidenced, a fixed-effects model is upheld). Hedges and Vevea (1998) assert that one must be primarily concerned with the inferences to be made about the population of studies included in a meta-analysis. But others such as Chang (1993) demonstrate the disparity in power between fixed- and random-effects contingent on homogeneity of effect sizes. Without better-reasoned, more consistent and deliberate use of models, data are analyzed and interpreted unevenly and less cogently throughout the field, delaying theoretical advances.

Statement of the Problem

Growing concern about the appropriateness of the fixed-effects model for meta-analysis (National Research Council, 1992; Chang, 1993; Erez, Bloom & Wells, 1996; Abelson, 1996; Harwell, 1997, Kromrey & Hogarty, 1998; and Friedman, 2000) signals the need for more highly defined criteria to select homogeneity tests useful for complex effects characterizing educational data. Since the model directs the selection of statistics, the analysis is only as valid as the extent to which the model expresses the relationship between the population, treatment and other influences on performance outcomes.

Moreover, educational and psychological data most often possess skewed distributions (Lix & Keselman, 1998). The danger of using statistics sensitive to nonnormality and heterogeneity is that they become liberal, inflating Type I error, particularly with an unbalanced design (Lix & Keselman). Once distributional assumptions of a statistical procedure are violated, as is often the case, it is useful to know the subsequent behavior of such statistics. In this way, applied researchers can better assess the extent to which such analyses generate valid results (Keselman, Huberty, Lix, Olejnik, Cribbie, Donahue, Kowalchuk, Lowman, Petoskey, Keselman & Levin, 1998). The greater possibility for heterogeneity of effects in educational studies and the tendency for meta-analysts to favor use of the fixed-effects model suggest that meta-analysts are applying fixed-effects to collections of studies actually containing heterogeneous effects.

Furthermore, previously simulated meta-analyses present conclusions about homogeneity tests based on over-simplified conditions, not realistic to education. Some important investigations have used equal sample sizes and variances, as well as normally distributed primary samples. Langenfeld and Coombs

(1998) express concern for the effect of these data conditions on the resulting magnitude of effects: "our understanding of the influence of varying sample sizes, degree of variance heterogeneity, and type of distribution on ME [magnitude of effect] statistics is not well understood" (p. 15). Hunter and Schmidt (1994) echo concerns about nonnormal distributions' influence on the Q statistic, stating "The development of methods for sensitivity analysis for random effects models is an important, open research area" (p. 391). Equal primary study variances and sample sizes have been shown to permit unbiased estimators for primary study effects (Hedges & Olkin, 1985). The presence of unequal study sample sizes impacts the accuracy of the primary study effect sizes. The paucity of realistically simulated conditions in such research challenges the meta-analyst to apply models without the full benefit of a well-substantiated rationale.

Only realistic simulations based on conditions commonly found in educational settings can support valid conclusions about the efficiency of estimators for each model's homogeneity tests. Therefore, an important aspect of promoting appropriate model selection involves the comparative investigation of the robustness properties of each model's tests to well-controlled and diverse data conditions.

Prior studies comparing the performance of fixed- and random-effects Q tests have not simultaneously controlled for data conditions specific to education. To this point, unequal within-study variances (Hedges & Vevea, 1998), varied skewness and kurtosis (Chang, 1993) and random primary study sample sizes (Kromrey & Hogarty, 1998) have not been addressed simultaneously in a comparative analysis of random- and fixed-effects models. Furthermore, only Hedges and Vevea (1998) have controlled for varying degrees of between-studies variance, τ^2. As Harwell (1997) points out, nonnormal score distributions, unequal groups variances and unequal study sample sizes impact the t statistic in primary studies. As these are common conditions in educational settings, these conditions' effects on meta-analytic tests are important questions.

Purpose of the Study

The purpose of the study is to investigate data conditions typical in education settings to begin to establish a set of criteria facilitating deliberate model selection for optimal model fit. The responsiveness of four tests of homogeneity of effects are compared under conditions of varying degrees of heterogeneity of variance, primary study sample sizes, number of primary studies and dual violations of normality and

homogeneity of effects, as evidenced by statistical power and control of Type I error. Harwell (1997) recommended utilizing random-effects regression in addition to the Q statistic that has been typically generated with fixed-effects regression. Raudenbush (1994) and Bollen (1989) further advised applying a weighted least squares regression, when sample sizes across studies are unbalanced. This second recommendation is supported by Hedges (1982) who found it provided reasonable accuracy for model fit specification when sample sizes were as small as 10. However, scant information is available to indicate whether such a procedure provides adequate robustness for meta-analytic tests.

Research Questions

The study provided meta-analysts with more specific guidelines for the use of each of four tests of homogeneity of effects by addressing the following questions:

1) To what extent is the Type I error rate of the fixed-effects Q, permuted Q_{bet} , random-effects Q and conditionally-random procedure maintained near the nominal alpha level across variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness and kurtosis?
2) What is the relative statistical power of the fixed-effects Q, permuted Q_{bet} , random-effects Q and conditionally-random procedure given variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness and kurtosis?

Limitations

Computer programs simulated data analyzed within this Monte Carlo study in which distribution shapes, study sample sizes, extent of variance across studies and random variation in true effects were controlled. Moreover, the models defining the statistical tools to be used were methodically alternated.

Other potentially influential conditions were not examined. There were a handful of sample sizes and distribution shapes under investigation in the present study. Related to this issue, there were innumerable potential factors influencing the effect size in any given study (Fern & Monroe, 1996). For this reason, one can never be absolutely certain the variance is attributable to the independent variables under investigation.

The data being simulated looked like data typically collected in reading programs found in public school systems, including features of skewness and kurtosis, sample sizes and the numbers of studies sampled. Therefore, these findings should not be considered widely generalizable to other contexts without further investigation and juxtaposition to other contexts. Differing conditions change the robustness of the statistical properties. As Overton (1998) suggests, "...one of the most important considerations in selecting a meta-analysis model is the contextual conditions in which the effect of interest (e.g., a selection test's validity) is to be generalized in theory or in application" (p. 376). In other words, a statistic is only meaningful given the assumptions applicable to the situation in which it is being used. No single statistic can be expected to be robust to all conditions.

Bradley's conservative criterion of acceptable Type I error was applied for purposes of evaluating the performance of each of the tests under the conditions being investigated. However, a number of alternative criteria could have been applied. This particular criterion was utilized based on typical best practices of researchers.

Finally, this study was based on a Monte Carlo simulation, the results were derived from approximations of the data, not real world observations.

Definition of Terms

Several terms employed in the review of literature and description of the study design are defined below:

Asymptotic Robustness Theory (Bentler, 1994) & (Wang, Fan & Willson , 1996) referred to the robustness of certain parameter estimates which maintain independence from the assumptions of normality, homogeneity of variance and homoscedasticity, due to the nature of their computation, as sample sizes increase.

Central Limit Theorem - as n increases, the influence of any outliers in the population diminishes to the extent that the distribution of scores assumes a normal distribution. Or effects of nonnormality in the population diminish as the sampling size (n) increases (Glass & Hopkins, 1984).

Estimate of the true effect θ – Referred to as Ti and consists of both the true effect and the error in estimating the true effect.

Fixed Effects Model - "A statistical model which stipulates that the units of one or more of the factors under analysis (studies, in a meta-analysis) are the ones of interest, and thus constitute the entire population of units. Only within-study sampling error is taken to influence the uncertainty of the results of the meta-analysis" (Cooper & Hedges, p. 535).

Homoscedasticity- The error variance is constant for all observations (Chatterjee & Price, p. 38).

Meta-analysis - "the statistical analysis of the findings of many individual analyses" (Glass, 1977, p. 352).

Moderator Variable – A variable altering the effect of another variable on some outcome.

Monte-carlo Study - A procedure in which an empirical sampling distribution is drawn from a computer-simulated treatment population manifesting specific model violations, then compared against a theoretical (critical) distribution for a given significance level (Kennedy & Bush, 1985).

Random Effects Model - "A statistical model in which both within-study sampling error and between-studies variation are included in the assessment of the uncertainty of the results of the meta-analysis" (Cooper & Hedges, 1994, p. 539).

Robust Methods - Test statistics which generate probabilities with minimal or no discrepancies between the nominal and actual levels of significance when applied to a set of data comprised of characteristics deviating from those parameters recommended for their use (Kennedy & Bush, 1985).

Sensitivity analysis – the process of developing methods which minimize the discrepancy between p values and the nominal p values under a variety of distributions and maintaining high efficiency and stringency over an array of circumstances. According to Seltzer (1991), it involves applying a random-effects model to determine if much change occurs in the second stage model before drawing conclusions about the relationship between two variables. It may change judgments about the magnitude of the relationships (p. 174).

Standard error (of the statistic) estimates the standard deviation of a parameter estimate. It is the standard deviation of the sampling distribution (Winer, 1971).

Total variance of the estimate of true effect - (v*I) consists of both τ^2 and the within-studies or estimation variance (vi)

Type I error - rejection of a true null hypothesis resulting in the faulty conclusion of statistically significant treatment effects.

Type II error - acceptance of a false null hypothesis resulting in the failure to identify true treatment effects.

Unbiasedness - Winer (1971) states: "One criterion for the goodness of a statistic as an estimate of a parameter is **lack of bias**. A statistic is an unbiased estimate of a parameter if the expected value of the sampling distribution of the statistic is **equal to the parameter** of which it is an estimate" (p. 7). Therefore, **unbiasedness** is a characteristic of the sampling distribution, not only the nature of the statistic. So the extent to which a statistic produces an unbiased parameter estimate suggests that over the course of a large number of samples, the mean of these statistics will be equal to the true parameter. Otherwise, the statistic is biased.

Weighted Least Squares - a modified regression procedure applied to obtain weighted estimates of differentially reliable studies. Hedges (1982) also defined weighted least squares as an "estimation of linear model parameters by minimizing a weighted sum of squares of differences between observations and estimates" (p. 256). It is characterized by the following:

$$\hat{\beta}_w = (X^1WX)^{-1} X^1 WY$$

Chapter Two

Literature Review

The literature referenced in this study concerning the relative sensitivity of two fixed-effects tests, the random-effects Q and conditionally-random procedures will be presented in seven sections to extend a rationale for both addressing this research interest and applying the Monte Carlo design. First, the historical and philosophical evolution of meta-analysis is discussed. Next, the four tests are presented. Third, a discussion of the process of model selection, model fitting and hypothesis testing is offered, as well as the rationale for maintaining a flexible, iterative approach. Fourth, potential factors affecting the efficacy of homogeneity tests are presented. Fifth, research literature pertaining to the comparative sensitivity of each of the four tests of homogeneity of effects is surveyed. Sixth, calculation strategies for reducing estimator bias are presented. Finally, a summary of the recommendations of prior researchers is presented.

Historical & Philosophical Evolution of Meta-analysis

The objective of Science is not focused on the unearthing of Truth, as much as the construction of more heuristic problems and perspectives. That is the problems or hypotheses incorporate more overriding factors, not merely the particulars of a single context. The potential for generating "more general problems" propels scientific progress, not the discovery of certain Truth (Popper, 1968). In a similar fashion, Kuhn (1962) suggests the scientist is obligated to "understand the world" and enlarge the scope and precision of the system of its ordering. Popper continues by explaining, in fact, the need for complete objectivity prevents scientific declarations to be anything more than tentative. Similarly, Pearson (Inman, 1994) asserts the scientist works to summarize perceptual data, not necessarily to uncover Truth. It is a descriptive, not explanatory endeavor.

Some, Meehl (1978) points out, argue the social sciences possess few accumulated insights. Glass, McGaw & Smith (1981) echo this concern with respect to meta-analysis, "Although scholars continued to integrate studies narratively, it was becoming clear that chronologically arranged verbal descriptions of

research failed to portray the accumulated knowledge" (p. 12). Post hoc analysis of scientific disciplines reveals that progress ensues over the course of time and many contributions – not readily evidenced in a single event (Serlin, 1987). Kuhn (1962) points out that many question whether the social sciences have adopted a set of rules and standards ("a paradigm") to constrain the generally accepted practice of the scientific enterprise. Acceptance of a paradigm requires a group of practitioners to universally embrace the same set of theories and rules. Without the accumulation of experiences, the group does not possess a shared perspective. If probability is the science of tentative truths (Popper, 1968), statistics is the discipline of determining the extent to which those truths are possible.

Statistics is guided by an imperative to summarize data when possible. We employ this process as a means of extending information (with a given amount of error) about the characteristics of a sample to a population of larger interest. Estimation of a common parameter by combining multiple estimates from similar studies is valued as being a more efficient, accurate and "natural" process (Glass & Hopkins, 1984; Hedges & Vevea, 1998). The more we minimize error, the more we learn about the presumed characteristics of the population. Combining data from similar studies, as opposed to data derived from a single study, enhances population parameter estimates (Erez, Bloom & Wells, 1996), by checking the consistency of the effect across contexts given the multitude of potential moderators within each study (Raudenbush, 1994). Pearson (1904) first derived an average estimate of effects to obtain a typical effect. Tippett (1931) and Fisher (1932) are among the first to combine probabilities across studies. The fundamental objectives are to identify both the patterns across studies, as well as the aberrations from the same. As will be discussed, it is this philosophy which underlies the inferences embedded within the fixed- and random-effects (meta-analytic) models. However, these models maintain varying degrees of focus on patterns or aberrations.

Meta-analysis permits questions not possible within the context of a single study. Present synthesis methods facilitate tests of hypotheses not tested in primary studies (Cooper & Hedges, 1994). Estimating the generalizability of the effects, as well as the nature to which they are specifiable to groups, exemplifies this objective. With meta-analysis, there is an accumulation of studies' results across situations, making it possible to determine whether a relationship exists regardless of specific circumstances. It is this capability that makes meta-analysis more consistent with general scientific inquiry.

The accumulation of scientific knowledge requires the development of testable hypotheses incorporating prior knowledge and theory (Kuhn, 1962). The idea of synthesis evolved from testing for one overall level of significance to identifying an average true effect (s), as well as an explanation of any moderators contributing to the impossibility of a single average effect (Cooper & Hedges, 1994). Significance testing used within the context of a single study does not accomplish this purpose alone (Mulaik, Raju & Harshman, 1997). But meta-analysis integrates the results of multiple related studies to facilitate generalization about a problem or set of factors (Cooper & Hedges, 1994). Meta-analysis refers to "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3). It represents one class of research review, the integrative review (Bangert-Drowns, 1986). Cooper & Hedges (1994) describe research integration as an effort involving the determination of consistencies and the origin of variability across a set of related studies.

Redundancy and multiple contexts, both integral to drawing generalizations about the validity of some hypothesis (Cohen, 1990; Tukey, 1969), are two central features of meta-analysis. This method achieves two important objectives: the determination of result invariance and the enhancement of objectivity. Replication and integration of studies both entail redundancy, a useful property for establishing consistency of results. By integrating studies incorporating similar features, meta-analysis uses multiple studies to assess the robustness and generality of a relationship(s) (Rosnow & Rosenthal, 1989). Similarly, Cohen (1990) states "Only successful future replication in the same and different settings (as might be found through meta-analysis) provides an approach to settling the issue [of the stable influence of some treatment]" (p. 1311). The compilation of studies demonstrating consistent results, regardless of research design differences and treatment administration peculiarities, ensures the insignificance of the researcher's influence and chance, as well as significance of the effect. Because repetition (or repeated outcomes) permits the evaluation of variability and consistency, it is a cornerstone of scientific inquiry (Tukey, 1969; Hedges & Olkin, 1985). Repeated outcomes suggest greater reliability of the data. Popper (1968) asserts:

> "Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested – in principle - by anyone…Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable" (p. 45).

With respect to the second aim, objectivity involves invariance in the manner with which a researcher develops observations, separate and apart from his/her actions or predispositions (Mulaik, Raju & Harshman, 1997). Objectivity increases as the error inherent in any particular study is minimized. But even assuming the absence of researcher bias, invariance does not provide irrefutable support for the presence of an absolute relationship or set of results. As Popper (1968) suggests, other outcomes could still be possible – corroboration of a result does not nullify the possibility of another outcome, as well.

Regarding model selection in meta-analysis, there is much disagreement as to the degree of objectivity required for preliminary efforts taken at the first stage of the analysis involving the determination of homogeneous effects. As meta-analysis involves the compilation of many studies, summary statistics from each of the primary studies are converted into effect sizes. The first stage of the meta-analytic process involves the determination of whether there is a common population effect or multiple typical effects. The disagreement lies in which criterion to use for model selection in determining the homogeneity of effects – either *a priori* selection based on researcher inferences about the comprehensiveness of the sample or selection based on the results of an initial administration of the fixed-effects Q test. The appropriateness of the model selected bears significant implications for the determination of the magnitude of the effect(s) at stage 2 (Chang, 1993).

The null hypothesis for a typical meta-analysis includes the assumptions that there is no difference in the effects from one study to the next ($\delta_1 = \delta_2 = \delta_k$) and that each study's effect does not equal zero ($\delta_k \neq 0$). Another null hypothesis assumes the study effects across classes share a common effect. Properly interpreting the lack of evidence to the contrary does not mean one does not suspend the possibility that true differences exist. Chang (1993) finds that false rejection of the null at the first stage results in z tests (incorporating random-effects variances) with reduced sensitivity for the magnitude of the common effect at the second (p. 121). In an effort to replicate and summarize results, there is an attempt to draw more definitive conclusions about the relationships between a multitude of variables (Raudenbush, 1994), as well as compile, evaluate and integrate the corpus of research studies into a meaningful nexus of information (Cooper & Hedges, 1994). As Kirk (1996) contends, " What we want to know is the size of the difference

between A and B and the error associated with our estimate; knowing that A is greater than B is not enough' (p. 774).

Replication or, in the case of meta-analysis, synthesis of multiple studies contributes to either the corroboration or nullification of a consistent measurement which then generalizes more accurately to a greater number of subjects. Cronbach et al. (1972) stated, "A behavioral measurement is a sample from the measurements that might have been made, and interest attaches to the obtained score only because it is representative of the whole collection or universe" (p. 18). Combining information from similar studies in somewhat different settings provides more conclusive evidence regarding a hypothesis by further enhancing population parameter estimates, as opposed to those generated from a single study (Cohen, 1990; Erez et al., 1996). A single study fails to provide conclusive evidence for the generalizability of a finding, because it does not demonstrate the stability of a relationship outside of one context. In other words, the consistency of an outcome supports its non-chance occurrence and that it truly represents some expectancy.

Meta-analysis involves procedures for both computing an estimate of across-studies effect sizes and an overall significance level for multiple studies addressing similar treatments. A brief overview of the determination of a true estimate of effect versus an overall significance level can help to distinguish these two meta-analytic procedures. Glass is credited with applying a method of combining study results from differing scales of measurement. He demonstrated how Cohen's d, standardized mean difference, or a product-moment correlation coefficient could fulfill the need for a scale-free measure of effect magnitude. Glass, McGaw and Smith (1981) applied analysis of variance and multiple regression for meta-analytic purposes. Effect sizes were employed as the dependent or criterion variable and study characteristics inputted as the independent or predictor variable(s).

Tippett (1931) and Fisher (1932) were among the first to combine probabilities across studies. Tippett illustrated that if p-values are independent, then each originates from a similar distribution under the null hypothesis. Significance tests are a sort of measure indicating the extent to which one ought to be disinclined to accept the null hypothesis under consideration (Mulaik, Raju & Harshman, 1997). As Rosenthal (1978) illustrates, there are at least eight methods for obtaining an overall significance level. Some are appropriate with larger numbers of studies, others with fewer studies. The blocking method,

for instance, may be more powerful than most, but is computationally more complex, particularly with a larger set of studies.

Neither the combination of probabilities nor the generation of a single index is adequate, in and of itself, for the calculation of a common effect size. Each resolves only part of the question as to the possibility of the presence of the relationship and the comprehensiveness of the sample. Meta-analysis circumvents the issue of statistical appropriateness, by deriving its summary statistic from a multitude of parameter estimates or sampling distributions. As stated earlier, both confidence intervals and significance levels guide the generation and/or evaluation of a common effect size in meta-analysis.

There are two problems in the exclusive application of probabilities and statistics. Specifically, Chow (1998) suggests a potential conflict in assuming the presence of a single sampling distribution, supporting the need to evaluate treatment effects using both significance levels and effect size estimates. He states " The validity of statistical power is debatable because statistical significance is determined with a single sampling distribution of the test statistic based on null hypothesis, whereas it takes two distributions to represent statistical power or effect size" (p. 169). Although Fisher's method of combining probabilities may be best known, it is faulty in that, when there are two studies with equally significant outcomes in opposite directions, it supports the significance of either result (Rosenthal, 1978), failing to account for the direction of the outcome.

There is also a basic problem with attempting to interpret effect sizes in isolation. Meta-analysts do not test the opposite of the null, that there is an effect, because it is not an exact hypothesis, for there is no single suggested value (like zero for the null) (Mulaik, Raju & Harshman, 1997). If a treatment group mean is significantly greater than the control group mean, this result is not specified. But when one merely combines effect sizes without presenting confidence intervals, the overall effect can equal zero as all effects are taken into account. Carver (1978) explains:

> "Now if we reverse the question to ask what is the probability that two obtained groups were sampled from the same population, we have the question that most people want to answer and assume they have answered when they calculate the p value from statistical significance testing. In essence they are asking what the probability is that the null hypothesis, H_o, is true, given the type of large mean difference we have obtained, or, what is p (H_o |D^1)?" (p. 385).

Because significance levels draw a single conclusion about the probability of an outcome (s), variation across studies was perceived, prior to the use of meta-analysis, as obscuring rather than clarifying interpretation. Meta-analysis both identifies the variation and its source(s) (Cooper & Hedges, 1994). Cooper & Hedges further assert, that past syntheses were often conducted using the "wrong datum (significance tests rather than effect sizes) and the wrong evaluation criteria (implicit cognitive algebra rather than formal statistical test)" (p. 523).

Comprehensiveness, not consistency alone, best imparts scientific truth (Light & Smith, 1971; and Glass, McGaw & Smith, 1981). Meta-analysts strive to account for discrepancies across studies, in an attempt to identify moderating variables, thereby providing a more complete representation of the population of interest. This objective cannot be accomplished by deriving an overall significance test in and of itself. To this end, meta-analysts call for the combined presentation of both estimates of effect and an overall significance level, accompanied by confidence intervals (Light & Smith, 1971; and Rosenthal, 1978). The presentation of effect sizes in conjunction with significance results helps to contextualize the interpretation of the latter (Rosenthal, 1978; Serlin, 1987; and Cohen, 1990).

Heterogeneous effects generate an estimate of the effect-size, at the second stage of the meta-analysis, which is an average effect (μ_δ) from a distribution of random effects, not a single effect (δ) from a set of equal effects (Chang, 1993). An average effect size computed across a series of heterogeneous effects is a distinctive statistic from the estimated effect size of a common population estimate. The pooling of effect sizes across studies, in the case of the later, assumes a consistent treatment administration to a single population. When an a priori decision or one attained as a result of information gleaned from a test of homogeneity of effects indicates multiple independent populations, effect sizes are no longer pooled into a single estimate of effect. Rather, an average effect size can be computed with the understanding that the statistic does not represent the effect of a treatment on a single population. The effect size representing a whole population presupposes the same strength of influence for a treatment or independent predictor variable on the relatively similar set of characteristics displayed by a uniform population. In contrast, an average effect simply combines effects across studies irrespective of treatment differences and any true differences between groups within a population.

Combining p values involves a nonparametric test, whereas tests of homogeneity of effects are parametric. Becker (1994) describes the p value, or significance levels, as reflecting the probability of seeing a sample as unique as the observed sample assuming the verity of a null hypothesis. Large p's indicate samples are well represented by the null condition. Significance levels or p values are laden with information about sample size or degrees of freedom, as well as the population, not distinguishing these from the influence of the treatment alone. Unlike effect- magnitude indices, combined p's do not present measures of effects free from these other inputs. Only an analysis of effect magnitudes provides information about both the size and the strength of the average effect. Furthermore, these analyses can further pinpoint any moderator variables contributing to inconsistent effects. The omnibus test wherein p values are combined presents information about whether or not the effect is different from zero.

Purpose of Homogeneity Tests

The question driving meta-analysis is whether there is one average pooled effect or multiple average effects. Two questions are subsumed within the one about the breadth of the population – a substantive question with implications for the partialling of variability and interpretation of results. Is one or are several populations being represented by the sample? This question is determined by another – is there variability across effects or merely sampling error? As will be discussed in greater detail, this question is the reason for conducting regression. In integrating studies' results, the objective is to determine whether a single average effect best describes a treatment effect on a population or a random distribution of effects.

Homogeneity tests permit inferences as to whether a set of studies share a common true effect, facilitating greater generalization about a treatment's efficacy from a single context to a broader array of settings. A frequently applied test of homogeneity, Q (Hedges, 1982), tests the assertion that effect sizes are equal (H_o: $\delta_1 = \delta_2 = \delta_3 = ... = \delta_{k,}$), with variability assumed constant across studies.

When the assumption of homogeneity is violated, the Q test possesses an approximate chi-square distribution with a noncentrality parameter. This noncentrality parameter cannot be theorized, because it consists of a combination of multiple sampling distributions. The only alternative is to estimate power using a simulation technique. In another method, the analyst further accounts for the variability in studies' effects by incorporating sampling error and between-studies variance. Three

models apply to meta-analysis: fixed-effects, random-effects and conditionally-random models. Distinguishing features of the first two models will be discussed shortly. Each model employs at least one statistical procedure. For the present study, 2 fixed-effects tests (not including the basic, large-sample derivation), 1 random-effects test and 1 conditionally-random procedure are presented.

Hedges and Vevea (1998) define Q, the general test of homogeneity of effects, as "a comparison of between- to within-study variance" (p. 490). Hedges (1982) states, "The test of homogeneity of effect (Hedges, 1982a) provides a method of empirically testing whether the variation in effect estimates is greater than would be expected by chance alone. If the null hypothesis of homogeneity is not rejected, the reviewer is in a strong position vis-à-vis the argument that studies exhibit real variability, which is observed by coarse grouping" (p. 246-7). Harwell (1997) explains that "…retention of Ho is typically followed by pooling the di and testing the weighted average d+ against zero" (p. 220). Chang (1993), among others, elaborates a two-stage process for determining a population effect.

According to Mulaik, Raju & Harshman (1997), there has to be some criterion for determining whether a small positive variance between effects is small enough to approximate zero. Variance between studies is estimated and then evaluated to determine if the effects are significantly different. The question becomes how one can select the inference without knowing whether there is one or several population effects. Faulty estimation of a common effect can result in the blurring of true differences across studies (Hedges & Olkin, 1985) and the use of erroneous z tests to determine the magnitude of effect (Chang).

The two stage process of effect size estimation involves both Chang's (1993) study of Type II error consequences in accurately computing z tests and the determination of the effect for a particular study, as well as the estimation of a population effect(s) across studies. First, one determines whether studies share a common effect sizes. This step involves the use of homogeneity tests. Second, one computes the magnitude of the average effect.

Despite the possible commission of Type II error, consideration of the influence of violations of statistical assumptions is limited. The recent and infrequent consideration of these assumption violations demonstrates that further study and vigilance is warranted. Concern related to the drawing of valid generalization across studies dates back centuries (Legendre, 1805). Rasmussen & Dunlap (1991) assert many data analysts demonstrate limited concern for the distributional features of their data. Similarly,

Wolf (1990) expresses concern for how few meta-analysts recognize the consequences of violated assumptions on the performance of meta-analytic tests and their failure to investigate the impact of such violations. As recently as the mid-90's, a review of over 400 psychological and educational journals revealed a similar lack of consideration in studies employing various univariate and multivariate designs. The violation of normality received attention from 11% (46 of 411) of the authors, while only 16% (69 of 411) addressed any assumption violations whatsoever (Keselman, Huberty, Lix, Olejnik, Cribbie, Donahue, Kowalchuk, Lowman, Petoskey, Keselman, & Levin, 1998).

Though there is ongoing dispute as to whether one determines the model (whether fixed, random or mixed) prior to or after applying the test of homogeneity, both the theory and data conditions need to be considered as an integrated whole. For this reason, we now turn to the data-analytic limitations and robustness features of each of five tests/procedures. Each is more or less appropriate under certain conditions.

Introduction of Tests

Q tests the null hypothesis that study effects are equal or there is no statistically significant variance among population effects. Essentially, it sums the total of the differences between each population effect and the average of all population effects dividing by the standard deviation of the estimate of the population effect. According to Hedges and Vevea, the Q test can be interpreted as an analysis of between- to within-study variance. If the effect sizes vary, Q has a noncentral chi-square distribution with k-1 degrees of freedom and a noncentrality parameter, λ. $\lambda.=\Sigma\ (\delta_i - \delta.)^2/\sigma^2(d_i)$, where d_i equals the unbiased estimate of the population effect, δ_i. When study sample sizes decrease, Q no longer retains a sampling distribution with a chi-square spread and well-defined degrees of freedom. For this reason, estimated Type I error rates depart sharply from nominal values (Harwell, 1997).

For the appropriate use of Q, primary studies are independent, normally distributed and share a common variance. Typically, there are a more limited number of studies included in the meta-analysis. The population is well defined and finite. The only source of variance under consideration is sampling error or variance introduced by uncontrollable and unknown factors. The Q test incorporates weighted variances in order to account for greater or lesser precision across studies included in the meta-analysis. In this set of circumstances, both the Type I error rate and power of Q approximate theoretical values. But unequal

group variances and sample sizes, as well as nonnormal distributions disrupt this congruence between estimated and theoretical values of Q (Harwell, 1997).

Harwell (1997) found the larger the variance ratios, the smaller the estimated power values. As within group variances became progressively divergent, estimated power values diminished. This finding held for all k, distributions and variance ratios and results from the incorrect pooling of within-study sample variances for the generation of d. Unequal group variances result in d's denominator being overestimated, diminishing the value of d and its power.

Specifically, Q requires the maintenance of the following assumptions to perform optimally: 1) independence of scores in primary studies; 2) a normal distribution; and 3) common variance. To better present the disparity between fixed- and random-effects models, a brief overview of their application follows. For both traditional Q and the random-effects procedure, first apply the regression equation

$$d_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + u_i + e_i$$

Where d_i = the effect size estimate for the *i*th *study*,
β_o = the grand mean effect size
β_1 = the expected mean difference in effect sizes between studies of different classes
X_{i1} = some amount of the first study characteristic in the *i*th *study*
u_i = the residual or component of the score effect size not explained by X, and
e_i = the error of estimation

Calculation strategies used to augment precision. Generalized Linear Models are the largest class of statistical models. They include the more specialized classical linear models – those models restricted to linear relationships. In a classical linear model, there is a systematic component (entailing the X model matrix and β or vector of parameters equivalent to the means) and a random component. The latter part assumes independence and constant variance of the errors or homoscedasticity. There are known covariates influencing the mean and these are measured without error. Unbiased β estimates generated from the minimizing of the least-squares criterion possess minimum variance. Least squares estimates depend on the assumptions of common variance and the independence of the observations from their mean value. Often, these are referred to as ordinary least squares.

The presence of either unequal variances or correlated observations signals the inappropriateness of ordinary least squares (Draper & Smith, 1998). In addition, Lix and Keselman (1998) advise against the

use of usual least squares estimates under conditions of nonnormal distributions and unbalanced study designs (in meta-analysis this situation arises when the number of subjects in each condition of a study varies across studies).

The unbalanced design contributes to the uneven precision of estimated effect sizes. In identifying the extent of variance across study effects, the primary objective is to incorporate estimates of variance that have minimal bias and a degree of precision corresponding to the kind of information yielded by each particular study. The precision of these estimates depends upon the study's sample size and on the extent of heterogeneity across the true effect sizes (Raudenbush, 1994). Studies with larger sample sizes permit more precise estimators of effect and need to be more heavily weighted than estimators derived from studies with smaller sample sizes (Hedges & Olkin, 1985; and Hedges & Vevea, 1998).

The extent of the precision is captured by employing weighted estimators responsible for minimizing variance by using weights inversely proportional to the variance in each study (Hedges & Vevea, 1998). For estimates of effect, the non-systematic variance is "…inversely proportional to the sample size of the study on which the estimate is based" (Hedges & Olkin, 1985, p. 11). The resulting procedure, weighted least squares regression, originates from the class of Generalized Linear Models. Before addressing the statistical procedure used to calculate these weighted estimates, a brief discussion of the Generalized Linear Model is presented.

The Generalized Linear Models extend beyond the classical linear models in that they permit one to consider the patterns of how moderating variables systematically affect the variation in some treatment outcome. As a result, they define both linear and non-linear relationships. Specifically, Generalized Linear Models are applied because they permit two extensions beyond the normal distribution and the identity function between the random component and the link between the random and systematic components. First, the distribution may be derived from an exponential family. Second, the link function "may become any monotonic differentiable function" (McCullagh & Nelder, 1983, p. 27).

Regardless of the use of a fixed- or random-effects model, once the individual variances evidence large differences in variance from one study to the next, residual variances will be unequal (Raudenbush, 1994). In either case, weighting the d's will be more appropriate. Using the weighted least squares will pinpoint the source of variability, as the residuals from the regression help determine the form of the

variances. This procedure generates negatively biased standard error estimates. The bias diminishes as λ_i , the unbiased estimator, approaches 0 or as the residual variance of λ_i diminishes.

The distinction between random- and fixed-effects weighting involves the additional variance component, σ^2_θ, added to the denominator of the random-effects version. Specifically, the random-effects weights are: $w^*_1 = 1/ v^*_1 = 1/(\sigma^2_\theta + v_i)$ whereas, the fixed-effects weights are: $w_i = 1/ v_i$

The weighted mean is incorporated into the calculation of the maximum likelihood estimator:

$$d^+ = \Sigma w_i T_i / \Sigma w_i$$

Note: d^+ is the weighted estimation of the estimated population effect, calculated by multiplying the sum of the weights by the estimated population effect and dividing that product by the sum of the weights.

In conclusion, weighted least squares utilizes study characteristics as the predictors of study outcomes, as well as the estimators of effect size variance unexplained by the model (Raudenbush, 1994). Applying weighted least squares permits improved validity of the overall results, by enhancing the precision of variance estimates, and facilitates further explanation of variance by identifying sources of variability.

The fixed-effects homogeneity test is generally referred to as Q or H (Chang, 1993). Notice the random-effects version of this algorithm incorporates two units in the error term, $u_i + e_i$. In the fixed-effects equation, u_i is absent. The u_i term refers to the true score (effect) variance, whereas the fixed-effects model only considers the sampling error, e_i, discounting the possibility of true variance across studies. As will be seen, this additional component of uncertainty plays a considerable role in increasing the variance and power of the random-effects statistic relative to traditional, fixed-effects Q. The fixed-effects Q test follows:

$$Q = \Sigma_i (d_i - d_+)^2/\sigma^2(d_i) \qquad \textbf{Fixed-effects Test}$$

where d_i = estimate of the population effect size, the minimum variance unbiased estimator of δ_i.

d_+= average of the d_i's $= \dfrac{\Sigma\, \sigma^{-2} (d_i)\, d_i}{\Sigma\, \sigma^{-2} (d_i)}$

$$\sigma^2(d_i) = \text{within-study variance} = \frac{n^E + n^C}{n^E\, n^C 2(n^E + n^C)} + \delta_i^2$$

After controlling for the study characteristics, the variance for d_i is

$$V^*_i = \text{Var}\,(u_i + e_i) = \sigma_\theta^2 + vi$$

With a balanced design and no predictors, proceed as follows

$$Q = \Sigma w_i (d_i - d^+)^2$$

Note: d^+ is the weighted estimation of the estimated population effect and was calculated by

$$d^+ = \frac{\Sigma w_i d_i}{\Sigma w_i}$$

If the design were unbalanced, compute the weights for each study effect accordingly:

$w_i = 1/v_i$ for random-effects weights $w_i = 1/(v_i + \sigma_\theta^2)$

Hedges and Vevea (1998) describe how the additional variance component in the random-effects test makes it a more conservative, less powerful, test than the fixed-effects Q. They explain: "Because the additional component of variance is the same for all studies, it both increases the total variance of each effect size estimate and tends to make the total variances of the studies (the v_i*) more equal than the sampling-error variances (the v_i)" (p. 492). Once Q is calculated for fixed-effects, this statistic can be used to calculate c which then permits the computation of τ^2 for the random-effects procedure.

$$C = \Sigma w_i - \frac{\Sigma (w_i)^2}{\Sigma w_i}$$

Next, test the significance of the effect-size variance component (τ^2) or that H_o: $\tau^2 = 0$

$$\tau^2 = Q - (k-1)/c$$

If this value is larger than zero then one can no longer assume the presence of a single effect. One then recomputes the random-effects weights using the estimate of τ^2.

d^+ for random-effects = random-effects weighted mean effect size or $\frac{\Sigma w_i d_i}{\Sigma w_i}$

Using d^+ for random-effects, one constructs the confidence interval around μ.

Lower limit = d^+ (random-effects) – 1.96(SE)< = μ < =d^+ (random-effects) + 1.96(SE) = Upper limit

Conceptually, the random-effects model interprets the collection of studies as part of a wider and unknown universe, as opposed to the fixed-effects model by which hypothesis testing is restricted to the immediate sample. For this reason, the random-effects statistic accounts for variance in a unique expression. Although it has been treated as a statistic not requiring an assumption of normal distribution of the study effects, there has been recent cause for suspicion of the same. Hedges (1992) suggests this estimator may not be unbiased under conditions of nonnormal random effects.

The between-studies variance component is the distinctive feature of the random-effects homogeneity test (referred to by this study as Q_+, but variously referred to in other literature as H_+),

differentiating the random-effects Q both theoretically and algorithmically. This element is variously referred to as the "between-studies variance component", "estimator of population variance", "estimator of the variance of population effects", "estimator of the population variance component", "heterogeneity of effects" and "treatment x studies interaction" and is variously expressed as τ^2, σ_θ^2, $\sigma_{\alpha\beta}^2$, σ_δ^2 and $\sigma_{\alpha\beta}^2$. For purposes of this study, it will be referred to as the "between-studies variance component" , "heterogeneity of effects or τ^2.

The between-studies variance or τ^2 represents the part of the variance that consists of systematic error. It defines the degree of variation across studies relevant to study-specific treatment effects or refers to the variance of the population from which the study-specific effect parameters are sampled (Hedges & Vevea, 1998). It is added to the sampling error to compute the estimate of the Total variance of the average effect. Moreover, τ^2 entails the variance of the distribution of the errors of θ_i. Typically, this statistic has increased variance due to the added uncertainty built into its algorithm. As stated above, the between-studies variance component (τ^2), is responsible for enlarging the standard error of the mean, making it substantially larger than in the fixed-effects test. This difference results from the increased between-study heterogeneity in the effects (Hedges & Vevea, 1998). In fact, Hedges and Vevea explain that the between-studies variance component is approximately two thirds as large as the average estimate of the sampling error variance. τ^2 is valued as 0 when Q – (k-1) becomes negative because it cannot be negative and is either present or absent (Hedges & Vevea, 1998).

Though it may not be partitioned within the model, Hedges and Vevea conduct their simulation controlling for varying degrees of the between-studies variance, τ^2. When $\tau^2 = 0$ with an unconditional inference, fixed-effects Q generates nominal probability; however, when $\tau^2 > 0$, Q generates a lower than nominal probability. But the same is true to a lesser extent for random- and conditionally-random procedures, particularly when k is small and heterogeneity is large (Hedges & Vevea, 1998). If one assumes τ^2 to be small when it is not, the result will be the underestimation of the variance v* (the sampling variance of the random-effects estimate) (Hedges & Vevea, 1998).

The random-effects algorithm includes the estimate of population variance, the component accounting for the added uncertainty.

$$Q_+ = \Sigma\,(d_i - d_+)^2/\sigma^2(d_i\,|\delta_i) \qquad \textbf{Random-effects Test}$$

Note: The primary difference between the traditional Q and Q_+ is the inclusion of the variance of the sample effect sizes while holding constant the population effect size(s), δ_i. This modification creates the variance of the conditional distribution of d_i given δ_i.

Hedges and Olkin (1985) explain the algorithm applies expected values of the mean squares. These values are represented in terms of variance components. Sample values are substituted for these expected values and used to solve for the variance components (please refer to the procedural description of the fixed-effects test). Weighted least squares are typically applied to enhance the precision of the estimates of variance for study effects. This procedure results in the unbiased estimates of the variance components. A more in-depth discussion of weighted least squares is presented at the end of this chapter.

First, it is important to distinguish the conditionally random-effects Q as a procedure and not a test statistic. Rather, it is a sort of protocol concerning the treatment of the choice to conduct either a random- or fixed-effects test of homogeneity. Hedges and Vevea (1998) state:

> " If the analyst chooses to make conditional inferences (by conditioning on the studies in the data set), the statistical model has been determined because the effect parameters are treated as fixed for inference. If the analyst chooses to make unconditional inferences, the statistical model treats the effects as a sample (even if no real sampling has been done), and thus, they are treated as random effects" (p. 495).

When conditional inferences are made (θ are fixed), random- and conditionally random-effects procedures overestimate confidence intervals (they are too wide). Such overextension depends on K and the degree of between-study heterogeneity (Hedges & Vevea, 1998). The conditionally random-effects procedure performs similarly to the fixed-effects Q when τ^2 is not statistically significant, behaving similarly to the random-effects Q when τ^2 is significant (Hedges & Vevea, 1998).

According to Hedges and Vevea (1998), the label of "conditionally-random" refers to the "choice of random-effects" being predicated on the test that (τ^2) is greater than zero. Because it is a procedure conditioned on the outcome of the null hypothesis of homogeneity of effects, it mediates the fixed- and random-effects approaches. According to Hedges and Vevea (1998), if one is not certain as to the homogeneity of the population, one applies the conditionally-random procedure. But optimally, according to Hedges and Vevea, the researcher will determine the model first and then select the appropriate

procedure based on this *a priori* decision. There is the equivalent practice using the test for determining the homogeneity of effects across studies. Hedges and Vevea credit Chang (1993) with the identification of the two-stage process applied by a number of meta-analysts for the purpose of determining the model in use based on the homogeneity/heterogeneity of effects at the first stage.

If an unconditional inference is desired, it means the effect parameters are being treated as a sample from a population and estimates the mean and variance of that population. The population of effect parameters from which the observed effects are collected is a random sample. The test being conducted concerns the mean effect size

Ho: $\mu=\mu_0$ (usually equal to 0)

In order to explain the specific use of the conditionally-random procedure, Hedges and Vevea (1998, p. 503) describe the decision point in the following manner: "In the conditionally random effects procedures, the random-effects procedures are used if Q is statistically significant, and the fixed-effects procedures, are used otherwise. Thus, the expression for $v_\bullet^C$ conditional on Q …" (p. 503).

For use in making either conditional or unconditional inferences apply:

$$v_\bullet^C = \begin{cases} vQ/k(k-1) & \text{if } Q>/Q_{.95} \\ v/k & \text{if } Q< Q_{.95} \end{cases}$$

This procedure assumes Q to have a chi-square distribution with k-1 degrees of freedom and independent of the estimate of the population effect, d. The sampling distribution of Q is a non-central chi-square variate with a non-centrality parameter. A confidence interval is built around the upper and lower limits of either the mean of the k effect-size parameters, θ, or the mean of the population from which the k effect-size parameters $\theta_1,\ldots,\theta_k$ in the k studies being analyzed were sampled, μ. The selection of the mean depends upon whether the test pertains to conditional or unconditional inferences.

The confidence intervals vary for conditional and unconditional inferences.

$L^C = T - z_{\alpha/2}$ square root $v_\bullet^C < \theta < d + z_{\alpha/2}$ square root $v_\bullet^C = U^C$

$L^C = T - z_{\alpha/2}$ square root $v_\bullet^C < \mu < d + z_{\alpha/2}$ square root $v_\bullet^C = U^C$

Notice that the confidence intervals are built the same for both conditional and unconditional inferences, respectively, with the exceptions of the parameter estimate about which the confidence interval is built and

the formulas used to compute standard error. The θ represents the ("mean of the k effect-size parameters", p. 495) one true estimate of the treatment effect uniform across all studies or the difference between the population of all studies' μ^T and μ^C, whereas the μ refers to the "mean of the population from which the k effect-size parameters $\theta_1, \ldots, \theta_k$ in the k studies being analyzed were sampled" (Hedges & Vevea, 1998, p. 498).

Hedges & Olkin (1985) refer to $Q_{between}$ test as the between class goodness-of-fit statistic Q_B. The $Q_{Between}$ statistic is a chi-square distributed test. The permuted version will be investigated here. Permuted $Q_{Between}$ is a randomization test designed to generate a larger, empirical sample. It tests the null hypothesis that the average effect size is the same across classes of studies. It remains consistent with the fixed-effects model because the only classes that are included in the test are those in the sample (it is a test of between-class homogeneity). It is not possible to permute studies that are not present. It should be noted that testing this null does not yield any direct variance estimates – only a probability statement about the similarity/difference between average effect sizes, not the difference between two or more parameters. For example, if one conducts such a test, they would assume that two grades (3^{rd} and 4^{th} grades) of students had the same effect size. The question becomes what is the probability that we would see data like that which is observed, if grade level is not a moderator variable and assuming the reality that population effect sizes are identical. So this probability statement is obtained without the need for any distribution assumptions. The null hypothesis is expressed as $\mathbf{H_o: \delta_1 = \delta_2}$

The test statistic is based on the total weighted sum of squares. The denominator is the normalized weighted sum of squares of the effect size indices about the grand mean d^{++}. Having a common effect size across studies indicates the $Q_{between}$ possesses an approximate chi-square distribution with k-1 degrees of freedom. The permuted version applies a permutation strategy instead of a using a chi-square distribution.

Kromrey & Hogarty (1998) find this procedure to be more robust to the dual violations of normality and homogeneity of variance. When the $Q_{between}$ Test is employed using a permutation strategy, instead of a chi-square distribution, the Type I error control is well maintained. The primary limitation

noted by their study is its inability to generate a sufficient data set with which to test at the .05 alpha level under conditions of small K (generally 5 or less).

According to Noreen (1989), a randomization test is a "procedure for assessing the significance of a test statistic [and] involves randomizing the ordering of one variable relative to another" (p. 12). Orderings are permutations of the variables relative to each other. Randomization tests are nonparametric and based on an empirical, rather than theoretical distribution. "Non-parametric" refers to the manner in which the nature of the population distribution is not specified explicitly. Almost all permutation tests are non-parametric and vice versa. They do not require random sampling, but are based on random assignment within the study.

"Distribution-free", as opposed to "non-parametric", refers to a test's significance level and is not predicated on the form of the population from which the sample is selected. Randomization tests are distribution-free, while maintaining the data's scale values. Permutations are not quite distribution-free because they still require distribution symmetry. But all distribution-free tests are permutations. The permuted data are computed from other possible random assignments based on a randomization scheme, yielding a new sampling distribution. Test statistics derived from the simulated sampling distributions are compared against the observed test statistic. Based on the proportion of those statistics equal to or greater than the observed statistic, a significance level is computed.

There are two classes of randomization tests: exact and approximate. An exact randomization is generated when all possible permutations are completed. An exact test refers to the probability of causing a Type I error that is exactly alpha (Good, 1994). In contrast, an approximate randomization is the random shuffling of the possible permutations. Not all permutations are conducted. Noreen describes the purpose of approximate tests being to increase the sample (number of orderings), thereby improving the precision of the approximation. Furthermore, approximate tests serve as a time saving measure in conducting permutations, as they do no require the generation of every possible order.

Permutation tests are more complex to implement than conventional statistical tests in that they require the generation of a new sample. Good (1994) provides an outline for conducting a permutation test:

1. Analyze the problem – What are the null and alternative hypotheses?
2. Select a test statistic

3. Calculate the statistic based on the original observation labels.
4. Rearrange the labels and recalculate the statistic for this set. Repeat until the entire set of permutations has a test statistic – generating a new sampling distribution. When one compares the test statistic of the shuffled data (the theoretical) against the estimated test statistic of the original data, and the former is greater than the latter, one is added to the nge counter. If shuffled data ends up being greater than the number of shuffles NS, one computes the significance level. You're counting how many randomizations to give a test statistic – larger than in the original data to get a combined significance level at the end.
5. Decide on the validity of the null hypothesis based on the permutation distribution. Looking to see if the original statistic is an extreme value within the permutation distribution. If so you reject the null hypothesis.

As an example, there may be 4 possible selections, but 24 permutations of these...

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

Permuting this test involves the reordering of study effects by dividing the factorial of K, the number of study effects, by the factorial of R, the number of studies in the smaller group (either treatment or control), which is multiplied by K minus R. For example, if the number of study effects equals 10 and the number of studies in the smaller group equals 4, the number of combinations of K taken R at a time will result in the number of permutations to be conducted. In this case, 210 permutations would be simulated, by computing,

$$K!/ (R!(K-R)!) \text{ or specifically } 10!/ (4!(10-4)!) = 3{,}628{,}800/24(720)=210$$

Once the number of permutations is computed, the permutations are run using the $Q_{between}$ test as follows:

$$Q_{bet}=\Sigma_i (d_i^{+} - d^{++})^2/\sigma^2(d_i^{+})=\Sigma_i\Sigma_j (d_i^{+}-d^{++})^2/\sigma^2(d_{ij}) \qquad \mathbf{Q_{between}\ test}$$

With the summation over *I* classes, and *j* studies in each class.
Where d_i+ refers to average weighted effect size for class *I*,
d^{++} represents the grand mean effect size, and
d_{ij} represents the effect size for the *j*th study in the *i*th class (Hedges & Olkin, 1985).

Permutations may be more powerful than parametric statistics, as they operate using symmetric distributions and/or distributions with small shifts in value. Good (1994) explains that "A most powerful unbiased permutation test often works in cases where a most powerful parametric test fails for lack of knowledge of some yet unknown nuisance parameter" (p. 2). To insure the validity of a permutation comparing samples of two populations requires that both the treatment and the control samples be drawn from the same distribution.

According to Good, permutation tests can be employed with mixed subpopulations of heterogeneous data. This was how Kromrey and Hogarty applied these tests.

Further study of the fixed-effects Q test relative to random-effects Q, the conditionally-random procedure and the permuted $Q_{between}$ test is warranted on the basis of several studies and reviews. As mentioned, Chang's (1993) study illustrates the influence of faulty decisions about homogeneity of effects at the first stage on the computation of magnitude of effect, as power is affected. Subsequently, Harwell's (1997) study of the traditional homogeneity of effects Q test suggests it has unstable Type I and Type II error control under violations of normality, unequal variance and unequal sample size in primary studies. Finding significant influences on Type I error and power, Harwell recommends meta-analysts consider using the random-effects procedure as an alternative to the traditional fixed-effects Q test, particularly under conditions of non-normality and heterogeneous effects. Based on the performance of the traditional fixed-effects Q under conditions of heterogeneity of effects, unequal primary studies' variance and nonnormality, both Harwell (1997) and Kromrey & Hogarty (1998) advise against its exclusive use.

These studies provide the context and structure for the present study design. The research questions directing the present study draw their value from these prior conclusions, designs and methodologies. A brief overview of the variables, inferences and methodology incorporated by these studies follows.

Chang (1993) conducts one of the most extensive power analyses comparing fixed-effects Q and random-effects Q using the following variables: normal or non-central chi-square distributions, variance of parameter effects, the number of studies included in the meta-analysis, the total sample size of a single study and study effect sizes. She also completes a regression to determine the most influential factors contributing to the power of both tests. Lastly, Chang conducts an analysis to determine the extent to which a faulty decision about the homogeneity/heterogeneity at the first stage of meta-analysis results in making a faulty decision to reject or accept the obtained z test value at the second stage.

Harwell (1997) investigates the Type I and Type II error rates of the Q test, by controlling the following variables: skewness and kurtosis, variance ratios within primary studies, the number of studies in the meta-analysis, the N of a single study and the study effect sizes (For the specific values for each variable, refer to the table at the end of this chapter). He conducts this test, under inferences commonly associated with the fixed-effects model. Harwell further manipulates the positive and negative pairing of

the sample sizes and variances. Harwell employs a nominal alpha of .05, a standard power of .8 and Bradley's (1978) criterion for intermediate stringency of α +/- ¼ α, determining the number of recommended simulated meta-analyses to be 5,000 (see Robey & Barcikowski, 1992, p. 286).

Harwell uses an unspecified random-number generator from the 1986 Numerical Recipes to simulate standard-normal deviates. He applies Fleishman's method for transformation of the same. Harwell replicates the nonnormal distributions by transforming normal random variates based on the Fleishman (1978) technique, referred to as the Power Method.

Kromrey and Hogarty's (1998) study extends the investigation first conducted by Harwell, by controlling the same variables, but furthering the inquiry by comparing Q to two more chi-square distributed tests and four permuted tests. The permuted tests include the permuted version of $Q_{between}$, as well as gamma, trimmed d and Cohen's d. In addition to unequal variances, Kromrey and Hogarty (1998) simulate all of the studies using the heterogeneous variance conditions. Like Harwell, these researchers focus on applying tests under inferences consistent with the fixed-effects model. The other major research question this study addresses pertains to the robustness of some common effect size indices: Hedges g, Cohen's d, Trimmed-d and γ_1 .

Keeping the conditions consistent with Harwell's (1997) study, Kromrey and Hogarty vary the same variables and maintain the same values for each level of each variable. Kromrey and Hogarty (1998) use the RANNOR random number generator in SAS to generate normally distributed random variables. Different seed values are used in each execution of the program to yield the random numbers. Nonnormal distributions are replicated by transforming normal random variates derived from RANNOR based on the Fleishman (1978) technique, referred to as the Power Method. Using SAS/IML version 6.12, Kromrey and Hogarty are able to verify the accuracy of the data analysis by comparing the results to the GLM procedure.

Type I error rates are computed for all seven procedures examined by Kromrey and Hogarty, drawn from either 1,000 or 5,000 randomly generated samples for each condition of the study. They apply Bradley's (1978) liberal criterion of robustness to assess the Type I error control for each test under the given conditions and determine the proportion of conditions with adequate Type I error control for each test.

Kromrey and Hogarty (1999) also investigate the power, as well as the Type I error control of both the effect size indices commonly applied in meta-analysis and the tests of homogeneity previously examined in their 1998 study. They employ the same five variables and methodology described in their earlier study. Again, 5000 simulations are conducted.

The Hedges and Vevea (1998) study is unique in that Hedges and Vevea deliberately and methodically test the robustness of the fixed-effects Q, random-effects Q and conditionally-random test under both the inferences associated with fixed- and random-effects. They control two variables in the form of the magnitude of between-studies variance and K, the number of studies included in the meta-analysis. Given that factors such as nonnormality, within-study variance, or within-study sample sizes are not varied, this study provides limited information about robustness issues typical to educational studies.

In addition to concerns raised by the limited robustness of the Q test and other homogeneity tests, concern about the exclusive and indiscriminate use of a single model has been expressed. Initially, several had urged meta-analysts to consider applying the random-effects model (National Research Council, 1992; Erez, Bloom & Wells, 1996; Abelson, 1997). By doing so, researchers avoid the use of the fixed-effects model as the primary default. However, this initiative was more of a general call for applying models with greater discrimination – nothing more specific in terms of conditions best suited for the application of one model or the other.

Despite the preference for its use, the sensitivity of the fixed-effects Q test to heterogeneity of variance and primary study nonnormality is suspect (see Harwell, 1997 and Kromrey & Hogarty, 1998). As the traditional Q test is not always appropriate, it is important to investigate alternatives. Specifically, it is important to know how these operate under dual conditions of heterogeneity of variances and nonnormality, using random sample sizes within and across studies. Before addressing conditions appropriate for use with specific tests, the meta-analytic models are presented in detail.

Distinguishing random- and fixed-effects model. As mentioned previously, the primary features distinguishing the two models are the breadth of the inferences about the sample of collected studies, the degree of variability of the study effects, the number of studies included in the meta-analysis and the treatment of uncertainty. Each of these elements has implications for the other aspects of the model. For

instance, as the number of studies increase so does the potential for expanding the generalizability, as well as increasing the variability of the same. In turn, the treatment of the uncertainty about the error is affected.

The breadth of the inferences translates into how variance is characterized. The sort of inferences made refer to whether the sample is considered to be representative of one single, well-defined population or a wider, less clearly delineated universe of several populations. Invariant or predictably varied study characteristics reflect the need for fixed-effects approaches. In contrast, multiple, unidentifiable sources of variance may be best treated using random-effects approaches. Random-effects tests possess a mathematical mechanism for enhancing the sensitivity necessary to accommodate the increased ambiguity inherent in the model - τ^2, or between-studies variance component. In the random-effects model, true effects originate from a distribution of effects with some variance. The study effects, δ_i, vary randomly around one grand mean, μ_δ. There are two sources of variation in the population effect sizes: 1) the variance in population effects parameters in the population distribution of the effect sizes; and 2) the variance in the estimator about the true parameter value for a study (Chang, 1993, p.26).

As a result of the differences in the treatment of variance, the tests corresponding to these two models tend to yield noticeable differences in power. Tests associated with fixed-effects models can produce narrow confidence intervals, as they do not incorporate this between-studies variance (Erez, Bloom & Wells, 1996). In contrast, confidence intervals generated from random-effects widen as τ^2 increases.

Many researchers maintain that assuming a fixed-effects model is justifiable only if the null test for the homogeneity of effects is maintained and no moderating variables are suspected (Chang, 1993; Matt & Cook, 1994; Erez et al., 1996). Some would argue there is no purpose in combining studies with little or no variability in treatment administration. Either no additional information is contributed (Erez et al.) or the between-studies variance [(τ^2) or (σ^2_θ)] is considered to be trivial or nonexistent.

Another consideration related to variability across study effects pertains to the number of available studies. Raudenbush (1994) suggests the decision to employ fixed- or random-effects be predicated, in part, on the number of available studies. His reasoning stems from the concern that random-effects is not as precise with small numbers of studies. In fact, others have produced evidence suggesting the sensitivity of certain homogeneity tests depends upon the relationship between large k and small N (Chang, 1993; Harwell, 1997; Hedges & Vevea, 1998; and Kromrey & Hogarty, 1998). Raudenbush further recommends

meta-analysts account for unidentifiable numbers of potential moderators of a true effect(s), by treating the true effect(s) of a series of studies as random. In general, the potential for a greater number of moderators increases as a meta-analysis incorporates more studies.

For the RE summary to be valid, it relies on both accurate estimation of τ^2 (otherwise expressed as $\sigma^2_{\theta,}$ σ^2_{δ} or τ^2) and an adequate number of studies. Moreover, valid generalization is predicated on a clearly-defined population (Raudenbush, 1994). If $\tau^2 = 0$, there is a common effect or the conditional variance of d = the unconditional variance of d. It is the variance of the population distribution of effects. If homogeneity between effects is not present, then a reviewer can categorize the effects by group, testing each for homogeneity of effects. Essentially, the fixed-effects model assumes the presence of one universal effect. In contrast, the random-effects model assumes each treatment produces its own effect and is derived from a universe of similar, but distinct, treatments. For this reason, the effects are most accurately modeled as a distribution of true effects.

A final consideration involves the modeling of uncertainty. In the fixed-effects model, there is one source of uncertainty concerning participant sampling. It relates to within-group sampling error, corresponding variance is σ_E^2. Conversely, the random-effects model includes 3 sources of uncertainty: within-group sampling error, σ_E^2, the random effect of the study, β, and the interaction between the treatment and the study or $\alpha\beta$. Winer (1971) contends it is imperative that experimental procedures closely reflect mathematical models to ensure valid prediction of experimental results. In other words, sampling methods must be accurately expressed in the algorithms used to model the variables being investigated. The meta-analyst must have a clear conception of the degree of representation emulated by the collection of studies to a specified population. If significant heterogeneity is present, it is a fairly clear indication that a single treatment is not responsible for the resulting effect. Because errors' influence is accounted for in the expression and analysis of the random-effects model, it has been argued that it is more consistent with most other statistical methodologies than the fixed-effects approach (Erez et. al, 1996). As mentioned earlier, the fixed-effects approach assumes a uniform model for any given set of studies. Both Erez et al. (1998) and Harwell (1997) encourage meta-analysts to utilize random-effects models in an attempt to discourage the inappropriate use of fixed-effects statistics. The National Research Council (1992) also recommends

reviewers apply the random-effects model with greater frequency to avoid the more restrictive assumptions underlying the fixed-effects model.

Model selection. Model selection is usually determined by either the tenability of the assumption of homogeneity of study effects (Chang, 1993; Erez, Bloom & Wells, 1996) or the researcher's theoretical inferences about the relationship of the sample to a population (Hedges & Vevea, 1998). As will be discussed, model selection bears important implications for the power of the test and its sensitivity. For this reason and the fact that inferences are based solely on hypothetical judgments, not certain truths, a homogeneity test is probabilistic and not absolute, equal consideration of both is necessary. Such a decision does not nullify the influence of theoretical judgments when variables are interpreted as being sampled from either a larger population of studies or a fixed and clearly defined population (Shadish & Haddock, 1994).

As specified in chapter 1, meta-analysis involves a two-stage process. First, the reviewer attempts to determine whether effects are equal (or whether there is variance between study effects). In other words, is there a common population effect size? The second hypothesis refers to the question of whether the true effect is greater than zero. If the homogeneity assumption is maintained, is the common effect size equal to zero? Or to put it another way, does the treatment have a significant, non-random, effect on the population to which it was administered? These hypotheses are expressed as the following:

H_{o1}: $\delta_1 = \delta_2 = ...\delta_k = \delta$ (There is no difference between study effects.)

H_{o2}: $\delta = 0$ (The common effect size is equal to zero.)

The models being discussed differ primarily in how they characterize the uncertainty of the variance across study effects.

With respect to the initial selection of either fixed- or random-effects models, inappropriate model use at the first stage bears consequences for the decisions made at the second stage of the process (Chang, 1993; Hedges & Vevea, 1998). Specifically, Chang finds when homogeneity of effects is falsely rejected, the application of a z test at the second stage correlates with more inflated Type I and Type II error than if the same assumption is falsely maintained. This result is especially affected by a large number of studies each with small sample sizes. Hedges & Vevea note similar outcomes using unconditional inferences (typically associated with the random-effects model). When a fixed-effects test

is applied, inflated Type I error will ensue, unless there is perfect homogeneity of effects. Under such conditions, the random- and conditionally-random procedures provide results closer to the nominal α than fixed-effects tests. Given the importance of the decision and the differences inherent in each model (both in terms of the inferences made and the partitioning of variance), we turn our attention to the issue of how model selection is presently being conducted in the field of meta-analysis.

Two schools of thought present distinctly different rationales for addressing the decision of model selection. Conventional statistical wisdom suggests that tools are selected based first on the theoretical purpose and later modified by the data's distribution characteristics (Tukey, 1969; and Cohen, 1990). In meta-analysis, there appears to be a more distinct and less integrated process adopted by each of two camps, focusing primarily on either theory or data-analytic concerns. One group of synthesists demands *a priori* selection of a model based on inferences formulated from theoretical knowledge of the sample of studies and the degree to which it represents the population. The other group employs the test of homogeneity of effects to determine the appropriateness of the model based on the likelihood of its true description of the variability across sample effects - an indicator of the presence of one or several population effects. It is noteworthy that Kromrey and Hogarty (1999) conclude that the traditional Q test of homogeneity of effects has limited, if any, statistical power rendering it a poor instrument for this purpose.

The opposing viewpoints are captured in the following:

> Abelson (1997) posits: "Empirically, considerable heterogeneity of effect sizes is quite often found in meta-analyses. We can argue abstractly all we want, but in the end, we must attend to behavior of our methods when confronted with real data. The assumption of constant true effect sizes is rarely sustainable' (p. 123-4).
> In contrast, Cooper & Hedges (1994) state: "Conceptual criteria would be applied first, with the model chosen according to the goals of the inference. Only after considering the object of the inference would empirical criteria influence modeling strategy…" (p. 526).
> Lix & Keselman (1998) state: "…the researcher needs to be clear on the goals of data analysis prior to choosing a particular method of statistical inference…" (p. 411).

In the first camp, inferences about the population are restricted to the group of values of predictor variables represented in the sample. Generalizations about treatments apply to similar treatments, even if not controlled in the study. Inferences are based only on studies collected in the sample. Hypothesis testing relates only to the present collection. Generalizing beyond the collection is possible only subjectively.

Alternatively, the collection of studies may be viewed as a result of chance. Generalizing beyond the immediate set of studies is necessary in setting up inferences about the sample and is done statistically. Population values of effect are random samples from a distribution of effects.

Other factors used to distinguish meta-analytic models are the data conditions under which the statistics operate most effectively (Kesselman, Huberty, Lix, Olejnik, Cribbie, Donahue, Kowalchuk, Lowman, Petoskey, Keselman, & Levin, 1998), the implemented sampling procedure and the number of studies incorporated in the meta-analysis. In these cases, data analytic goals are clarified, followed by the selection of a statistical method (Lix & Keselman, 1998). In contrast, another perspective asserts that the theory directs the use of a particular sampling procedure and the parameters used to define the extent of a sample's representation of the population (Serlin, 1987).

Rasmussen and Dunlap (1991), Raudenbush (1994) and Hedges and Vevea (1998) agree that the decision to apply one model over another hinges, in part, on the number of studies to be included in the MA. Smaller numbers of studies would be more validly analyzed, using fixed-effects. In contrast, larger numbers of studies would permit valid generalization to a large possible universe or population.

Both the theoretical and the data-analytic perspectives overlook the tentativeness and reiterative nature of the scientific process. As McCullagh and Nelder (1983) assert, one never knows with certainty whether a model is accurate. In the case of the theoretical camp, one cannot be certain whether the selected studies are inclusive of a single population or multiple populations. Those in the data-analytic camp are initially applying a test based on a fixed-effects model's partitioning of variance. Further, this test has been shown to have little, if any, power to detect true differences in study effects (Kromrey & Hogarty, 1999). As the departure of these meta-analysts' approaches suggest, the choice between fixed and random model use is debatable because it pertains both to the nature of scientific inquiry, as well as the type of data deemed applicable for answering research questions (Hedges, 1994).

A similar dichotomy exists in the field of Measurement pertaining to generalizability theory. When conducting a D-study for purposes of drawing data for decision-making, the set of measurement conditions are treated as either fixed or random. Specifically, when these conditions are viewed as fixed, the intent is to restrict generalizations to those conditions appearing in the study. However, treating the set

of conditions as random involves treating these as if they were a sample from a broader universe of conditions to which inferences will be made to the larger universe of conditions.

Though it is a process of quantifying observed objects, measurement, and ultimately model specification, is a system of human abstraction. Nunnally (1967) emphasizes that measurement (the "rules for assigning numbers to objects to represent quantities of attributes") relies on a system of abstraction to measure attributes of objects or people (p. 2). "If a measure is intended to fit a set of axioms for measurement (a model), the closeness of the fit can be determined only by the extent to which relations in empirical data meet the requirements of the model" (p. 8). The underlying implication is that measurement and model development is a subjective and imperfect endeavor.

The extent to which a sample represents a population is predicated on the sampling procedures, the degree of disparity in primary studies' administration of the treatment, and the modeling of the uncertainty. Meta-analysts first establish their theoretical inferences by determining the criteria used for primary study selection, as random sampling is not typically possible. Serlin (1987) suggests, " Theory must guide the selection of a sampling procedure and theory must determine in what ways the sample should be representative" (p. 366). Initially, the reviewer must establish a criterion for the selection of primary studies. This criterion dictates the characteristics of the sample of accumulated studies. In turn, sampling procedures influence the characteristics of the sampling distribution of the statistic.

When one attempts to analyze a set of data with tools selected solely on the basis of a desired set of inferences, tools and data are not necessarily compatible. Therefore, the tools may not generate a valid analysis. Any random sample bears a unique set of distribution characteristics (Keselman et al., 1998). Independent of sampling, Keselman et al. explain that " Every inferential statistical tool is founded on a set of core assumptions. As long as the assumptions are satisfied, the tool will function as intended. When the assumptions are violated, however, the tool may mislead" (p. 351). As a result of any mismatch between the sampling procedures, in particular, and the population of interest, the sample data can possess unexpected distribution characteristics.

To promote optimal model selection, it is necessary to understand the principles defining a well-specified linear model. The value of a model lies in both its ability to promote the summary of data based on its differentiated presentation of systematic effects as well as the nature and magnitude of random

variation (McCullagh & Nelder, 1983). Furthermore, a good model is characterized by fitted values which minimize some criterion (either a discrepancy or closeness measure); parsimony of parameters included in the model; and scope. Finally, they explain that value lies in describing not just the systematic variation of the data under immediate investigation but patterns potentially occurring in future data of a similar nature.

Based on these objectives, a presentation of the process initiated with model selection and culminating in sensitivity analysis follows. First, a model can be defined as a symbolic representation of the elements specifying the relationship between the x's and y's, including the characterization and partitioning of systematic and nonsystematic variance. The statistics are employed as the vehicle through which one makes probability statements about the viability of each of those elements in relationship to each other as expressed by the model**.** In other words, the statistic is a mathematical abstraction of the empirically-based model (Stevens, 1968).

The model is based on the researcher's theory-based perceptions of how the variables relate to each other. Theories are hunches about some phenomenon based on observations (data). Furthermore, the scaffolding for these definitions is based on former researchers' hunches. The data is collected based on operational definitions and measures of some behavior (these too are developed based on the perception of the researcher). And the statistics used to analyze the data are constructed on the basis of a set of assumptions about the data. Keeping the connection of the nature of model development and the origin of the measures and statistics in mind while reviewing the process of their development and utilization may illuminate the influence each contributes in conducting tests of parameter estimation, as well as the subsequent interpretation of data analyses. Finally, the influence of previously unspecified moderating variables changes the dynamics of the relationship captured by the model.

Model selection is the first step in a process of establishing a model for use in identifying statistics appropriate for testing the relationship characterized by the model. It is the determination of a general class of models. Secondly, model specification is conducted for explanatory meta-analysis. Most meta-analysts treat the moderator variables as if they were causes in different effect sizes. For this purpose, model specification is a strong assumption. Having the right regressor or moderating variables and correct functional form – linear or non-linear and/or non-additive – follows model selection. Model specification, as defined by Hedges and Olkin (1985), is an expression of the way in which "estimates of regression

coefficients approach their respective population values" (p. 172). They further explain that the "estimates of coefficients in linear models are consistent when the variables that actually determine the dependent variable are included in the model" (p. 172). Model fitting is the third step of estimating the regression coefficients. Model fitting involves generating estimates of the parameters and checking the residuals by way of sensitivity analysis. Model fitting subsumes model checking. Model fitting refers to the determination of the distance or closeness of the theoretical values (derived through the use of the model and observed outcomes) from the representation of the linear relationship of the observed y's and the selected covariate x's. According to Hedges (1994), specification of the population directs how the synthesis results will be interpreted.

The relationship between theory and sampling makes model selection inextricably related to the sensitivity of the statistical test, as sampling produces the distribution characteristics. To properly interpret the goodness of a model, one needs a fair and representative sample of the population, as well as a statistic that effectively filters the noise while remaining sensitive to the detection of the signal. Theoretical assumptions about the population guide the sampling procedure. Determining from where the sample is to be drawn requires the *a priori* demarcation of the parameters of the population. The sampling distribution of the statistic is based, in large part, on the distribution of the sample. In turn, the statistic is used to make a decision about the likelihood of the relationship between sample and population. Presumably, the statistic reflects the principle theoretical underpinnings of a given model. The assumptions are essentially the rules for use, employed because they are the conditions under which the statistic is best suited to detect true variance.

Ideally, a statistical test would possess two primary features. The α or significance level of this statistic would equal zero and the power, 1-β, would equal 1 or 100%. However, without perfect knowledge about the true state of the conditions being investigated, confirming these parameters is impossible (Good, 1994). A statistic is unbiased when it's average (derived from multiple samples) equals the parameter it estimates. As Winer (1971) explains, "One criterion for the goodness of a statistic as an estimate of a parameter is lack of bias. A statistic is an unbiased estimate of a parameter if the expected value of the sampling distribution of the statistic is equal to the parameter of which it is an estimate." (p. 7). Unbiasedness refers to a feature of both the sampling distribution as well as the statistic. Efficiency refers

to how precise the statistic estimates a parameter. That is the more precise an estimate, the more restricted the confidence interval or smaller the standard error. The third property to be considered when evaluating estimators concerns the consistency of each estimator's successive approximation to the parameter as sample size increases. As the sample size increases, regardless of the degree of bias, an estimator will more closely approximate the parameter's value. As variance across study effects diminishes, the precision of the population effect estimate increases.

As violated assumptions can result in misinterpretation of statistical results, application of a corresponding model statistic that fails to parallel the parametric conditions will lead to faulty interpretations of the results. Assumptions are the conditions for appropriate use of inferential statistics (Keselman et. al., 1998). Similarly, when a statistic reflecting one model's treatment of the variance of effects does not parallel the reality of the data's population derivation (either from a single population or distribution of population effects), a faulty interpretation results (Becker, 1994). With respect to meta-analytic model selection, Erez et. al. (1996) points out that homogeneity of effects rarely occurs due to the variety of constructs studied in psychology and other social sciences. Moreover, certainty about the inclusion of all relevant studies is not possible (Glass, McGaw & Smith, 1981; Abelson, 1997). In the presence of uncertainty wherein one pools effects and derives estimates of these, the possibility remains that a variety of effects impacts these estimates via an unreported moderating variable (Mulaik, Raju & Harshman, 1997). For this reason, some argue against the assumption of homogeneous effects (Abelson, 1997). Others (Hedges & Vevea, 1998) state heterogeneity alone is not a reason to avoid using fixed-effects. Rather, if a limited set of studies is drawn and inferences are restricted to that set, selecting a fixed-effects model is the only reasonable option.

Ultimately, it is important to understand how the model contributes to test sensitivity. According to Good (1994), sensitivity analysis is the process of developing methods which minimize the discrepancy between p values and the nominal p values under a variety of distributions and maintaining high efficiency and stringency over an array of circumstances. A test's power is determined by how likely it is to detect true differences between populations. The sensitivity of one test is compared to that of the others (Good, 1994). Hedges and Olkin (1985) have described that failure to differentiate systematic variation from estimates of error undermines the sensitivity of the statistical test for systematic variation. In fact, the extent

to which the systematic variance is partitioned from the nonsystematic variance, the random error, determines the sensitivity of the test and the extent of bias. Because the test is devised to correspond to and test the underlying assumptions of the model, higher Type II error rates result from misspecified variance in the model.

As a result of the fixed-effects model's failure to differentiate between-studies variance from sampling error, the fixed-effects Q possesses little, if any, power to detect heterogeneous effects. An example of the effect of poor model specification on test sensitivity lies in the fixed-effects model and the Q test. As will be described, Kromrey and Hogarty (1999) find the Q test has no practical utility for detecting true differences in study effects, as its power is low when applied to data with heterogeneous variances. It also evidences unstable Type I error control. For a test to be useful, it needs to control both Type I and Type II errors. For these errors to be well controlled by the statistic, it requires proper partitioning of the systematic variance from the non-systematic variance. Kromrey and Hogarty (1998) also find the iterated Q (Hedges & Olkin, 1985) to have no appreciable improvement in Type I error control under similar conditions.

The literature supplies little explicit attention to determining the necessity for or superfluity of direct model to statistic correspondence. This consideration is important and bears directly on the model selection process. More discussion is warranted, but it is beyond the purview of the present study. The understanding permitted by simulating and reporting the effect of common data conditions on the four tests' sensitivity will provide an apples-to-oranges knowledge of the relative sensitivity of each. Realizing the elusive nature of Truth, it is valuable to identify actions clarifying the relationship, if any, between applied models and statistics.

Based on the numerous points for misinterpretation due to the extent of subjectivity involved in each process from model development and model specification to test selection, maintaining a flexible, open-ended approach to model selection and test use is crucial to the integrity of the overall scientific endeavor. Similarly, McCullagh and Nelder (1983) recommend an iterative process whereby model selection is followed by model checking to be reiterated as long as necessary until the best plausible model is identified and verified (model fitting), before summarizing the results and drawing conclusions. There is a need for ongoing model checking, while maintaining flexibility in model selection. The selection process

would be best guided by treating both inference goals and data conditions with equal consideration. Returning to the objective of Science referred to at the beginning of this chapter: it is to develop more heuristic problems, not to suggest certainty of a single theory (Popper, 1968). That is the problems or hypotheses incorporate more overriding factors, not merely the particulars of a single context. As Kuhn (1962) suggests, the job of scientists is to magnify the scope and precision of ordering a system of understanding. Due to the need for flexibility in model development and implementation, as well as the many factors contributing to test sensitivity, it is most important to maintain an approach of informed vigilance, as opposed to adherence to a single model and test. But one cannot place too much "emphasis on inference", because rigid use of a statistic results "in a loss of flexibility in data analysis" (Cohen, 1990, p. 1310).

In an effort to support well informed, context-based decisions of meta-analytic test use, some general concerns related to the minimization of bias in homogeneity tests will be presented in the following section. The process of refining model to statistic correspondence will gain greater clarity by understanding conditions where Type I error is minimized and power is increased, as well as becoming knowledgeable about the conditions that increase the precision of tests while improving model fit. Such knowledge will facilitate future decision-making about the tenability of models. To this end, this study attempts to address issues for better refining our knowledge of the application of four tests of homogeneity to specific data conditions.

By now, it should be clear that any judgments made about the adequacy of one statistic relative to another are predicated on the specific model and data conditions under consideration. Additionally, once distribution assumptions of a statistical procedure are violated, as is often the case, it is useful to know the impact on subsequent performance of these statistics. In this way, applied researchers can better assess the extent to which such analyses generate valid results (Keselman, Huberty et al., 1998).

Implications for Test Selection

Because the problem of interest, in meta-analysis as in all data-analysis, is one of isolating and controlling "sources of artificial variation across studies" (Cooper & Hedges, 1994), it is important to recognize all possible sources of variation, including sampling error, true variance due to treatment differences, variance due to differences in populations and bias introduced by the test. The question

becomes one of determining whether the variance across study effects is attributable to sampling error or something meaningful to the true variance (either due to differences in treatment or differences in the treatment's effect on multiple populations). Depending upon this initial determination of variance, the researcher will either pool the effects or develop a regression model to classify the remaining variance. It should be noted in the case of the permuted $Q_{between}$, one starts out by testing to see if the average effects across classes are equivalent. The moderator variables have been identified *a priori*. Using tests that minimize bias is integral to the validity of the analysis.

As described earlier, bias contributes to Type I and Type II error, wherein the investigator fails to accurately estimate the parameter(s) of interest, thereby drawing faulty conclusions about the accuracy of the null hypothesis. In general, larger sums of squares are associated with smaller standard error of the regression coefficients. So extreme values for X enhance statistical significance tests (Pedhauzur, 1982). Furthermore, when unequal error variances go unchecked, parameter estimates have large standard errors, though appearing unbiased, resulting in diminished precision and tests with low sensitivity (Chatterjee & Price, 1977). The precision of an estimator is generally measured by the standard error of its sampling distribution (Winer, 1971). A smaller standard error enhances the precision of the estimate.

For clarification, a description of the true states of reality assumed by each error type follows. The Type I error assumes the null hypothesis is false, even though it is true (there is no difference between effects and there is one true effect, but one rejects it instead). Underlying a Type II error is the assumption that the null hypothesis is true, when it is, in fact, false (there is a true difference between effects, but one accepts the false null). In general, power bears a positive concomitant relationship to sample size, effect size and α. As these three variables increase, power increases.

When assumptions are violated or certain factors present, several researchers (Chang, 1993; Harwell, 1997; Hedges & Vevea, 1998; and Kromrey & Hogarty, 1998) find that both the traditional fixed-effects and random-effects Q tests are subject to inflated Type I error. Inflated Type I error is especially problematic because it suggests good power, though power values are likely to be artificially enhanced (Chang, 1993). With respect to many tests of homogeneity, the null hypothesis states there is a uniform treatment effect or there is no significant between-studies variance. A difference suggests the possible presence of a true treatment effect. At the meta-analytic level, the researcher is interested in identifying

differences, if any, between groups from one study to the next. Significant variation across studies reflects multiple, not single, effects may have been manifested. Multiple effects do not render meaningful generalizations, unless moderator variables are isolated and analyzed.

Now discussion is turned to the conclusions drawn in the literature about the specific conditions and their influence on the control of Type I and Type II errors for each test of homogeneity under present investigation. Additionally, those conditions and tests requiring further study will be presented as identified by the authors of the previous research or as indicated by the omission of variables in those studies. As revealed by the Review of Educational Research survey mentioned previously, Q is presently used as the common default. Therefore, factors influencing its use will be discussed first, so as to prepare the reader for the conclusions drawn about the comparative performance of the three alternative tests examined in this study. Before proceeding to the specific conditions for test application, general information about some factors relevant to the use of all tests of homogeneity is presented.

Typically, data collected from educational and psychological settings are characterized by skewed distributions (Lix & Keselman, 1998). Distortions in data-analysis most often arise when skewness or kurtosis is greater than 1 in absolute value (Wang, Fan & Willson, 1996). Most traditional homogeneity tests lack robustness to violations of normality to begin with, resulting from the relationship between the sampling variance of the sample variance and the kurtosis of the population (Raudenbush & Bryk, 1987). The presence of both skewness and kurtosis exerts greater influence on an analysis than either distribution trait alone (Wang, Fan & Willson, 1996). Normality is an important consideration due to the implications for the control of Type II errors (Keselman et. al, 1998 and Wilcox, 1995).

Kromrey & Hogarty (1998) find sampling distributions of most mean difference indices tested reflect increases in positive skewness when samples are generated from populations with heterogeneous variances. Though the standardized mean difference typically reveals the degree of overlap between distributions of experimental and control group scores, it cannot provide a valid summary if data are not normally distributed. As Winer (1971) describes, parameters provide a description of the population distribution. The frequency distribution determines the number of parameters needed to depict the population. A normal distribution requires two parameters, whereas skewed distributions require more

(Winer, 1971). Only a monotonic transformation of the data can permit an unbiased estimator to adequately estimate the effect (Hedges & Olkin, 1985).

In general, increasing the N, using random sampling, mitigates the effects of a nonnormal distribution (Snedecor & Cochran, 1989). There is a corresponding increase in power, as N increases. (Note: there are circumstances under which increasing the total N does not convey protection against the effects of skewed distributions.) Chang (1993) elaborates this point, "Using the asymptotic theory to obtain power for homogeneity test would give conservative power estimates for data with small samples or non-normal population effects" (p. 59). Moreover, increased sample sizes contribute to improved accuracy, diminishing the effects of random error. In a related manner, Kromrey and Hogarty (1998) find: "All of the indices, as expected, showed a decrease in sampling variation with increasing sample size…bec[oming] more symmetric and more mesokurtic with larger samples" (p. 8).

As the estimates of population effect incorporate effect size indices and these are the data upon which the tests of homogeneity are conducted, understanding the sensitivity of the effect size index is important, when considering the extent to which homogeneity tests remain robust under various conditions. Heterogeneous variance tended to result in increases to the mean and variance of Hedges' g and Cohen's d (Kromrey & Hogarty, 1998). The effect-size index, Hedges' g, upon which the fixed-effects Q test is computed, operates on the assumption that experimental and control group data are normally distributed (Hedges & Olkin, 1985). Kromrey and Hogarty explain that standardized mean difference indices are especially susceptible to heterogeneous variances in that they exhibit increases in positive skewness. Moreover, there is a concomitant relationship whereby increases in population distribution skewness seems to be accompanied by increases in the mean effect size for g and d, as well as increases in variance of these indices. Kromrey and Hogarty (1998) conclude there is no major difference in Type I error control from one effect size estimate to the next. Similarly, Kromrey and Hogarty (1999) find few differences in the power of three commonly used effect size indices, g, d and trimmed-d. Trimmed-d provides the most power applied to data from nonnormal distributions. Hedges'g presents the most power with normal distributions.

With respect to significance tests, few understand the role of sample size (Abelson, 1997). Sample size often determines the extent of variance. Typically, a smaller sample size is associated with greater

variance. Again, the larger the sample size, the greater the power of a test. As already mentioned, large sample theory suggests that large samples mitigate the influence of skewed distributions. For purposes of this study, four issues pertaining to sample size are of present interest: Total size of the experimental and control groups, differences in within-study groups of the sample, the order of the discrepancy, if any, in size between the first and second groups within a study, the size of the first and second groups relative to the size of the variance between these two groups and the ratio of sample size to k.

Large differences in sample size across studies contribute to heterogeneous error variances (Hedges & Olkin, 1985). Hedges & Olkin (1985) describe: "In fact, the nonsystematic variance of estimates of effect is inversely proportional to the sample size of the study on which the estimate is based. Therefore, if studies have different sample sizes, as is usually the case, effect estimates will have different error variances. If the sample sizes of the studies vary over a wide range, so will the error variances" (p. 11). If considerable heterogeneity is present, procedures permitting explanatory analyses should be selected.

Those studying the behavior of the homogeneity tests tend to use equal sample sizes, though this condition rarely occurs in the actual literature. According to Harwell (1992, 1997), equal sample sizes minimize the possibility for inflated Type I error for t tests with unequal variances provided a normal score distribution. In general, increasing the sample size minimizes the influence of random error. However, as mentioned above, skewed distributions can exacerbate unequal variances even under conditions of equal group sample sizes. Harwell (1997) found equal sample sizes between experimental and control groups act as a partial safeguard against Type I error inflation when distributions are normal and variances are unequal. But equal sample sizes fail to provide any safeguard against inflated error rates in the presence of combined unequal variances and skewed distributions.

It would seem that larger variance, regardless of the larger sample size, increases Type I error, as it pertains to chi-square distributed statistics. This pattern does not reflect the tendency of t-tests. As Glass & Hopkins (1984) suggest: "…the true probability of a type I error is always less than the nominal probability when the larger n and larger variance are paired" (p. 238). Therefore, the robustness of the t-test cannot be assumed for all of the tests of homogeneity.

How well the systematic variance is partitioned from the nonsystematic (random) variance – the error – determines the sensitivity of the test and the extent of bias. Between-studies variance is the systematic part of the variance. τ^2 (between-studies variance) is a more comprehensive estimate than between-class effect size differences, in that it represents the variance in the true effect sizes of different implementations of the treatment. Beyond any moderating variables, part of the systematic error is due to aspects of treatment logistics, as well as the time and location of the measurement of the effect of such treatment on a given performance.

The δ_k (between-class effect size differences) can be employed as an explicit test for a moderating variable. The systematic difference in average effect size is broken down to groups of studies, but grouped according to some variable. For example, it is the difference between the average effect size when given to first graders and the average when given to third graders, otherwise referred to as δ. Harwell (1997) suggests that a δ_k of 0 produces results similar to those demonstrated by Hedges and Olkin (1985), using a $\delta_k = .25$. The $\delta_k = 0$ case permits the estimation of Type I error rates. And $\delta_k = 1$ permits the estimation of power. No other specific conclusions are presented with respect to between-class effect size differences for each of the chi-square and permuted tests.

Generally, increasing population effect size results in the increase of sampling variability within the sampling distributions of the effect sizes (those based on mean differences); thereby, increasing positive skewness and leptokurtosis (Kromrey & Hogarty, 1998). In fact, a single study effect size can significantly influence meta-analytic results (Fleiss & Gross, 1991).

Within a given study, if the treatment and control groups have different population variance then the size of the population variance and the sample size of each group have implications for Type I error control. Groups with larger variance and smaller sample sizes represent a negative pairing. Positive pairings consist of smaller variance and smaller sample size. Positive pairings will yield a conservative test and negative pairings will yield a liberal test. In the latter case, Type I error becomes even more inflated than in the presence of equal sample sizes and unequal variances (Harwell, 1997; and Kromrey & Hogarty, 1998). When negative pairings are combined with variance ratios of 4:1 and 8:1, all of these estimated Type I error rates are inflated well beyond α=.05. Large sample sizes do not seem to neutralize this effect.

By contrast, positive pairings with the same variance ratios tend to yield conservative Type I error rates (Harwell, 1997; and Kromrey & Hogarty, 1998). Unequal sample sizes, where the group with the larger population variance is paired with smaller sample sizes than the second, help to maintain better Type I error control.

With respect to the t-test, Glass and Hopkins (1984) explain that it is robust to heterogeneity, as long as the sample sizes from the two groups are equal. In the t-test case, pairing a larger set of n's with a smaller variance results in an underestimation of the true alpha at .05. So we have conservative Type I error, associated with Type II error. But when the sample effects of the group with the smaller sample sizes also come from a population with smaller variance, the Type I error rate quickly increases, as the n ratio gets smaller (the first n is smaller). When sample sizes are equal, there is homogeneity for all practical purposes.

Now that factors relevant for the use of all homogeneity tests have been presented, results from studies examining the specific conditions governing the efficient use of each test are elaborated. It should be noted that not all factors considered for one test have been controlled in the study of every other statistic. Therefore, certain potentially influential factors are not addressed in this section, but are proposed later.

Conditions Affecting the Use of Q

According to Harwell (1997), the performance of Q relies on large sample theory, wherein within-study sample sizes are large enough to support a noncentral t distribution. Based on Harwell's (1997) study, the Q test has conservative Type I error when sample sizes are less than 40, particularly when study sample sizes and K have a ratio less than 1. Sample sizes less than 20 contribute to low power for all but dramatic heterogeneity of effects. Though equal sample sizes between experimental and control groups can often diminish inflated Type I error when there are unequal variances in the primary studies, it has relatively little influence when skewed distributions are involved. Kromrey and Hogarty (1998) find that increasing K, increasing variance heterogeneity and increasing nonnormality most noticeably affect the Q test's sensitivity.

When the group with the larger population variance (whether it is the experimental or control group), considered the first group, has a smaller sample size than the second group, Type I error control is better maintained. As suggested in the discussion of normality, there are conditions under which larger

study samples do not afford any enhancement in robustness (Abelson, 1997; and Kromrey & Hogarty, 1998). When smaller study sample sizes are present, the Q test's sampling distribution no longer parallels the chi-square distribution. As a result of this discrepancy, the Type I error rates substantially depart from nominal α. Moreover, the greatest degree of inflation of average Type I error results from negative pairings of sample size and variance. Conversely, small unequal variance and small unequal sample sizes generate conservative Type I error rates and lower power for all variance ratios (Harwell, 1997; and Kromrey & Hogarty, 1998). The observed effect size has a larger actual magnitude when there is a smaller N, assuming the same p values as a study with a larger N. The Q test has conservative Type I error when sample sizes are less than 40, particularly when sample sizes and the number of studies in the meta-analysis (k) have a ratio less than 1. There is higher power with increasing K, once N is larger than 40 (Harwell, 1997).

In general, as variance heterogeneity increases, Type I error control diminishes for Q (Hedges & Vevea, 1998; and Kromrey & Hogarty, 1998). Large variance ratios yield small power (Harwell, 1997). Harwell (1997) and Kromrey and Hogarty (1998) further conclude that within primary studies, small unequal variance and large unequal sample sizes yield minimal departures from nominal α for smaller variance ratios, but inflated Type I error for variance ratios of 4:1 or 8:1. Note: When there are large variance ratios (e.g. 8:1), and effect sizes are incorrectly pooled in d, estimated power shrinks. This happens because the denominator of d is overestimated, incorrectly reducing the value of d and diminishing power (Harwell, 1997). Hedges and Vevea (1998) find that Q is least robust to $\tau^2 > 0,$ in that Type I error rate exceeded nominal α, particularly when unconditional inferences are in place.

Skewness, when matched with unequal primary study variances, contributed to inflated Type I error regardless of sample size and K (Harwell, 1997). Even given equal sample sizes, skewness will continue to erode Type I error control, particularly when unequal variances are present. Furthermore, inflation increases as skewness and heterogeneity increase. Harwell finds the same outcome for N=200. But Q has conservative error rates for k=30 when there were skewed distributions. In fact, it is likely that any chi-square distributed test will have increased sensitivity under conditions of skewness and kurtosis (Wang, Fan & Willson, 1996).

Generally, when K is small, variance estimate truncation at zero can lead to bias. It is minimized as K increases (Hedges & Vevea, 1998). When N is held constant, increasing K tends to generate more conservative Type I error rates (Harwell, 1997). Specifically, as K increases, even with conditions of homogeneous variance, Q's Type I error control diminishes (Kromrey & Hogarty, 1998). Kromrey and Hogarty notice that using the same conditions, but adding the influence of increasing variance heterogeneity, Type I error control erodes more drastically. Kromrey and Hogarty (1998) point out that for Q, control of Type I error worsened with non-normal distributions and larger values of K. According to Harwell (1997), the primary determinant is larger K because of the increase in noncentral t variates. As δ and N increase, power improves (per Harwell, regarding Q).

When study sample sizes and k have a ratio less than 1, Harwell says Q yields conservative Type I error. Large K with small N contributes to inflated Type I error (per Harwell, 1997). Chang (1993) also concludes that estimated power surpasses theoretical power. With a *large number* of studies where the sample sizes within each study is small (10 or less), large sample theory for effect sizes needs to be modified (Hedges & Olkin, 1985). When you increase k for a fixed N, the problem worsens for Q because there are more noncentral t variates (Harwell, 1997). That is the influence of large sample theory (wherein sample sizes for each group are at least 10 and the magnitude of the effect sizes is no greater than 1.5) may no longer be possible, rendering the estimates of effect sizes less accurate. Once you increase the within-study sample sizes for a fixed k, the d's correspond more closely to a normal distribution with estimated Type I error rates closer to α. Using a weighted linear combination of estimators d_i will permit estimates to approximate the δ^* (the maximum likelihood estimator).

Bias is not necessarily eliminated when a large number of estimators, K, are averaged, as each estimator's bias can be in the same direction. To this purpose, Hedges and Olkin advise applying the unbiased estimator d_i when studies have small sample sizes. With respect to this concern, Chang (1993) concludes though the noncentral chi square distribution (based on asymptotic theory) enhances robustness for reviews with large samples and evenly distributed parameter effects, it results in conservative power for reviews with small samples or nonnormal population effects.

For equal within-study sample sizes but unequal variances, Type I error becomes increasingly inflated with skewed distributions and increasing variance ratios for fixed N and K and more pronounced as

k increased with a fixed N and variance ratio. So the problem (resulting in inflated Type I error) seems to be the unequal variance that further compounds with skewed distributions and increasing K (Harwell, 1997). Estimated power is less than theoretical (nominal) power when N is less than 40, particularly for small N/K ratios. Note: Chang's results indicated that the estimated power was larger than theoretical for the same set of conditions (per Harwell, related to Q). As K increased and N decreased, greater departures arose between the theoretical and simulated power (Chang, pertaining to Q).

Both Harwell (1997) and Kromrey and Hogarty (1998 and 1999) find that Q has inflated Type I error and minimal power under conditions of heterogeneous variances and nonnormal distributions in the primary studies. In a study comparing fixed-effects Q, the random-effects Q, and conditionally-random Q, Hedges & Vevea (1998) conclude that Q most closely approximates the nominal value of α when homogeneity of effects is present. With conditions of heterogeneous effects, the conditionally-random and random-effects Q more closely approximate the nominal value of α. Similarly, Chang (1993) finds that fixed-effects Q is not appropriate for use under conditions of heterogeneous effects.

Although Hedges and Vevea (1998) suggest the application of Q for situations of complete homogeneity of effects, Kromrey and Hogarty (1998) find Q to have diminished Type I error control even under this condition, as K increases from 3 to 30. In fact, only permuted tests maintain Type I error regardless of the presence of heterogeneity of variance and K (provided that K was equal to or greater than 10). Further, Kromrey and Hogarty (1998) find regular $Q_{between}$ tends to maintain better Type I error control over more conditions than the Q test. So for conditions of homogeneous effects and when K is less than 10, the regular $Q_{between}$ is preferable. Therefore, there seems to be some disagreement between the suggestions being made by Hedges and Olkin (1985), Hedges and Vevea (1998) and Kromrey and Hogarty (1998).

Chang (1993) explains the additional variance component included in the random-effects Q test overestimates variance when true population effects are equal. In such a case, the fixed-effects Q should be applied (p. 86). Hedges and Vevea conclude that the random and conditionally-random procedures are most appropriate for between-studies heterogeneity or when the between-studies variance component is greater than zero. But when this variance component equals zero (that is there is no difference among effect size estimates), the fixed-effects Q most optimally controls Type I error.

Based on the aforementioned effect on Q's sensitivity, Kromrey and Hogarty (1998) strongly advise against the use of the Q test of homogeneity when treatment effects are computed from primary studies involving either nonnormal distributions and/or heterogeneous variances. The presence of either or both of these conditions tends to result in inflated Type I error rates. Given these conditions, they recommend employing the permuted $Q_{between}$ test. Additionally, Chang's (1993) study illustrates how large K paired with small N or extreme parameter effects present the greatest challenge to maintaining Type I error control, when applying fixed-effects Q. Moreover, as K increased and N decreased, greater departures arose between the theoretical and simulated power.

When the $Q_{between}$ Test is employed using a permutation strategy, instead of a chi-square distribution, the Type I error control is well-maintained. But this test bears the limitation of insufficiency with small numbers of studies (k=5 or less) included in a meta-analysis, because there are too few permutations of the data to permit a test at an alpha level of .05.

Unequal variances matched with positive pairings of variances and sample sizes generate conservative Type I error rates for the fixed-effects Q (Kromrey & Hogarty). Harwell also finds the proportion of the variances and sample sizes (e.g. small unequal variances to small unequal sample sizes) generates conservative Type I error rates and lower power across all variance ratios. Furthermore, the size of the variance ratio used in conjunction with the positive or negative pairings of sample sizes and variances affects the Type I error rate, as well. Specifically within primary studies, small unequal variance and large unequal samples sizes generated minimal departures from nominal α for smaller variance ratios, but inflated Type I error for larger variance ratios (4:1 and 8:1).

Conditions Indicating the Use of Random-effects Q

As stated previously, Hedges and Vevea (1998) recommend using either the random-effects or conditionally-random procedures when the between-studies variance component (τ^2) is greater than zero, the true population parameter is greater than zero and unconditional inferences are employed. Under these circumstances, the fixed-effects Q, conditionally-random procedure and the random-effects Q all produce inflated Type I error. But the random-effects Q test produces rejection rates much lower than either the other two tests (Hedges & Vevea, 1998). When the between-studies variance is underestimated, the random-effects Q produces less inflated α estimates. Additionally, as both K and between-studies variance

increase, the random Q most closely approximates the nominal α, though still exceeding it (Hedges & Vevea, 1998). Random Q does not produce this result for conditions with small N.

Additionally, researchers such as Raudenbush (1994) tend to advise use of the fixed-effects model for small collections of studies and random-effects for larger collections of studies. It is unclear at what value of K this specific demarcation lies. Raudenbush describes scenarios based on two studies (clearly fixed-effects) or several hundred (clearly random-effects). Beyond those extreme parameters, there is not much more detailed criteria, other than to suggest looking to the magnitude of the treatment and control group differences.

As Hedges and Vevea (1998) illustrate, the primary difference between the fixed- and random-effects estimates of variance is that within the random-effects' procedure, the estimate of τ^2 is added to the variance before dividing it by k. According to Erez, Bloom and Wells (1996), random-effects confidence intervals widen as τ^2 (the between-studies variance) increases. In contrast, the fixed-effects confidence intervals tend to be too narrow, as they do not include between-studies variance. Mulaik, Raju and Harshman (1997) point out that narrow confidence intervals reflect high power and wide intervals suggest low power. Chang's (1993) study seems to confirm these findings, as she explains: "Unlike for the fixed-effects models where simulated power values were sometimes higher than theoretical power values; for random-effects models, a strong two-thirds (9/28) of the discrepancies reflected lower simulated power values" (p. 63).

Specifically, the random-effects approach recognizes multiple sources of variance by including systematic error. Between-study heterogeneity in effects contributes to differences in standard errors from the random- to the fixed-effects model. The random-effect's larger variance component due to the addition of τ^2 increases the standard error of the mean (Hedges & Vevea, 1998).

Heterogeneity of effects and the number of studies included determine how conservative Type I error becomes when employing random- and conditionally random-effects procedures in a conditional inferences situation (Hedges & Vevea, 1998). Small k does not permit more accurate estimation of the variance component which in turn affects the precision of the weights. The estimate of the variance component (τ^2) is used to generate weights used in calculating the mean estimate of the average effect size and its total variance.

When Not to Use the Random-Q Test

Hedges and Vevea (1998) suggest the application of Q for situations of complete homogeneity of effects, as the fixed-effects Q is the only model tested by them that provides rejection rates exactly equal to nominal α (Hedges & Vevea, 1998). It should be noted that Kromrey and Hogarty (1998) recommend the use of regular $Q_{between}$ and permuted alternatives of homogeneity tests. But in general, fixed-effects procedures should be applied to conditions calling for conditional inferences and homogeneity of effects. Applying either random-effects Q or the conditionally-random procedure will undermine power and produce Type II error. But when heterogeneity is present with conditional inferences, these two models (random-effects Q and conditionally-random test) have even lower rejection rates than the fixed-effects Q. In this case, random-effects Q produces even less inflated Type I error than the conditionally-random procedure (Hedges & Vevea, 1998). But because the random-effects Q does not test the inferences underlying the fixed-effects model, it's use under these conditions is not recommended.

Additionally, under heterogeneity, random-effects theoretical power is less than for fixed-effects, as there is a larger variance included in the denominator of the magnitude-of-effects (z) test (Chang, 1993). Relative to the random-effects z test, the fixed-effects z consistently has greater power.

There are some other factors worth mentioning. The random-effects Q test is limited in that the estimate is not precise when there are a small number of studies, due to the limitations of the incorporated τ^2. Power is also curtailed when there is large k with small samples (Chang, 1993). Moreover, when both K and between-study heterogeneity are large, the width of the confidence intervals becomes overextended (Hedges & Vevea, 1998). Such an outcome results in rejection rates being too low for random-effects and conditionally-random procedures.

Conditions Indicating the Use of Conditionally-Random Q

Conditionally-random or random-effects procedures should be employed under unconditional inference situations. Using fixed-effects procedures under these conditions will yield inflated Type I error rates (Hedges & Vevea, 1998). If (τ^2) is greater than zero, the fixed-effects procedure confidence intervals exhibit lower than nominal probability content i.e. the confidence intervals are too narrow. But given the same scenario using random or conditionally-random procedures,

As with the random-effects Q test, the conditionally-random procedure is susceptible to inflated Type I error and low power when conditional inferences are employed. However, the conditionally-random procedure is less sensitive than the random-effects Q. Moreover, the conditionally-random procedure tends to underestimate nominal α, when the true population effect is equal to 0 (perfect homogeneity of effects). In fact, it yields rejection rates progressively more conservative as τ^2 exceeds 0. In other words, the conditionally-random procedure tends to produce Type II error (low power), though to a less extent than the random-effects Q test.

When the true population effect is greater than 0, employing conditional inferences, use of the conditionally-random procedure tends to result in inflated Type I error for all conditions of τ^2 (whether equal to or greater than 0). In fact, so do the fixed- and random-effects Q tests. But the conditionally-random procedure produces less inflation than the fixed-effects Q, and more inflation than the random-effects Q.

When to Use Conditionally-random Q

The conditionally-random procedure produces rejection rates not as extreme as the fixed-effects Q, but more so than the random-effects Q, when between-studies variance becomes greater than zero (Hedges & Vevea, 1998). This pattern is more pronounced when the true population effect is greater than zero and as K increases. On the other hand, when the true population effect is zero (homogeneous effects) and between-studies variance increases, the conditionally-random procedure more closely approximates α and produces less Type II error than the random-effects Q (Hedges & Vevea, 1998).

When Not to Use Conditionally-random Q

As stated previously, Hedges and Vevea (1998) recommend using either the random-effects or conditionally-random procedures when the between-studies variance component (τ^2) is greater than zero, the true population parameter is greater than zero and unconditional inferences are employed. Under these circumstances, the fixed-effects Q, conditionally-random procedure and the random-effects Q all produce inflated Type I error. But the random-effects Q test produces rejection rates much lower than either of the other two tests (Hedges & Vevea, 1998). When the between-studies variance is underestimated, the random-effects Q produces less inflated α estimates. Additionally, as both K increases and between-

studies variance increases, the random Q most closely approximates the nominal α, though still exceeding it (Hedges & Vevea, 1998). Note this does not include a situation with small N.

Conditions Indicating the Use of Permuted Q

This test has not been compared to either the random-effects Q test or conditionally random-effects procedure. In fact, Kromrey and Hogarty (1998) may have produced the only study to simulate and analyze the performance of the permuted version of the $Q_{between}$ test. Therefore, further simulation would enhance the reliability of these results, in and of itself. The other tests under investigation in their study are the Q test, iterated Q and regular $Q_{between}$, as well as 3 permuted indices. All of the permutation-based tests under investigation in the Kromrey and Hogarty (1998) study outperformed the chi-square tests, in general, for each variable under consideration.

In testing an alternative hypothesis and utilizing a different distribution strategy, the permuted $Q_{between}$, maintains better control of Type I and Type II error than the three investigated chi-square tests (Q, iterated Q and regular $Q_{between}$). Permuted Q tests a different null from the other types of homogeneity tests. It also tests a different null from that tested by other fixed-effects tests. Instead, it tests for between-class homogeneity.

Chi-square tests are greatly affected by nonnormality and large sample sizes (Wang, Fan & Willson, 1996). The $Q_{between}$ statistic tests a different null hypothesis, essentially determining the presence of moderating variables. Kromrey and Hogarty explain that increasing K seems to be associated with inflated Type I error rates for both the Q test and iterated Q, whereas the $Q_{between}$ (though still inflated) maintains better Type I error control. Additionally, the permutation strategy further frees the statistic from many of the assumptions typically held by chi-square and normal distributions. As long as K was at least 10, the permutation tests maintain Type I error control across population shapes. They also maintain Type I error control for all conditions with K=10 and 30, regardless of the extent of variance heterogeneity.

Despite the caution that "robust" tests applied to unbalanced designs under conditions of heterogeneity and nonnormality may exhibit liberal Type I error rates (Lix & Keselman, 1998), permuted $Q_{between}$ seems to maintain extraordinary Type I error control. Kromrey and Hogarty (1998) find this procedure to be most robust to the dual violations of normality and homogeneity of variance. When the $Q_{between}$ Test is employed using a permutation strategy, instead of a chi-square distribution, the Type I error

control is well maintained. The permuted $Q_{between}$ derives part of its robustness from the superior properties of the regular $Q_{between}$, as well as its permutation strategy. In order to promote a deeper understanding of the comparative performance of the permuted $Q_{between}$, results of the comparative performance of three chi-square (including regular $Q_{between}$) based tests follows.

Though increasing heterogeneity results in increasingly inflated Type I error rates for each of the chi-square tests (Q, iterated Q and regular $Q_{between}$, to a lesser extent), all permutation strategy-based tests maintain better Type I error control. Regardless of the degree of heterogeneity, permuted $Q_{between}$ maintains Type I error control, as long as the number of studies is 10 or greater. The Type I error rate closely approximates the nominal α level, .05.

Based on Kromrey and Hogarty's (1998) study, the $Q_{between}$ test outperformed both the Q test and iterated Q with proportions of conditions ranging from .75 to .396. Although some inflation of the Type I error rate was evidenced, violations of normality affected the $Q_{between}$ Test less than the homogeneity tests.

The $Q_{between}$ Test's ability to maintain adequate Type I error control is not necessarily enhanced by larger samples in the primary studies. Regardless of sample size, it permitted greater Type I error control than did the Q test or Iterated Q test. Unequal sample sizes, wherein the first group has a smaller sample size than the second group, may have contributed to better Type I error control for the two homogeneity tests, as well as the $Q_{between}$ Test (to a lesser extent).

Nonnormality produces inflated Type I error rates for two versions of the Q test of homogeneity. But the $Q_{between}$ between maintains better Type I error control under the same condition (Kromrey & Hogarty, 1998). This test is less likely to be influenced by sample size than regular or iterated Q tests (Kromrey & Hogarty, 1998).

When to Use the Permuted $Q_{between}$

Where fixed-effects inferences are most appropriate, the number of studies exceeds 9 and the researcher suspects the presence of moderator variables, applying the permuted $Q_{between}$ appears to be the optimal choice. But assuming that the fixed-effects model is appropriate for collections larger than 9, application of permuted $Q_{between}$ appears to be substantiated by the results from the Kromrey and Hogarty (1998) study. In contrast to Q, permuted $Q_{between}$ maintains Type I error control near the nominal level under all investigated conditions of varied skewness and kurtosis. Moreover, use of this test does not

require the determination of homogeneity/heterogeneity, as it tests a different null hypothesis. And thus far, permuted $Q_{between}$ demonstrates robustness to all variations of within-study sample sizes.

When Not to Use the Permuted $Q_{between}$

When the $Q_{between}$ Test is employed using a permutation strategy, instead of a chi-square distribution, the Type I error control is well-maintained. But this test bears the limitation of insufficiency with small numbers of studies (k=5 or less) included in a meta-analysis, because there are too few permutations of the data to permit a test at an alpha level of .05**.** Kromrey and Hogarty (1998) conclude that small K, fewer than 10 studies, generates too few permutations to permit a valid test at the nominal α level .05. They suggest this restriction to be the major limitation of the permuted $Q_{between}$.

Another consideration pertains to the appropriateness of applying the test under conditions suggesting multiple treatment effects. Although it tests for between-class moderating variables, the underlying assumption is more consistent with the fixed-effects model because the only classes included in the test are those in the sample (it is a test of between-class homogeneity). It is not possible to permute studies that are not present. It should be noted that testing this null does not yield any direct variance estimates – only a probability statement about the similarity/difference between average effect sizes, not the difference between two or more parameters. The test assumes population effects are identical. So when the meta-analyst hypothesizes the presence of multiple treatment populations, and wishes to draw inferences beyond the immediate collection of studies, applying some other model, such as random-effects would be more appropriate. If the permuted $Q_{between}$ were applied instead, the test would not be testing the parameters expressed by the model.

Primary Within-study Sample sizes(REQ)

The increase in the variance of population effects or the sample sizes yield increased mean power estimates. Chang (1993) concludes that only Total Sample Size has a significant influence on power. Random samples within primary studies are not considered. She explains there may be the possibility that simulated power underestimates actual power for small samples and large K.

Between-Studies Variance (τ^2)(REQ)

Hedges and Vevea (1998) find that as between-studies variance equals 0, there is inflated Type I error, particularly for random-effects Q. Based on her regression analysis of the factors involved in the

random-effects test of homogeneity, the variation of effects and the total sample size in the model are most responsible for explaining the power of the random-effects test (Chang, 1993, p. 78). Similarly, Raudenbush (1994) notes the precision of the estimates of study effects reflect both the study sample size and the degree of heterogeneity across the true effects.

As the number of studies in the meta-analysis increases, only the random-effects Q and conditionally-random procedure continue to approximate the nominal value. However, substantial heterogeneity of effects combined with increasing numbers of studies can widen the departure from the nominal value. Hedges and Vevea contend the robustness of either the random-effects Q or conditionally-random procedure, although more appropriate with heterogeneous effects, relies on the combined quantities of both heterogeneous effects and increasing numbers of studies.

Between-study heterogeneity in the effects contributes to differences in standard errors from a RE to a FE model (Hedges & Vevea, 1998, p. 494). The RE's larger variance component due to the addition of τ^2 increases the standard error of the mean. But precise estimation of τ^2 is primarily contingent on K (Hedges & Vevea, 1998). If K is small, precision of weighted estimates will be undermined, despite large study sample sizes. However, when K is larger than 20, biases are minimized.

Number of studies in the meta-analysis (k)(REQ)

The power of random-effects Q was most explained by the spread of parameter effects and total sample size (Chang, 1993). Hedges and Vevea (1998) point out if both K and between-study heterogeneity are large, the width of the confidence intervals becomes overextended. Such an outcome results in rejection rates being too low for random-effects and conditionally-random procedures.

Under conditions of large within-study sample sizes, the K plays a crucial role for enhancing the precision of the estimate of the between-studies variance component (Hedges & Vevea, 1998, p. 493). Conversely, Chang (1993, p. 63) finds that simulated power may underestimate actual power for small samples with large k. But according to her regression analysis (p. 78), k had no effect, due to its inclusion in the computation of total sample size. For random-effects Q, Chang (1993) concludes there are no significant associations between the k and N and frequency of significant power discrepancies. In other words, the dependence of power discrepancies on sample sizes N did not vary with k. Chang's study reveals that when k is larger and the average within-study N is smaller, simulated power was higher than

theoretical power for the random-effects Q. In other words, random-effects Q produces inflated Type I error.

Between-Studies Variance (τ^2)(CRQ)

The conditionally-random procedure produces rejection rates not as extreme as the fixed-effects Q, but more so than the random-effects Q, when between-studies variance becomes greater than zero (Hedges & Vevea, 1998). This pattern is more pronounced when the true population effect is greater than zero and as K increases. On the other hand, when the true population effect is zero (homogeneous effects) and between-studies variance increases, the conditionally-random procedure more closely approximates α and produces less Type II error than the random-effects Q (Hedges & Vevea, 1998).

Number of studies in the meta-analysis (k)(CRQ)

As K increases under conditional inferences and a true population effect greater than zero, the conditionally-random procedure produces increasingly inflated Type I error. However, as the between-studies variance increases, it tends to produce progressively less inflated Type I error rates. K does not appear to have an appreciable influence when the true population effect is zero (homogeneous effects) and between-studies variance equals zero. Rejection rates diminish, as between-studies variance increases, producing conservative Type I error. However, when the true population effect is greater than zero and between-studies variance equals zero, rejection rates again become exceedingly greater than α, as K increases.

As K increases under unconditional inferences and a true population effect greater than zero, the conditionally-random procedure produces the same pattern i.e. it produces increasingly inflated Type I error. In this case, as the between-studies variance increases, it produces increasingly more inflated Type I error, as do the other two tests. Again K does not appear to have an appreciable influence when the true population effect is zero and between-studies variance equals zero. Rejection rates increase, as between-studies variance increases, producing inflated Type I error. When the true population effect is greater than zero (heterogeneous effects) and as between-studies variance increases, rejection rates increase for small K, but decrease for larger K.

Suggested Research

Although Chang conducts a thorough power analysis and comparison between the traditional and random-effects Q tests, she does not evaluate these tests' performance in terms of skewness/kurtosis, randomly assigned sample sizes and relative to the performance of the $Q_{between}$ test.

Kromrey and Hogarty (1998) state they do not explore the influence of randomly assigned sample sizes within primary studies. Also, their design does not control for varying degrees of between-studies variance. Although the permuted $Q_{between}$ test is shown to have more robustness under heterogeneous variances than Q, it has not been compared to the random-effects and conditionally random-effects procedures. Comparing it to random-effects and conditionally-random tests can lend greater insight into how all three of these tests respond to violations of normality, random samples and heterogeneity of effects relative to each other.

Hedges and Vevea (1998) do not vary the within-study variances, skewness/kurtosis or primary within-study sample sizes. Therefore, a potentially fruitful investigation involves comparing the random- and fixed-effects Q, conditionally-random procedure and the permuted $Q_{between}$ test.

Chang (1993) suggests investigating the influence of the between-studies variance (τ^2) component on the random-effects test to determine the magnitude of the effects. Both Chang and Harwell recommend further investigation of the random-effects statistic, particularly with respect to the impact of small n and large k on the between-studies variance component, τ^2. Specifically, Chang suggests that the theoretical power estimates presented a poor fit for the simulated random-effects values, calling into question the precision of the τ^2, contingent on K.

Like Hedges (1992), Chang acknowledges that the distribution of effects is not taken into account for random-effects Q. Hedges (1992) concedes that meta-analysts tend to accept that random-effects Q does not require the study effects to be normally distributed. However, if this issue has not been investigated, it would be a fruitful line of inquiry.

Neither Chang nor Hedges and Vevea investigate the combined influence of variance ratio to sample size pairings on the control of Type I error, particularly using random samples. Also, neither of these studies explore the influence of skewness and kurtosis of the within-study samples on the control of

Type I error and power. Further, permuted Q_{bet} has not been compared to either random-effects Q or the conditionally-random procedure.

Hedges and Vevea's study controls for the between-studies' variance heterogeneity and K, but does not account for within-study sample sizes or the influence of violations of normality. Furthermore, the conditionally-random procedure is compared only to fixed- and random-effects Q. Therefore, comparing the Type I error control and power of the conditionally-random procedure to the permuted $Q_{between}$, as well as the fixed- and random-effects Q tests under conditions of primary study nonnormality, unequal variances and unequal sample sizes, is a worthwhile line of investigation.

As mentioned previously, Kromrey and Hogarty (1998) recommend varying the sample sizes within the primary studies and permitting them to vary randomly. In addition, the permuted $Q_{between}$ Test has not been compared to the random-effects or conditionally-random procedures in terms of both Type I error control and power. Again, this test has not been submitted to conditions of varying between-studies variance.

Lastly, in reviewing the internet web sites providing commercial and freeware meta-analytic software, none appear to introduce the incorporation of τ^2 or the use of permuted $Q_{between}$. The commercial and best-funded programs include: www.meta-analysis.com, www.metawinsoft.com and www.weasyma.com. The first two programs provide both fixed- and random-effects options for analyzing categorical or continuous data, presenting transformations of common effect size indices. The last program only presents methods for evaluating categorical data, explicitly stating that methods for use with continuous data are not yet available.

Summary

Because Q, the traditional test of homogeneity, is susceptible to within-study variance heterogeneity and nonnormal score distributions (Chang, 1993; Harwell, 1997; Kromrey & Hogarty, 1998), researchers have begun investigating alternative procedures to accommodate these conditions. Though standardized mean difference effect size estimates such as Hedges' g (used to compute the Q test) remain sensitive to violations of normality and homogeneity, Kromrey & Hogarty (1998) find permuted $Q_{between}$ more robust than the traditional test. Another alternative, the conditionally-random procedure, can be effective with significant between-studies variance (Hedges & Vevea, 1998). Several researchers

recommend the random-effects model of the traditional test, when significant heterogeneity or a distribution of several population effects is suspected (Chang, 1993; National Research Council, 1992; Erez, Bloom & Wells, 1996; and Harwell, 1997). As suggested in Table 1, none of these procedures have been compared under violations of normality and homogeneity within primary studies, unequal and random sample sizes, and heterogeneous effects across schools.

Table 1 – Relevant Factors Examined By Other Studies

	Chang (1993)		**Harwell (1997)**	**Kromrey & Hogarty (1998)**		**Hedges & Vevea (1998)**		
Test	Q	RE	Q	Q	Permuted $Q_{Between}$	Q	CR	RE
Skewness & Kurtosis	normal, non-central chi-square		2/6; 0/25; 0/0; 1/3; 1.5/5	2/6; 0/25; 0/0; 1/3; 1.5/5				
Variances within and across studies	variance of parameter effects		1:1;2:1; 4:1; 8:1	1:1;2:1; 4:1; 8:1		τ^2=0; .33;.67;and 1.0		
K	2,5,10, 30		5,10, & 30	5,10, & 30		1-10, 20, 30, 40, 50, 60, 70, 80, 90, & 100		
N of a single study	20, 60, 120, & 200		equal & unequal within: 10, 20, 40 & 200	equal & unequal within: 10, 20, 40 & 200				
Study effect sizes	many effect sizes		0, .125, .375, .5, .75 & 1.0	0, .125, .375, .5, .75 & 1.0				

Chapter Three

Method

The effectiveness of each of four tests of homogeneity of effects under varying conditions will be evaluated, using computer-simulated data following the design, and analyses to be described. Evaluation of test effectiveness concerns the accurate verification or falsification of homogeneity of effects, as evidenced by the degree to which Type I and Type II errors are controlled, relative to nominal α.

Purpose

The purpose of the study was to investigate data conditions typical in education settings to begin to establish a set of criteria facilitating deliberate model selection for optimal model fit. The responsiveness of four tests of homogeneity of effects was compared under conditions of varying degrees of heterogeneity of variance, primary study sample sizes, number of primary studies and dual violations of normality and homogeneity of effects, as evidenced by statistical power and control of Type I error. Harwell (1997) recommended utilizing random-effects regression in addition to the Q statistic that has been typically generated with fixed-effects regression. Raudenbush (1994) and Bollen (1989) further advised applying a weighted least squares regression, when sample sizes across studies are unbalanced. This second recommendation is supported by Hedges (1982) who found it provided reasonable accuracy for model fit specification when sample sizes were as small as 10. However, scant information is available to indicate whether such a procedure provides adequate robustness for meta-analytic tests.

Comparative analyses are based on each test's relative Type I error rates and power. This study will provide meta-analysts with more specific guidelines for the use of each test by addressing the following questions:

1) To what extent is the Type I error rate of the fixed-effects Q, permuted Q_{bet}, random-effects Q and conditionally-random Q maintained near the nominal alpha level across variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness and kurtosis?

2) What is the relative statistical power of the fixed-effects Q, permuted Q_{bet}, random-effects Q and conditionally-random Q across variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness and kurtosis?

Design

This study is modeled after Harwell (1997) and Kromrey and Hogarty's (1998) experimental design. Specifically, it entails a 2 x 3 x 3 x 3 x 3 x 3 x 2 factorial design. The study also controls for between-studies variance, as suggested by Hedges and Vevea's (1998) study. The randomized factorial design includes seven independent variables: (1) number of studies within the meta-analysis (10 and 30); (2) primary study sample size (10, 40, 200); (3) score distribution skewness and kurtosis (0/0; 1/3; 2/6;); (4) equal or random (around typical sample sizes, 1:1; 4:6; and 6:4) within-group sample sizes; (5) equal or unequal group variances (1:1; 2:1; and 4:1;); (6) between-studies variance, τ^2 (0, .33, and 1); and (7) between-class effect size differences, δ_k (0 and .8). Data were obtained using two programs: one for null hypotheses (972 simulations) and the other for non-null hypotheses (486 simulations). Hence, the study incorporates 1,458 experimental conditions, illustrated in Figure 1. Simulated data from each sample are analyzed using each of four tests of homogeneity.

The dependent variable is, in part, the proportion of conditions with adequate Type I error control at the nominal alpha level of .05. Additionally, estimates of statistical power are computed for those conditions where tests maintained adequate Type I error control. These power estimates indicate the degree to which a test reflects sensitivity to significant heterogeneity of effects, in the presence of violated assumptions.

Harwell (1997) applies a criterion for determining inflated versus conservative Type I error rates. Specifically, the criterion includes the number of rejections above .056 are termed "inflated", whereas those empirical values below .044 are termed "conservative". The dependent variable is the proportion of meta-analyses leading to a rejection of the null hypothesis. This represents either the Type I error rate or power depending on the truth of the null hypothesis. Essentially, the effectiveness of each of the four tests of

homogeneity of effects is being evaluated based on this performance, as well as how large a discrepancy exists between the simulated data and nominal α values.

Table 2 – Study Design (shaded cells = Power estimates; white cells = Type I error rates)

Primary Within-study Sample Size	*Equal or Random Value of Within-study Sample Size*	*Skewness & Kurtosis*	*Variance Within Studies*	*Btwn-class Effect Size Dfrncs*	***For Each of the Four Tests***					
					τ^2=0		τ^2=.33		τ^2=1.0	
					κ=10	κ=30	κ=10	κ=30	κ=10	κ=30
10	**1:1**	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		1/3	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		2/6	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
	4:6	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						

Note: The number of cells will be quadrupled as the data for each of the 4 tests of homogeneity are entered.

Table 2 (continued) – Study Design (shaded cells = Power estimates; white cells = Type I error rates)

Primary Within-study Sample Size	*Equal or Random Value of Within-study Sample Size*	*Skewness & Kurtosis*	*Variance Within Studies*	*Btwn-class Effect Size Dfrncs*	***For Each of the Four Tests***					
					τ^2=0		τ^2=.33		τ^2=1.0	
					κ=10	κ=30	κ=10	κ=30	κ=10	κ=30
10	**4:6**	**1/3**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		2/6	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
	6:4	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		1/3	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						

Note: The number of cells will be quadrupled as the data for each of the 4 tests of homogeneity are entered.

Table 2 (continued)– Study Design (shaded cells = Power estimates; white cells = Type I error rates)

Primary Within-study Sample Size	*Equal or Random Value of Within-study Sample Size*	*Skewness & Kurtosis*	*Variance Within Studies*	*Btwn-class Effect Size Dfrncs*	***For Each of the Four Tests***					
					τ^2=0		τ^2=.33		τ^2=1.0	
					κ=10	κ=30	κ=10	κ=30	κ=10	κ=30
10	**6:4**	**2/6**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
40	**1:1**	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		1/3	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		2/6	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						

Note: The number of cells will be quadrupled as the data for each of the 4 tests of homogeneity are entered.

Table 2 (continued)– Study Design (shaded cells = Power estimates; white cells = Type I error rates)

Primary Within-study Sample Size	*Equal or Random Value of Within-study Sample Size*	*Skewness & Kurtosis*	*Variance Within Studies*	*Btwn-class Effect Size Dfrncs*	***For Each of the Four Tests***					
					τ^2=0		τ^2=.33		τ^2=1.0	
					κ=10	κ=30	κ=10	κ=30	κ=10	κ=30
40	**4:6**	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		1/3	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		2/6	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
40	**6:4**	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						

Note: The number of cells will be quadrupled as the data for each of the 4 tests of homogeneity are entered.

Table 2 (continued)– Study Design (shaded cells = Power estimates; white cells = Type I error rates)

Primary Within-study Sample Size	*Equal or Random Value of Within-study Sample Size*	*Skewness & Kurtosis*	*Variance Within Studies*	*Btwn-class Effect Size Dfrncs*	***For Each of the Four Tests***					
					τ^2=0		τ^2=.33		τ^2=1.0	
					κ=10	κ=30	κ=10	κ=30	κ=10	κ=30
40	**6:4**	**1/3**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		2/6	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
200	**1:1**	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		1/3	**1:1**	**0**						
				0.8						

Note: The number of cells will be quadrupled as the data for each of the 4 tests of homogeneity are entered.

Table 2 (continued) – Study Design (shaded cells = Power estimates; white cells = Type I error rates)

Primary Within-study Sample Size	*Equal or Random Value of Within-study Sample Size*	*Skewness & Kurtosis*	*Variance Within Studies*	*Btwn-class Effect Size Dfrncs*	***For Each of the Four Tests***					
					τ^2=0		τ^2=.33		τ^2=1.0	
					κ=10	κ=30	κ=10	κ=30	κ=10	κ=30
200	**1:1**	**1/3**	**2:1**	**0**						
				0.8						
			4:1	**0**						
				0.8						
		2/6	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
	4:6	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		1/3	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		2/6	**1:1**	**0**						
				0.8						

Note: The number of cells will be quadrupled as the data for each of the 4 tests of homogeneity are entered.

Table 2 (continued) – Study Design (shaded cells = Power estimates; white cells = Type I error rates)

Primary Within-study Sample Size	*Equal or Random Value of Within-study Sample Size*	*Skewness & Kurtosis*	*Variance Within Studies*	*Btwn-class Effect Size Dfrncs*	***For Each of the Four Tests***					
					τ^2=0		τ^2=.33		τ^2=1.0	
					κ=10	κ=30	κ=10	κ=30	κ=10	κ=30
200	**4:6**	**2/6**	**2:1**	**0**						
				0.8						
			4:1	**0**						
				0.8						
	6:4	**0/0**	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		1/3	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						
		2/6	**1:1**	**0**						
				0.8						
			2:1	**0**						
				0.8						
			4:1	**0**						
				0.8						

Note: The number of cells will be quadrupled as the data for each of the 4 tests of homogeneity are entered.

Seven experimental variables are being investigated: (1) number of primary studies included in the meta-analysis, (2) total sample size, (3) primary study distribution shape, (4) equal, random or unequal within-group sample sizes, (5) experimental to control group variances, (6) extent of heterogeneity of effects and (7) between-studies effect size differences. Three variables (1), (6) and (7) focus on aspects of the meta-analysis, whereas, (2) – (5) address features of the primary studies included in the meta-analysis.

In order to determine the extent to which the experimental conditions are representative of published meta-analyses, a survey was conducted of all articles published in the *Review of Educational Research* from 1995 to 2000. Out of a total of 15 meta-analyses, three employed random-effects tests either *a priori* or *post hoc*. The meta-analyses included between 13 – 180 studies. Because some of these syntheses included studies in which several effect sizes were computed to address multiple hypotheses, as many as 1,728 effect sizes were computed for a single meta-analysis. Six syntheses included 30 or fewer studies. Eight syntheses included 50 or fewer studies. Eleven syntheses included 60 or fewer studies. It should be noted that two of the 15 studies either did not conduct homogeneity tests or compute an average effect. These two meta-analyses were eliminated for this reason.

Sample size data were incomplete for six of the syntheses. Due to differences in design, some studies presented total primary study sample sizes, whereas others provided the sample sizes for the experimental and control groups. Total sample sizes ranged widely from 12 – 3,656. There were multiple sample sizes for syntheses addressing multiple questions.

Ten of the meta-analyses resulted in overall heterogeneity, requiring further evaluation of moderating variables. Three syntheses presented unclear information about the result. One study presented a homogeneous effect after the removal of outlying effect sizes. Only two studies presented a definitively homogeneous main effect. χ^2 and average effects were computed for most of the meta-analyses. Eight of the remaining 13 meta-analyses generated an average effect. These were computed for heterogeneous and homogeneous outcomes alike. The average effects (computed for 8 of the 13 studies) ranged from .34 to .79. The χ^2 s (computed for 4 of the 13 studies) ranged from 82.32 to 3,246.99. The levels selected for the variables being investigated in this study correspond with those presented in the extant literature.

Given that the between-studies variance component represents a significant difference in the delineation of uncertainty between the fixed-effects and random-effects models, further consideration of the

influence of varying τ^2 seems warranted. Based on previously applied τ^2 levels and a conversation with L.V. Hedges (personal communication, September 3, 1999) about the determination of typical τ^2 values, the decision was made to incorporate 3 specific values of τ^2 (0, .33 and 1.0).

After analyzing raw data from a Title I reading program administered throughout a public school district in Florida, it was apparent that reading scores were nonnormally distributed for this population. Scores were negatively skewed and leptokurtotic. Several researchers have addressed the concern of the influence of nonnormal population shape on the performance of homogeneity tests (Hedges & Olkin, 1985; Raudenbush & Bryk, 1987; Wilcox, 1995; Lix & Keselman, 1998; and Kromrey & Hogarty, 1998). Based on prior evidence in the literature and from the school district, 3 representative population shapes were selected [0/0 (normal skewness and kurtosis), 1/3 and 2/6 (extreme skewness and kurtosis)].

Sample

The data were generated through a Monte Carlo study. A Monte Carlo study uses computer models to simulate statistics' performance under various conditions. Snedecor and Cochran (1989) state: "An important use of tables of random digits and of computers is to draw repeated random samples of a given size from a population. By estimating a desired population characteristic from each sample we obtain the 'sampling distribution' of the estimates" (p. 15).

The effectiveness of the statistic is determined by the extent to which Type I and Type II error rates are controlled, relative to the theoretical or criterion alpha level. Data were generated using the SAS procedure Interactive Matrix Language. Hedges and Olkin (1985) refer to the SAS proc. Matrix: "A simpler alternative to the computation of estimates and test statistics is to use a program (such as SPSS or SAS proc GLM) that can perform WLS analyses" (p. 173). But for this study, the analyses were conducted, initially, using SAS/IML version 6.12. It performs the operations more rapidly than the GLM procedure. But the accuracy of the data analysis was verified by comparing the results to the GLM procedure.

In order to extend the studies conducted by Harwell (1997) and Kromrey and Hogarty (1998), this study utilized the same procedures for random number generation and transformation of nonnormal scores. Harwell used an unspecified random-number generator from the 1986 Numerical Recipes to simulate standard-normal deviates. He then applied the Fleishman's method for transformation of the same.

Following the procedure used by Kromrey and Hogarty (1998), the RANNOR random number generator in SAS were used to generate normally distributed random variables. Different seed values were used in each execution of the program to yield the random numbers. Nonnormal distributions were replicated by transforming normal random variates derived from RANNOR based on the Fleishman (1978) technique, referred to as the Power Method.

In conducting a Monte Carlo study, the study design focuses on establishing a criterion for determining the robustness of the tests, as well as the number of iterations necessary to ensure the reliability of the results. Based on Robey and Barcikowski's (1992) seminal work, the number of iterations necessary for a two-tailed test of departures of π, the estimate of the actual Type I error rate (computed as the observed proportion of the total number (n) of calculated test statistics greater than a critical test value under the null hypothesis), from α, the nominal Type I error rate, was determined. Consistent with Harwell's nominal alpha of .05, a standard power of .8 and Bradley's (1978) criterion for intermediate stringency of α +/- ¼ α, the number of recommended simulated meta-analyses would be 2,660 (see Robey & Barcikowski, 1992, p. 286). Though it may not affect appreciable differences in the consistency of findings, this study will employ the same number of iterations (5000) as found in Harwell's study.

In an effort to describe the realistic simulation of an educational meta-analysis, study characteristics are presented in the form of a linear model. In thinking about the model, one might think of the simulated meta-analysis as consisting of a collection of studies about a given reading program being administered to groups of first and third grade readers. The reading outcomes being investigated are differentially effective depending upon the student's grade, the moderating variable. The individual student scores will be generated through the implementation of the following model,

$$X_{ijk} = \mu + \alpha_{jk} + \varepsilon_{ijk}$$

Where X_{ijk} = the observed "score" for child i, in group j in study k,
μ = the grand mean of all scores for all children in both groups in all of the studies,
α_{jk} = the "effect" of treatment j being implemented in study k, and
ε_{ijk} = random error associated with this child's score.

Each simulation will represent an individual student's score, designated as part of either the control or treatment group. This latter feature will be expressed as an independent variable embedded within the α_{jk}. At the meta-analytic level, the characterization of "samples of studies" as either random- or fixed-effects

will arise from the value of τ^2 as either equal to 0 (fixed-effects) or greater than 0 (random-effects). τ^2 will be expressed in the α. The extent of heterogeneity of the "sample" will be manipulated during the simulation by varying the error term, ε. In a similar manner, the skewness and kurtosis of the "sample effects" will be controlled. K will be varied based on the number of iterations.

When considering a heterogeneous outcome or non-null condition, two additional components will need to be added. If heterogeneity is simulated with the influence of a systematic moderator, studies will be grouped according to the student's participation in either a 1st or 3rd grade reading program. The linear model can then be characterized as

$$X_{ijkm} = \mu + \alpha_{jkm} + \beta_m + \alpha\beta_{jkm} + \varepsilon_{ijk}$$

Where β_m represents a main effect for overall differences in 1st and 3rd grade reading students, and $\alpha\beta_{jkm}$ expresses the differential effectiveness of the reading program for use with 1st and 3rd grade students. The magnitude of this last component establishes the degree of falsity of the meta-analytic null hypothesis of homogeneity of effects.

The type of score distribution and group variance ratios are patterned after the Harwell (1997) and Kromrey and Hogarty (1998) studies. Kromrey and Hogarty (1998) recommended the inclusion of random within-group sample sizes, as included in the present study. Hedges and Vevea (1998) incorporated the between-study variances included in the present study. As between-study variance has been shown to significantly influence the resulting Q statistic (independent of primary study variance), it too has been included.

Test Statistics Examined

For each simulated meta-analysis, the significance probability of heterogeneous effects will be analyzed by employing each of four tests of homogeneity of effects. Tests of homogeneity of effects reflect ("whether sampling error alone accounts for this variation or whether features of studies, samples, treatment designs, or outcome measures also contribute to variation" and "indicates that more variance exists in effect size estimates across studies than predicted by sampling error alone" (Cooper & Dorr, 1995, p. 489). Further, "Homogeneity analysis compares the amount of variance exhibited by a set of effect sizes with the amount of variance expected if only sampling error is operating" (Cooper, Nye, Charlton, Lindsay & Greathouse, 1996, p. 251). Homogeneity tests, as a whole, apply either regression or

correlation coefficients to determine the average effect. In this study, regression coefficients are employed. The rationale for applying regression versus correlation is that regression coefficients provide a more apt measure of magnitude than do correlation coefficients (Abelson, 1997). Slope has dimensional units and r is susceptible to range restriction, whereas b is not.

For both traditional Q and the random-effects procedure, first apply the regression equation

$$d_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + u_i + e_i$$

Where d_i = the effect size estimate for the *i*th *study*,
β_o = the grand mean effect size
β_1 = the expected mean difference in effect sizes between studies of different classes
X_{i1} = some amount of the first study characteristic in the *i*th *study*
u_i = the residual or component of the score effect size not explained by X, and
e_i = the error of estimation

For each simulated meta-analysis, consisting of *k* studies, the Q test of homogeneity will be calculated. This statistic is derived by

$$Q = \Sigma (d_i - d_+)^2/\sigma^2(d_i)$$

d_i = minimum variance estimate of the sample effect; sample effect size for the *i*th study
d_+ = average weighted d_i; weighted average affect size across the *k* studies and

$$d_+ = \sum_{i=1}^{k} \frac{1}{\sigma^2}(d_i)/\Sigma\ 1/\sigma^2$$

The element v_i used to weight the reliability of each study is the variance of the effect size in the *i*th study, obtained as

$$\sigma^2 = n^T + n^C/\ n_i^T\ n_i^C + d_i^2/\ 2(n_i^T + n_i^C)$$

where n_i^T and n_i^C are the sample sizes for the treatment and control groups in the *i*th study.

The obtained test statistic, Q, is evaluated for statistical significance by comparing its magnitude to a critical value of chi-square with k-1 degrees of freedom. If the obtained Q exceeds the critical chi-square value, the null hypothesis of homogeneity of effects is rejected.

As mentioned previously, the random-effects model interprets the field of studies as part of a wider and unknown universe. The random-effects procedure bears one unique element, differentiating it both theoretically and algorithmically. This element is variously referred to as the "between-studies variance component", "estimator of population variance" "estimator of the variance of population effects" and "estimator of the population variance component". This element is added to the sampling error to compute the estimate of the Total variance of the average effect. Typically, this statistic has increased

variance due to the added uncertainty built into its algorithm. Often, differences between fixed- and random-effects standard errors are due to 'substantial between-study heterogeneity' in the effects (Hedges & Vevea, 1998, p. 494). Typically, the standard error is substantially larger than in the fixed-effects test. Its algorithm includes the estimate of population variance, the component accounting for the added uncertainty. The distinction of between-studies variance translates into added degrees of freedom in the denominator incorporated into the individual study variances, contributing to the larger standard error of the mean. Larger standard error results in wide confidence intervals and low power.

$$Q_+ = \Sigma (d_i - d_+)^2/\sigma^2(d_i|\delta_i)$$

Note: The primary difference between the traditional Q and Q_+ is the inclusion of the variance of sample effect sizes while holding constant the population effect size(s), δ_i. This modification creates the variance of the conditional distribution of d_i given δ_i.

Hedges and Olkin (1985) explain the algorithm applies expected values of the mean squares. These values are represented in terms of variance components. Sample values are substituted for these expected values and used to solve for the variance components. This procedure results in the unbiased estimates of the variance components. Specifically, if the design were unbalanced, compute the weights for each study effect accordingly (see Raudenbush, 1992):

for random-effects weights $w_i = 1/(v_i + \sigma_\theta^2)$, where v_i is defined as the conditional variance or the square of the standard error for a given study.

Once Q is calculated for fixed-effects, this statistic can be used to calculate c which then permits the computation of τ^2 for the random-effects procedure.

$$C = \Sigma w_i - \frac{\Sigma(w_i)^2}{\Sigma w_i}$$

Next, test the significance of the effect-size variance component (τ^2) or that H_o: $\tau^2 = 0$

$$\tau^2 = Q - (k - 1)/c$$

If this value is larger than zero then one can no longer assume the presence of a single effect. One then recomputes the random-effects weights using the estimate of τ^2.

d^+ for random-effects = random-effects weighted mean effect size or $\frac{\Sigma w_i d_i}{\Sigma w_i}$

Using d^+ for random-effects, one constructs the confidence interval around μ.

The conditionally-random effects Q test is a procedure for maintaining a degree of flexibility in the process of testing for homogeneity of effects. Initially, the traditional fixed-effects Q is used to test for the homogeneity of the effects. If the null hypothesis of homogeneity is maintained, the decision is made to combine the study effects and compute an average or common effect. But if the synthesist rejects the null, further testing for moderating variables is conducted using the random-effects Q and corresponding weighted least squares procedure.

The Permuted $Q_{Between}$ test is a randomization test designed to generate a more extensive sample from a limited number of studies, by randomly reassigning studies to each of two classes. It is a modification of the fixed-effects test of homogeneity first investigated by Hedges and Olkin (1985). Hedges and Olkin (1985) refer to the $Q_{between}$ test as the between class goodness-of-fit statistic Q_B.

The Permuted $Q_{Between}$ tests a different sort of hypothesis than the other homogeneity tests. It tests the null hypothesis that the average effect size is the same across classes. Specifically, it partitions the observed effect sizes into groups according to the hypothesized moderator variable and tests the tenability of the null hypothesis that population effects are the same in the subgroups. It is a simpler hypothesis in that it directly tests for moderating variables. However, this test maintains the fixed-effects assumption that all of the observed variation is sampling error. Moreover, the sample effect sizes originate from the same population. The variable effects have been observed. The test statistic is based on the total weighted sum of squares**.** The denominator is the normalized weighted sum of squares of the effect size indices about the grand mean d++. Having a common effect size across studies indicates the $Q_{between}$ possesses an approximate chi-square distribution with k-1 degrees of freedom.

According to Noreen (1989), a randomization test is a "procedure for assessing the significance of a test statistic [and] involves randomizing the ordering of one variable relative to another" (p. 12). Orderings are permutations of the variables relative to each other. Randomization tests are nonparametric and based on an empirical, rather than theoretical distribution. "Non-parametric" refers to the manner in which the nature of the population distribution is not specified explicitly. Almost all permutations are non-parametric and vice versa. They do not require random sampling, but are based on random assignment within the study.

Permuting this test involves the reordering of study effects by dividing the factorial of K, the number of study effects, by the factorial of R, the number of studies in the smaller group (either treatment or control), which is multiplied by K minus R. For example, if the number of study effects equals 10 and the number of studies in the smaller group equals 4, the number of combinations of K (now referred to as N) taken R at a time will result in the number of permutations to be conducted. In this case, 210 permutations would be simulated, by computing,

N!/ (R!(N-R)!) or specifically 10!/ (4!(10-4)!) = 3,628,800/24(720)=210

Once the number of permutations has been determined, one can then generate the data to be tested for homogeneity of effects. The algorithm for the $Q_{between}$ test follows:

$$Q_{bet}=\Sigma_i (d_i^{+} - d^{++})^2/\sigma^2(d_i^{+})=\Sigma_i\Sigma_j (d_i^{+}-d^{++})^2/\sigma^2(d_{ij}) \qquad \mathbf{Q_{between}\ test}$$

With the summation over *I* classes, and *j* studies in each class.
Where d_i+ refers to average weighted effect size for class *I*,
d^{++} represents the grand mean effect size, and
d_{ij} represents the effect size for the *j*th study in the *i*th class (Hedges & Olkin, 1985).

Kromrey and Hogarty (1998) found this procedure to be more robust to the dual violations of normality and homogeneity of variance. When the $Q_{between}$ Test is employed using a permutation strategy, instead of a chi-square distribution, the Type I error control is well maintained. Because it does not involve a normal or chi-square distribution, the Permuted $Q_{Between}$ is freed from many of the assumptions, such as normality, upon which other homogeneity tests are based. The primary limitation noted by their study is its inability to generate a sufficient data set with which to test at the .05 alpha level under conditions of small K (generally 5 or less).

Data Analysis

Each of the 1,458 experimental conditions extrapolated from the seven independent variables is to be analyzed using each of the four tests of homogeneity. Effectiveness of each test is to be evaluated based on the proportion of the 5000 simulations of each meta-analytic condition reflecting adequate Type I error control at the nominal alpha level of .05. As defined previously, rejections above .056 are termed "inflated", whereas those empirical values below .044 are termed "conservative". In addition, the estimated Type I error rates will be reported. For non-null conditions, power estimates will be calculated. The rejection rates of all variable combinations will be presented by inserting these in the design matrix, such as the one appearing in Figure 1 of this chapter. Furthermore, the marginal rejection rates for each

factor of the study, indicating comparative Type I error control, will be presented in the form of graphs. Both the proportion of conditions with adequate Type I error control and the average Type I error rate estimates will be presented graphically, categorized by each of the seven independent variables. Finally, box and whisker plots will be used to illustrate the distribution of the Type I error rates of each of the four tests. The organizing variable will be K, the number of studies included in the meta-analysis.

Chapter Four

Results

The purpose of this study was to investigate the power and Type I error control of the permuted Q, random-effects test, fixed-effects test and regular Q test under varying levels of heterogeneity of effects (τ^2=0, τ^2= .33, and τ^2=1) and at α level .05, as well as three variance ratios, two different numbers of studies (hereafter referred to as K) and 3 levels of sample sizes within studies (hereafter referred to as N=10, 40 and 200). Test performance was based on a set of criteria established by Bradley (1978) wherein the robustness of the test depends upon the range of p results falling around a preset α. The relative efficacy of these three tests (and one conditionally-random procedure) were compared between the K=10 and K=30 conditions and across variance ratios between control and experimental groups, increasing skewness/kurtosis and increasing sample sizes within meta-analytic studies. The comparison of the relative performance of these three tests and the conditionally-random procedure within each set of controlled conditions should enhance the appropriateness of practitioners' test selection for meta-analysis.

In particular, the research questions addressed were the following:

1. To what extent is the Type I error rate of the fixed-effects test (FE), permuted Q_{bet}, random-effects test (RE) and conditionally-random procedure (CR) maintained near the nominal alpha level across variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness and kurtosis?
2. What is the relative statistical power of the fixed-effects test (FE), permuted Q_{bet}, random-effects test (RE) and conditionally-random procedure (CR) given variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness and kurtosis?

Results of each of the 1,458 experimental conditions arising from the factoring of seven independent variables across each of the three tests of homogeneity and conditionally-random procedure are presented.

Effectiveness of each test was evaluated based on the proportion of the 5000 simulations of each meta-analytic condition reflecting adequate Type I error control at the nominal alpha level of .05. As defined by Bradley (1978), rejections above .055 are termed "inflated", whereas those empirical values below .045 are termed "conservative" for nominal α=.05. For nominal alpha level .10, rejections above .11 are "inflated", while those rejections below .09 are conservative.

As an overview, the box and whisker plots for each primary condition will be presented. All conditions will be incorporated while isolating the one variable of interest. These plots will be presented by the following 5 controlled variables: 2 plots for K=10 and K=30; 3 plots for (τ^2=0, τ^2=.33, and τ^2=1); 3 plots by (N =10, N=40 and N=200), 3 plots for (skewness/kurtosis = 1/1, 1/3 and 2/6); and 3 plots for each of the variance ratios of 1:1, 2:1 and 4:1, respectively. The estimated Type I error rates are reported. For non-null conditions, power estimates have been calculated. The rejection rates of all variable combinations are presented within the design matrices, appearing in Tables 1 through 6 of this chapter. Tables illustrating the proportion of conditions with adequate Type I error control and the average Type I error rate estimates follow. Lastly, box and whisker plots used to display the distribution of the Type I error rates of each of the three tests and conditionally-random procedure are presented.

Control of Type I Error Rate

For purposes of this study, the control of Type I error is being examined using box and whisker plots, proportion of simulations with adequate Type I error control, average Type I error rates for each condition and marginal error rates for individually simulated conditions. Type I error control must first be determined before further examination of power becomes relevant for any of the conditions with a δ or effect greater than 0. Type I error occurs in the meta-analytic case when the researcher has deemed that a differential treatment effect (a moderating effect) has occurred for separately defined groups when in fact no true difference in effect exists.

Box Plots for K

In both the box plots for K=10 and K=30 (Figures 1 and 3), the permuted Q between, random-effects test (RE) and the conditionally random procedure (CR) demonstrated markedly greater concentrations of conditions in which Type I error was better maintained than in either the regular Q test or the fixed-effects (FE) test. For this reason, only the box plots for the three better performing tests were magnified so that a closer comparison can be made between these. As the distribution covers a much broader range, magnification of the range is less necessary.

Figure 2 reveals the permuted Q between maintained Type I error control to the highest degree of the three best performing tests. For K=10, the random-effects test (RE) and the conditionally random (CR) procedure performed similarly, showing a larger spread in the distribution of Type I error rate estimates.

Comparing the K=10 (see Figure 1) to the K=30 (see Figure 3) condition, the regular Q test produced a greater number of conditions with lower, though still inflated, Type I error when K=10. At K=10, the median error rate for the regular Q test fell at .75, whereas at K=30 it was 1. Surprisingly, the FE test maintained a similar frequency of conditions with lower, though still inflated, Type I error, regardless of the K.

The RE and CR tests performed similarly at both K=10 (see Figure 2) and 30 (see Figure 4). At K=10, each had slightly inflated medians with inflated Type I error at .10. At K=30, the median for each decreased to approximately .065. As will be demonstrated later, this pattern of performance was borne out by the other analyses.

Of all of the tests, only permuted Q maintained Type I error control consistently across conditions of K. Although, deviations from nominal alpha were present, at least 50 % of the conditions fell within Bradley's criterion of acceptability. With only a few exceptions did any of the permuted Q's conditions exceed .06 (see Figure 4) at K=30.

Figure 1.
Distributions of Type I Error Rate Estimates Across Experimental Conditions for K=10

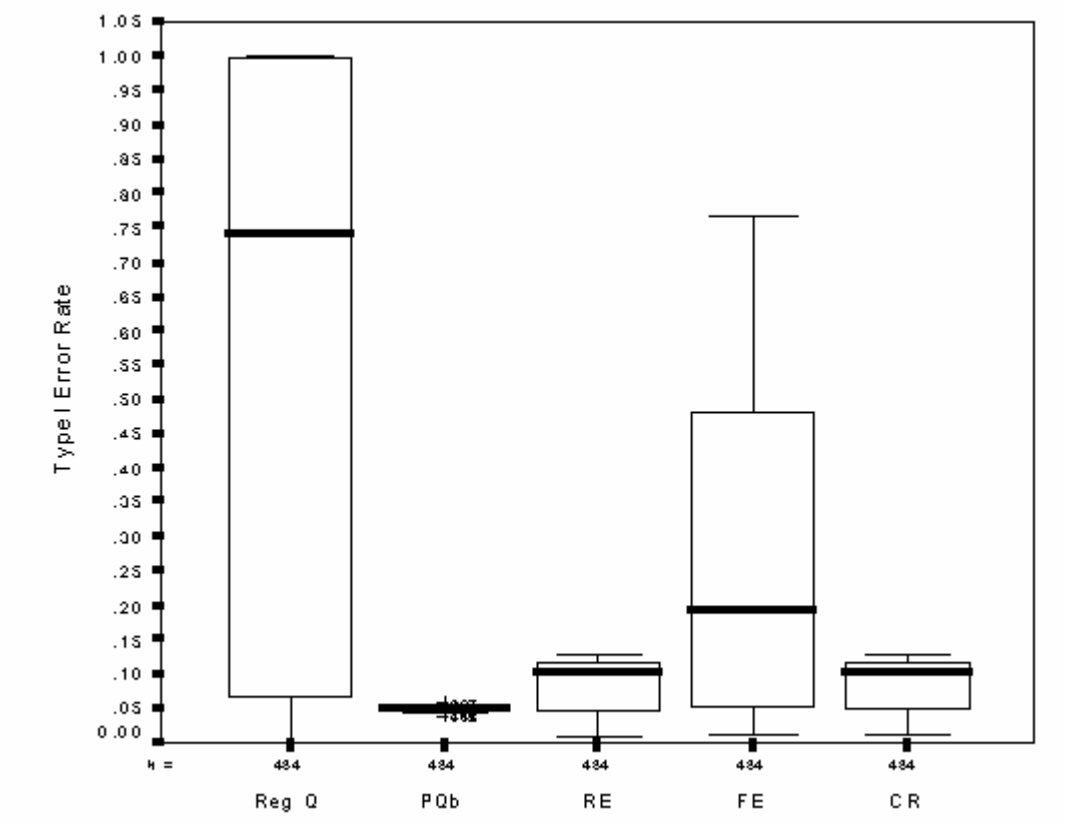


Figure 2.
Magnified Distributions of Type I Error Rate Estimates Across Experimental Conditions for K=10

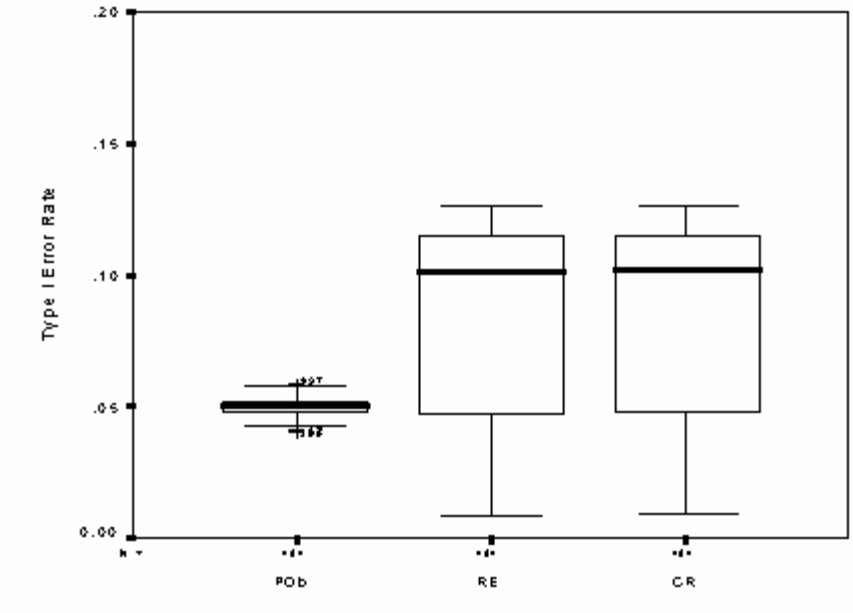


Figure 3.
Distributions of Type I Error Rate Estimates Across Experimental Conditions for K=30

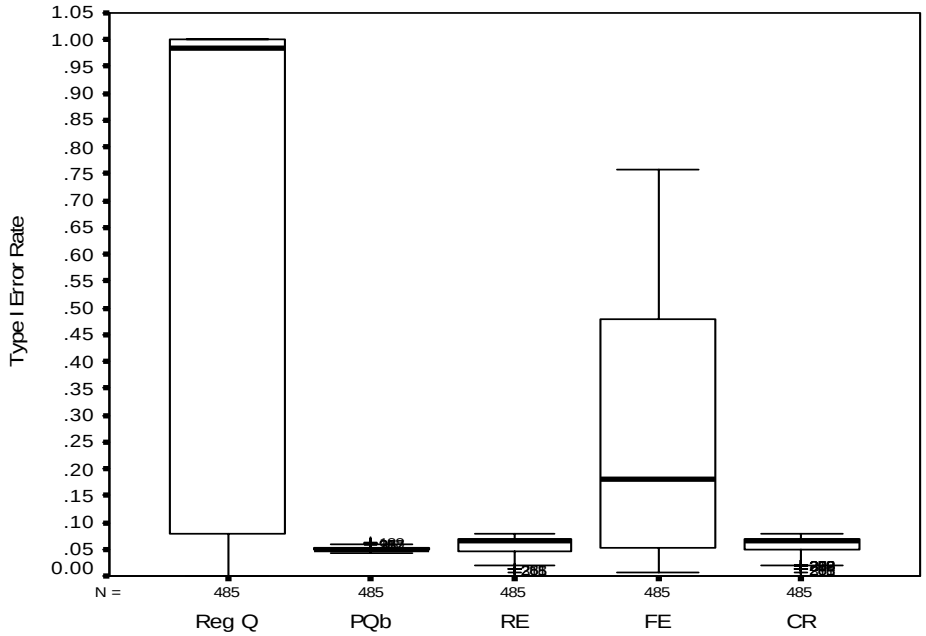
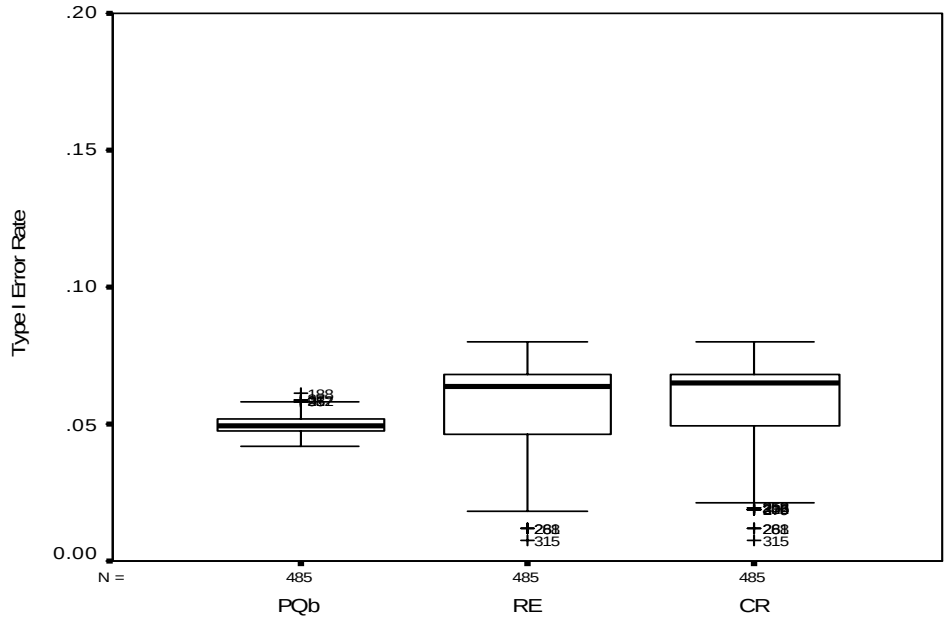


Figure 4
Magnified Distributions of Type I Error Rate Estimates Across Experimental Conditions for K=30



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Box Plots for τ^2

When τ^2 =0, the between-studies variance is essentially not a consideration. Under τ^2 =0 (see Figure 5), each of the tests tended to maintain Type I error rates around nominal α=.05. The RE, FE, CR and regular Q tests produced some conservative rates and some moderately liberal rates. As the reader will recall, the model underlying the regular Q test requires the test to be most sensitive to within-study variance. Therefore, reported outliers are to be expected.

Increasing τ^2 introduces between-studies variance. Because Q was constructed to be sensitive to heterogeneity across studies, Q cannot simultaneously maintain robustness to Type I errors to the same extent that other tests do. But Q's particular sensitivity to variance permits it to detect any heterogeneity present. The reader will discover how increasing τ^2 permitted this test to display increased power while other tests (including the most powerful one) lost power.

Similarly, the FE test did not maintain robustness when many factors were introduced. At τ^2 =0, FE performed much like the RE and CR tests. But with increases in τ^2 and the introduction of between-studies variance, the median error rate rose from just under .05 to .30 (for τ^2 =.33, Figure 7) and .50 (at τ^2 =1, Figure 9).

The RE and CR marginal error rates closely reflected the performance of the other. When τ^2 =0, both tests' median fell below the nominal α of .05. Therefore, these tests provided conservative estimates of α. When τ^2 values increased from 0 to .33 and 1.0, median estimates greater than .05 ensued. Though the tests produced marginal error rates less inflated than those resulting from the FE test, those rates still exceeded .05.

Only Permuted Q remained robust to Type I error across all τ^2 conditions. The median rejection rates closely approximated .05. Indeed, most rejection rates did not deviate far from this value.

Figure 5
Distribution of Type I Error Rate Estimates Across Experimental Conditions for $\tau^2 = 0$

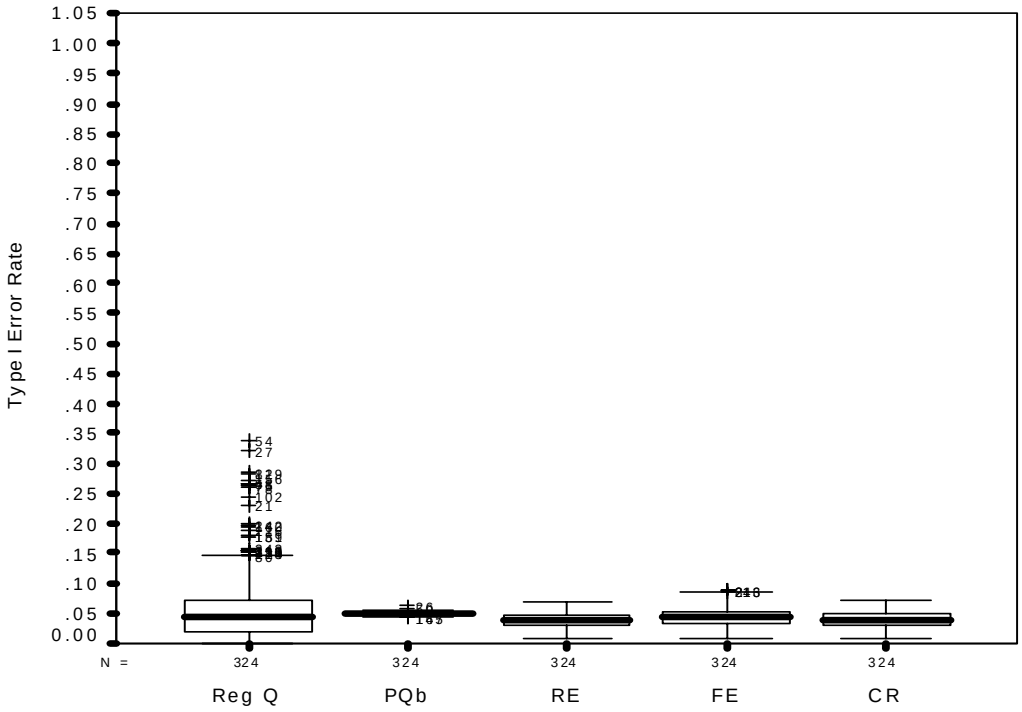
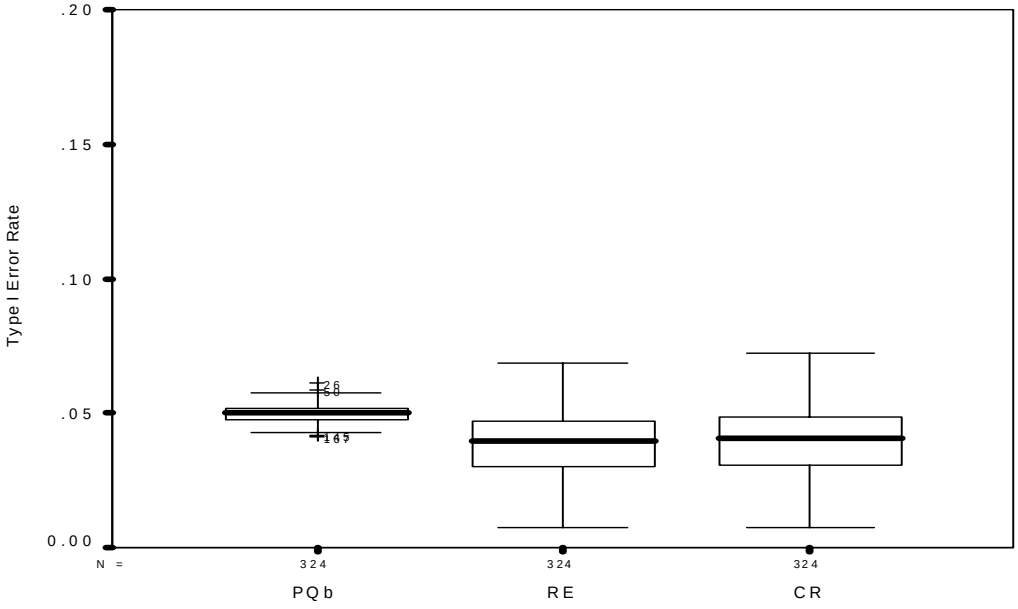


Figure 6
Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for $\tau^2 = 0$



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Figure 7

Distribution of Type I Error Rate Estimates Across Experimental Conditions for $\tau^2 = .33$

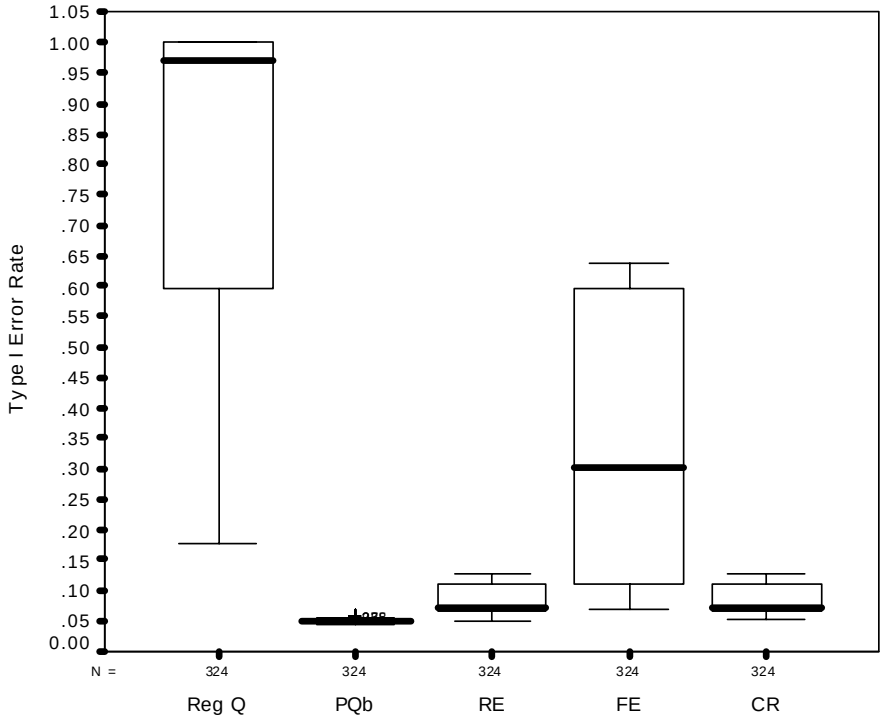
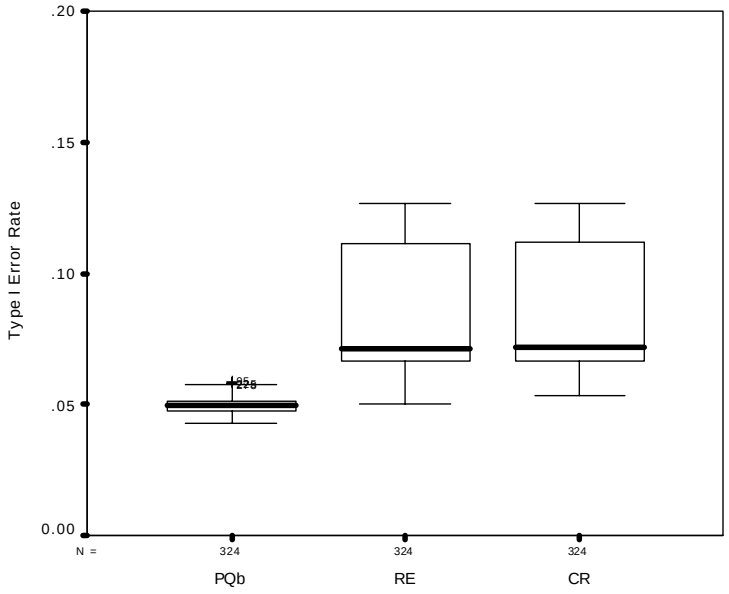


Figure 8

Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for $\tau^2 = .33$



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Figure 9

Distribution of Type I Error Rate Estimates Across Experimental Conditions for $\tau^2 = 1.0$

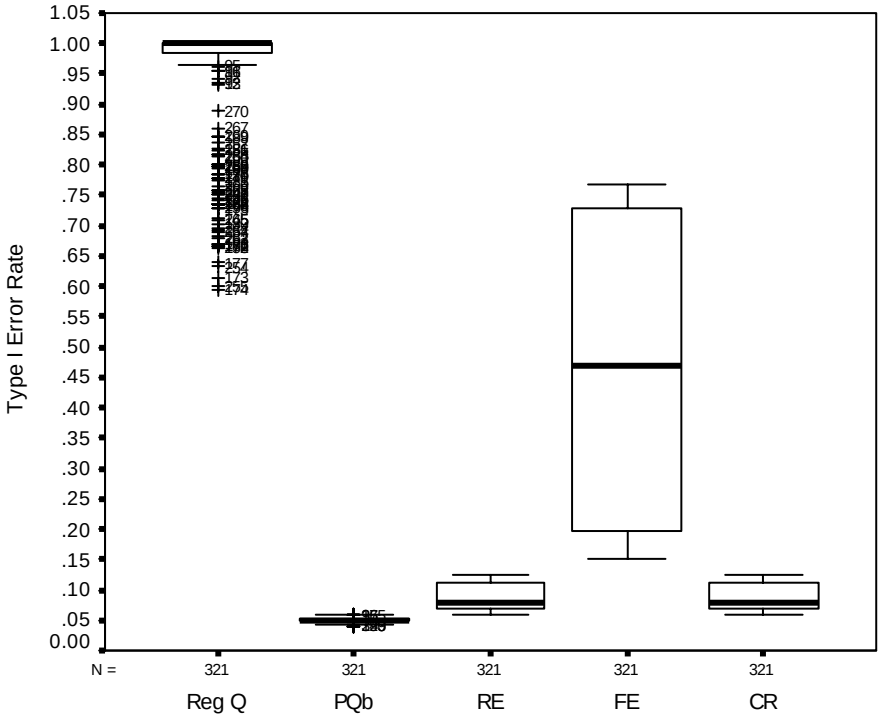
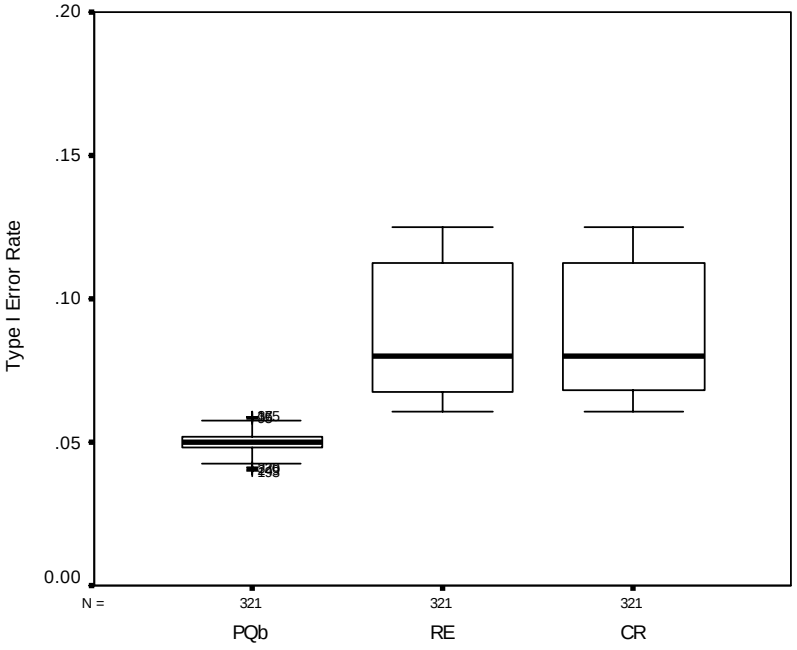


Figure 10

Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for $\tau^2 = 1.0$



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Box Plots for Primary Study Sample Size

Harwell (1997) suspected that small N paired with large K contributes to regular Q's greater tendency for inflation of Type I error. Applying the FE test while increasing primary study sample size resulted in increasingly inflated Type I error. Q's median rose from just over nominal α .05 to .30 (N=40, see Figure 13) and .60 (for N=200, see Figure 15). This result stands in contrast to Harwell's conclusion, except that K was not controlled for in this specific analysis.

As sample size increased, RE and CR performance again reflected the same pattern. At N=10 (see Figure 11), the median rates exceeded nominal alpha, but not .10. Although the median remained relatively unchanged at all 3 levels of N, the 3rd quartile rose from approximately .07 to .11, as N was elevated from 10 to 40. The dispersion of error rates did not change dramatically from N=40 to N=200.

Lastly, the permuted Q maintained adequate Type I error control across all 3 levels. In fact, the distribution of error rates remained unchanged from N=10 to N=200.

Figure 11
Distribution of Type I Error Rate Estimates Across Experimental Conditions for N = 10

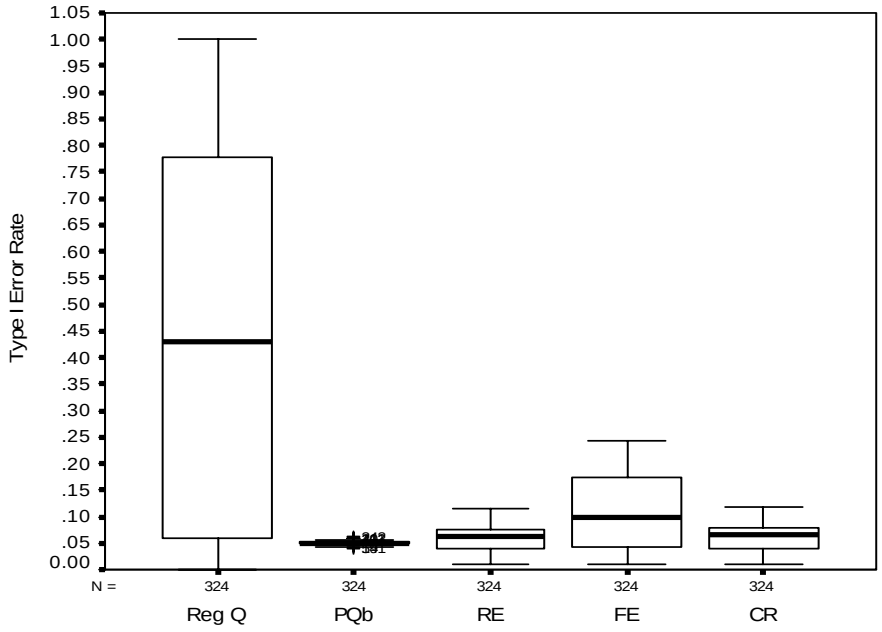
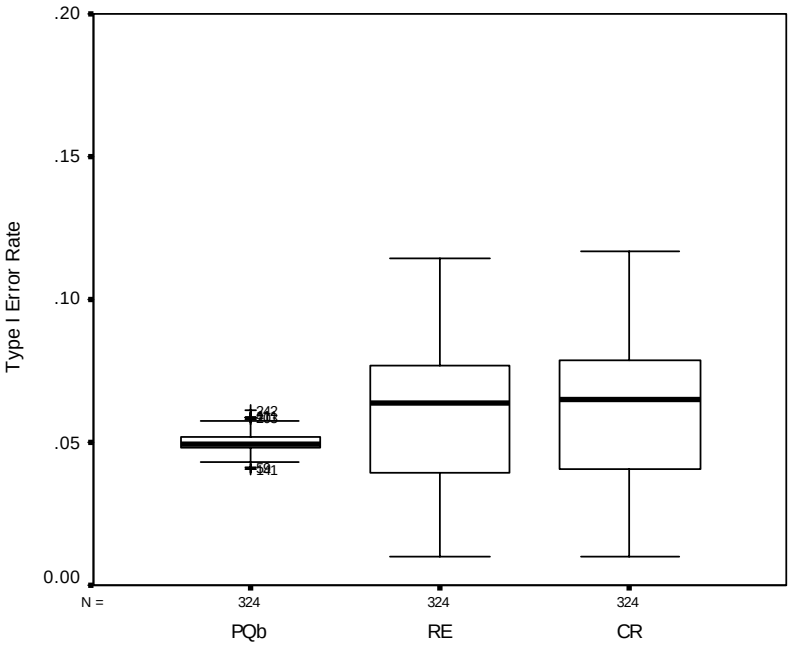


Figure 12
Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for N = 10



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Figure 13
Distribution of Type I Error Rate Estimates Across Experimental Conditions for N = 40

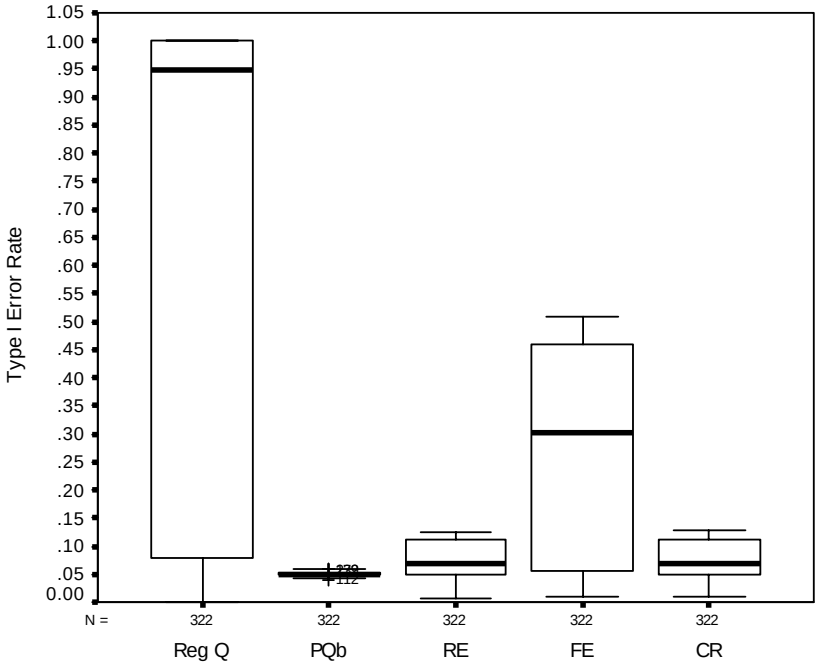
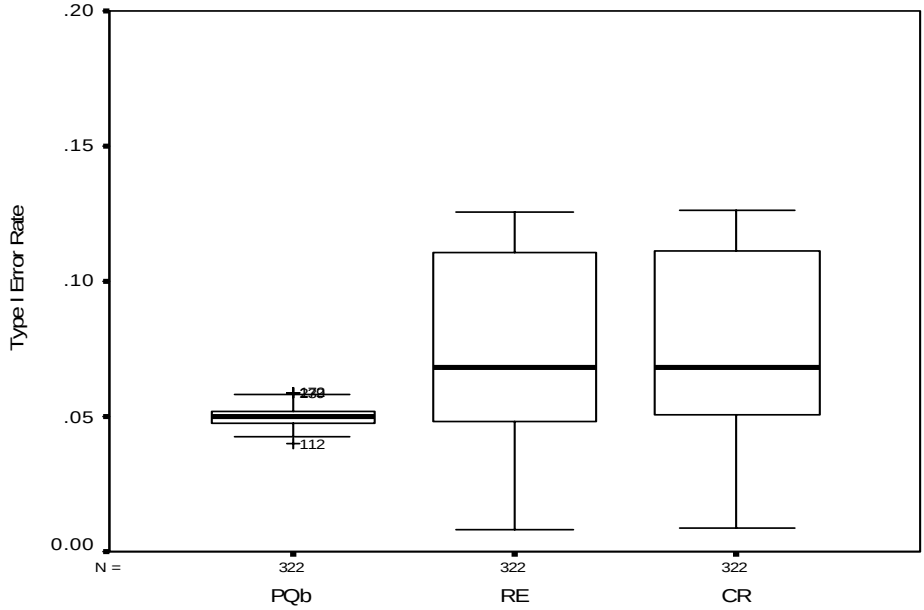


Figure 14
Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for N = 40



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Figure 15
Distribution of Type I Error Rate Estimates Across Experimental Conditions for N = 200

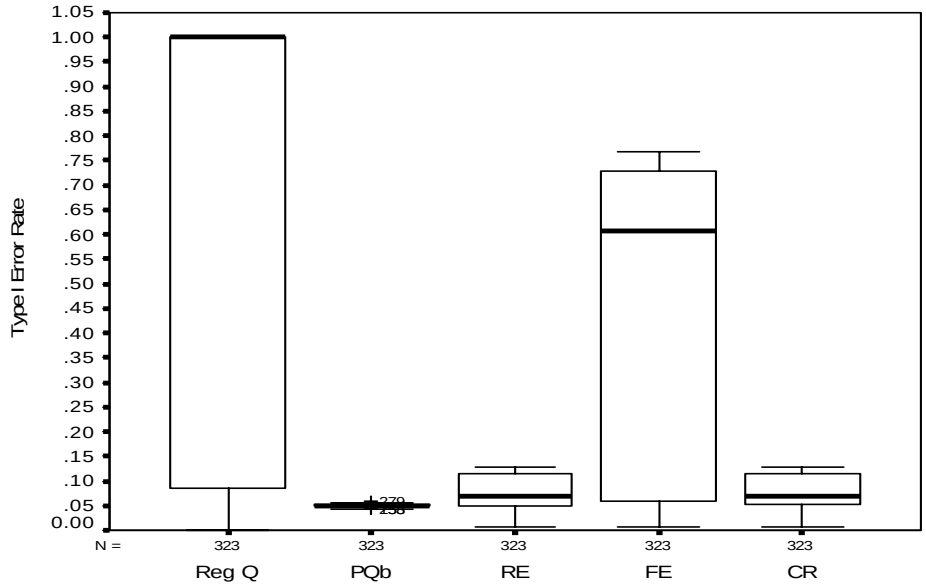
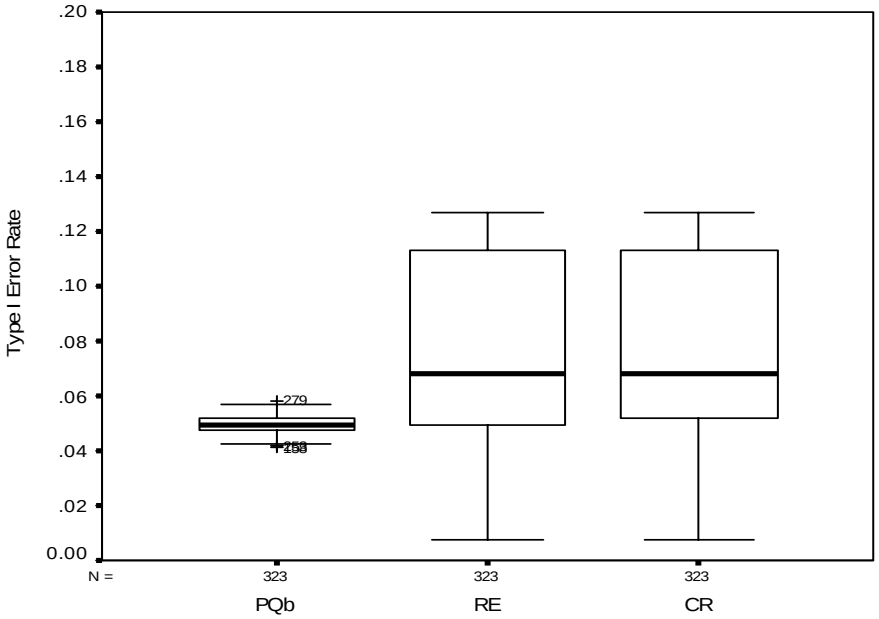


Figure 16
Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for N = 200



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Box Plots for Variance Heterogeneity

Median error rates tended to decrease for the regular Q and the FE tests (see Figures 17, 19 and 21), as the difference in population variances increased. In contrast, median error rates remained fairly constant for the RE and CR tests (refer to Figures 18, 20 and 22). What did change for these latter 2 tests was the line delineating the third quartile of error rates (.075 at sds=1/1; .09 at sds = 2/1; and .11 at sds=4/1).

Again, permuted Q maintained Type I error rates, with marginal error rates clustered tightly around the nominal alpha. This test appeared unaffected by changes in the variance from equal to unequal variances within groups.

Figure 17
Distribution of Type I Error Rate Estimates Across Experimental Conditions for sds = 1/1

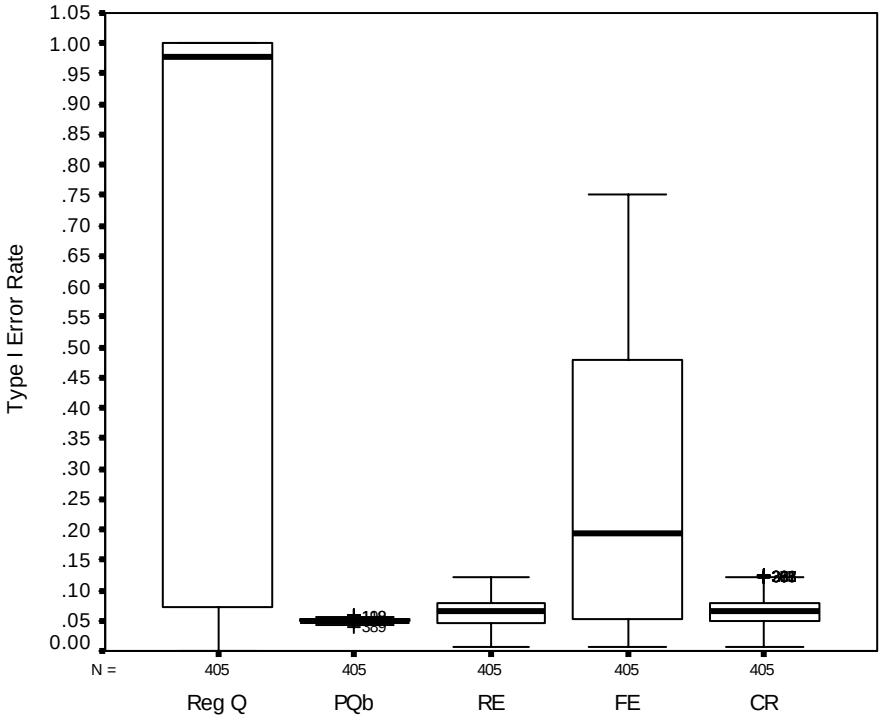


Figure 18
Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for sds = 1/1

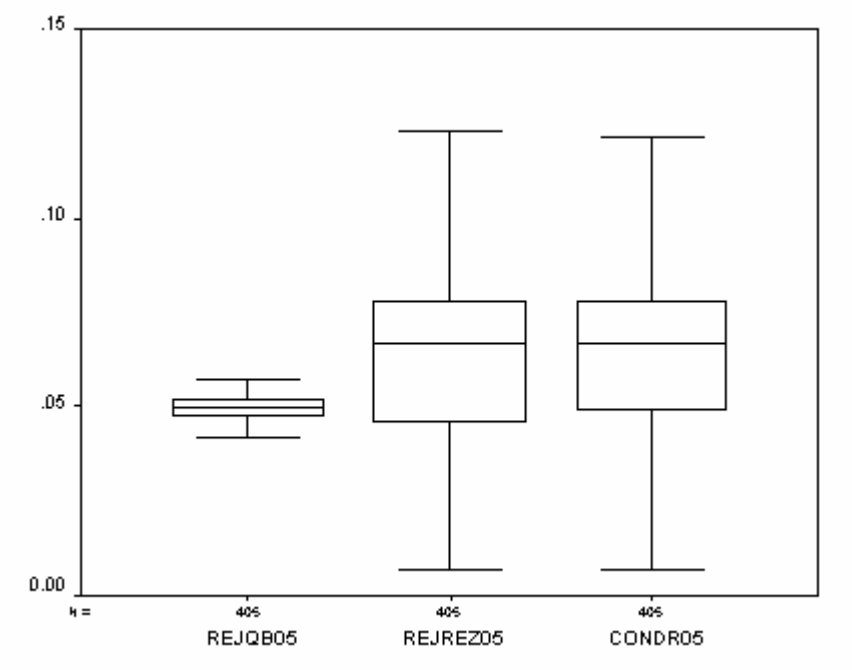


Figure 19
Distribution of Type I Error Rate Estimates Across Experimental Conditions for sds = 2/1

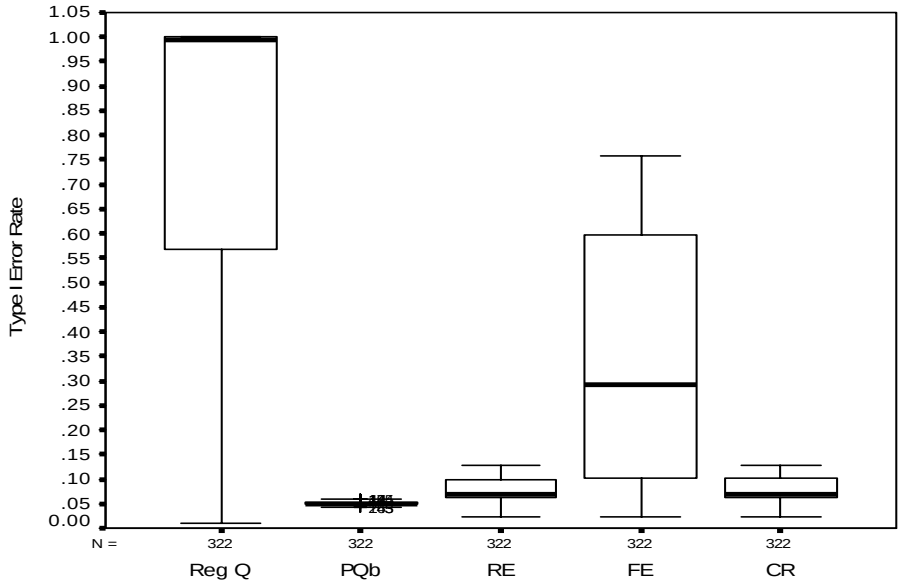
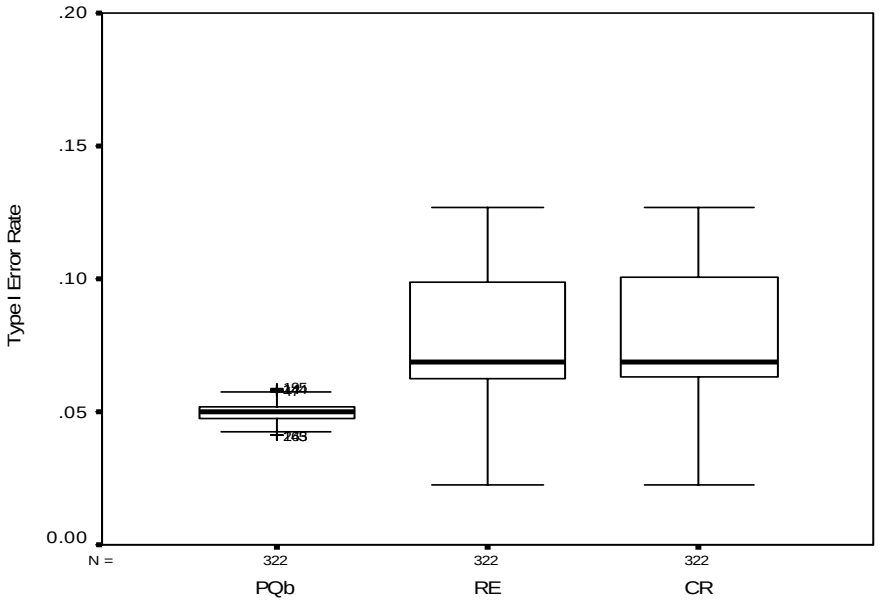


Figure 20
Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for sds = 2/1



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Figure 21
Distribution of Type I Error Rate Estimates Across Experimental Conditions for sds = 4/1

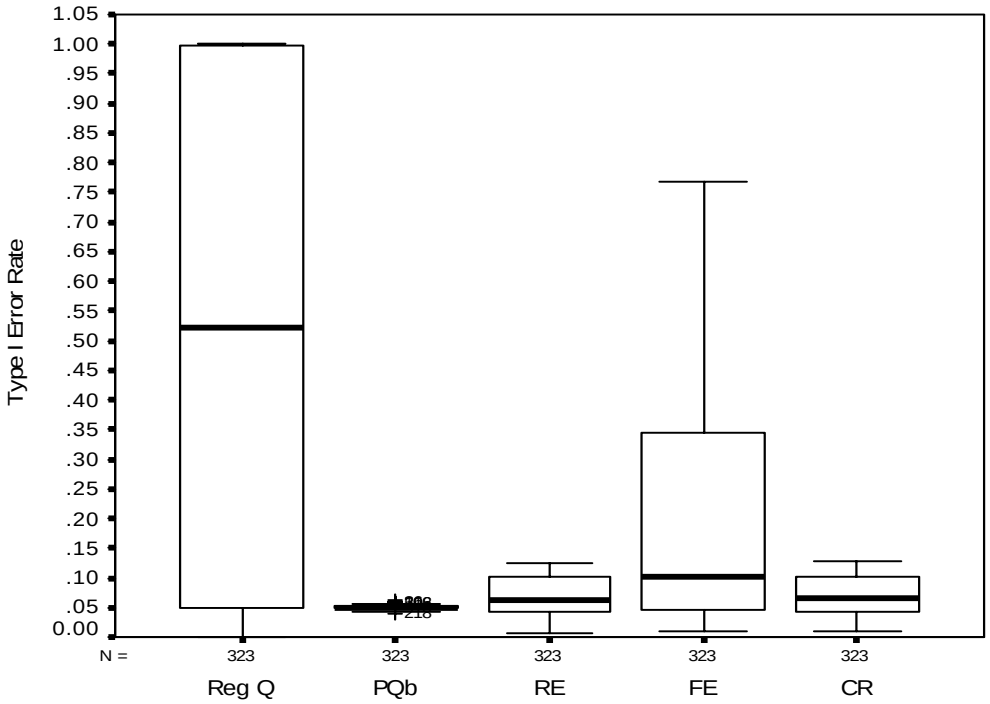
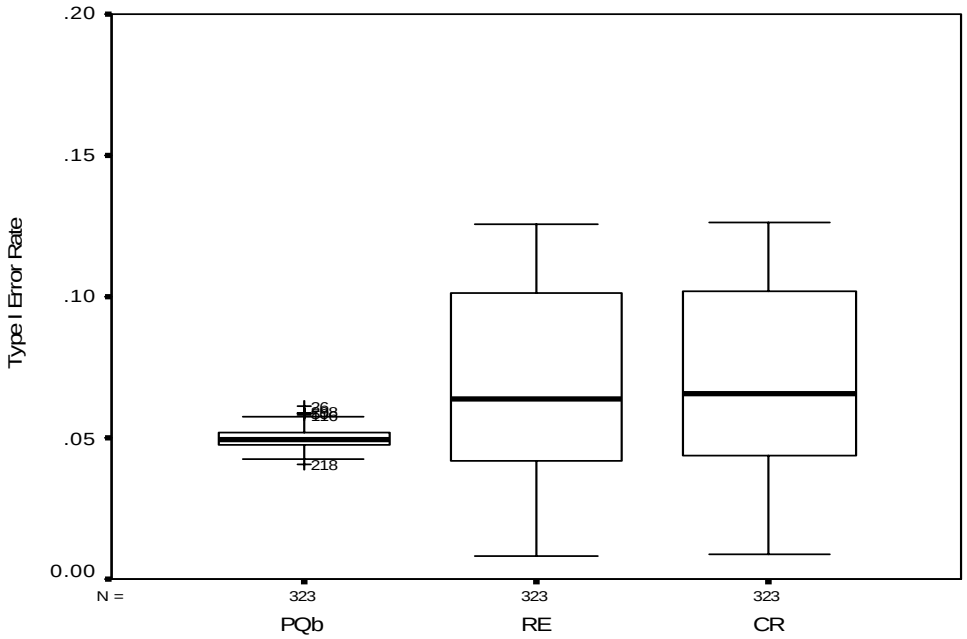


Figure 22
Magnified Distribution of Type I Error Rate Estimates Across Experimental Conditions for sds = 4/1



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Box Plots for Skewness/Kurtosis

Contrary to the findings established by Kromrey & Hogarty (1998) and Harwell (1997), changing the population distribution did not significantly alter Type I error rates for any of the 5 tests under consideration. It should be noted that the box plots provide a general overview and do not permit examination of multiple factors held constant all at once.

Permuted Q maintained adequate Type I error control, as marginal error rates concentrated around nominal alpha. In other words, Permuted Q was unaffected by the population distribution.

The RE and CR tests both demonstrated a median marginal error rate just over .05, but less than .10 (see Figures 24, 26 and 28). At least 25% of all error rates fell below .05. None of the error rates exceeded .15. This pattern remained constant across the 3 levels of skewness and kurtosis.

The median error rates for regular Q and the FE tests remained at .80 and .15, respectively (see Figures 23, 25 and 27). Because the application of none of these 5 tests resulted in varying error rates, given differing degrees of population shape, it can be concluded that distribution shape did not have a markedly significant effect on Type I error control.

Figure 23
Distribution of Type I Error Rate Estimates Across Experimental Conditions for skewness/kurtosis = 1/1

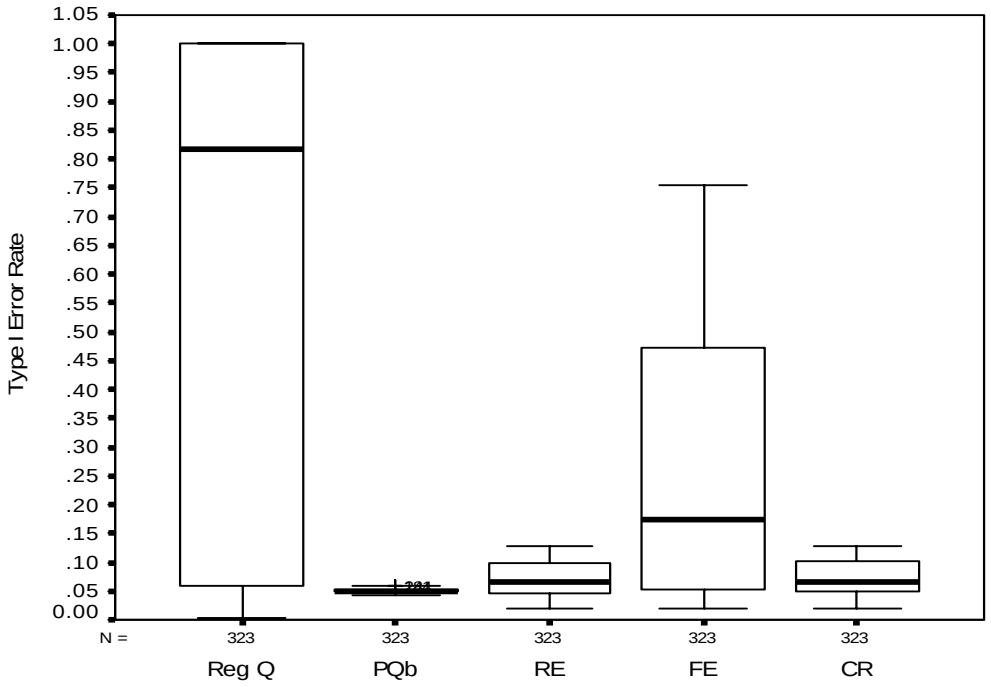
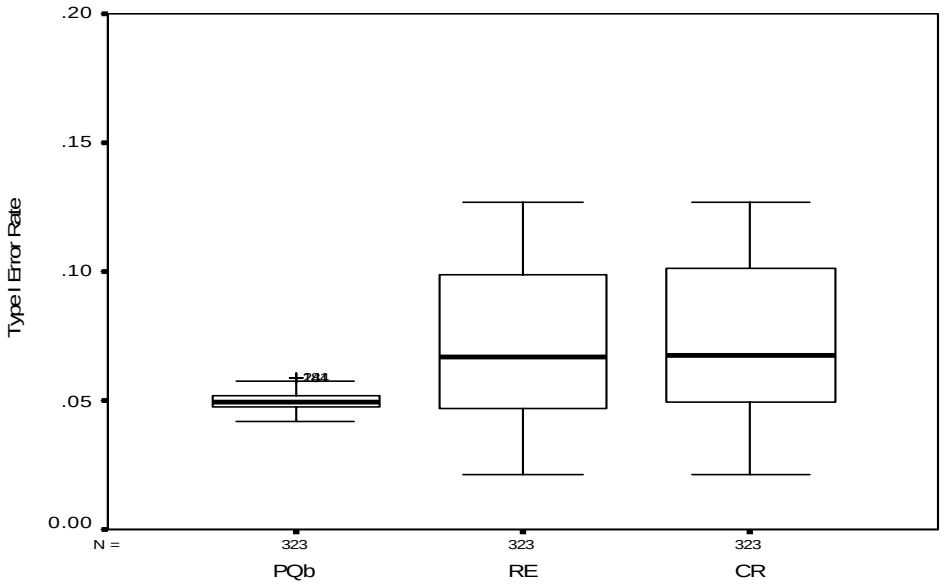


Figure 24
Magnification of Type I Error Rate Estimates Across Experimental Conditions for skewness/kurtosis = 1/1



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Figure 25
Distribution of Type I Error Rate Estimates Across Experimental Conditions for skewness/kurtosis = 1/3

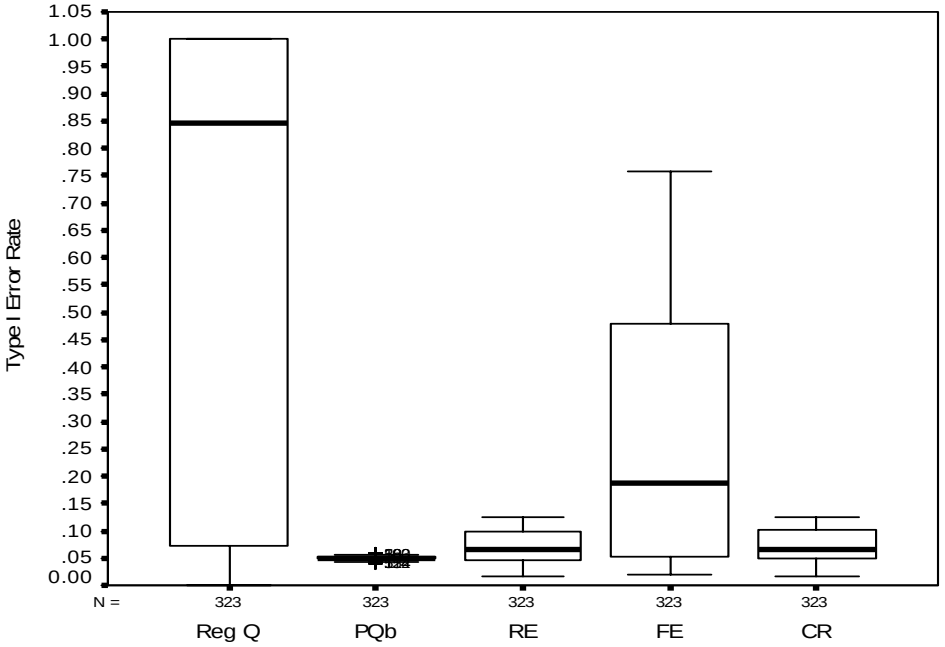
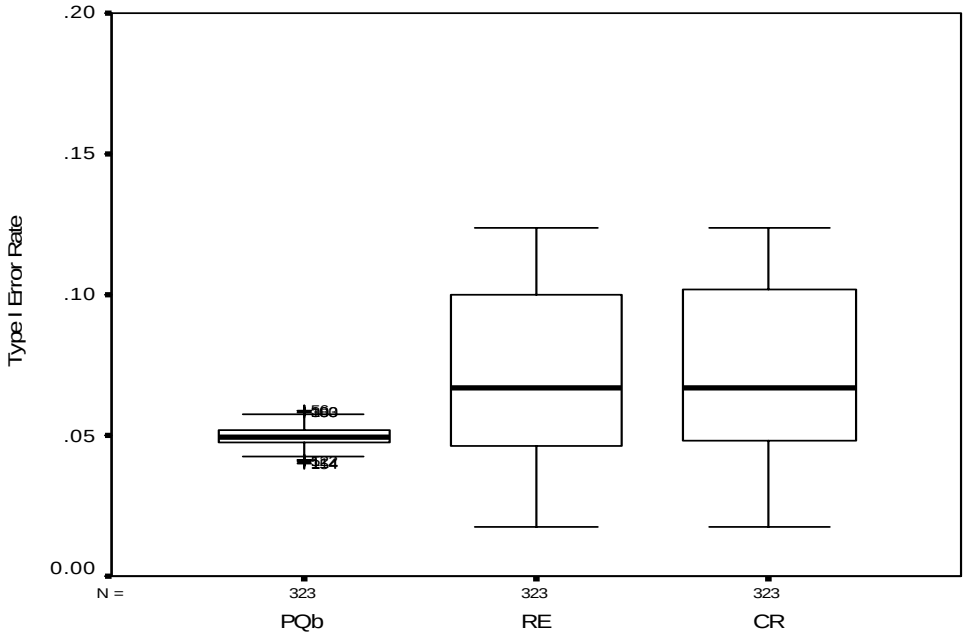


Figure 26
Magnification of Type I Error Rate Estimates Across Experimental Conditions for skewness/kurtosis = 1/3



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Figure 27
Distribution of Type I Error Rate Estimates Across Experimental Conditions for skewness/kurtosis = 2/6

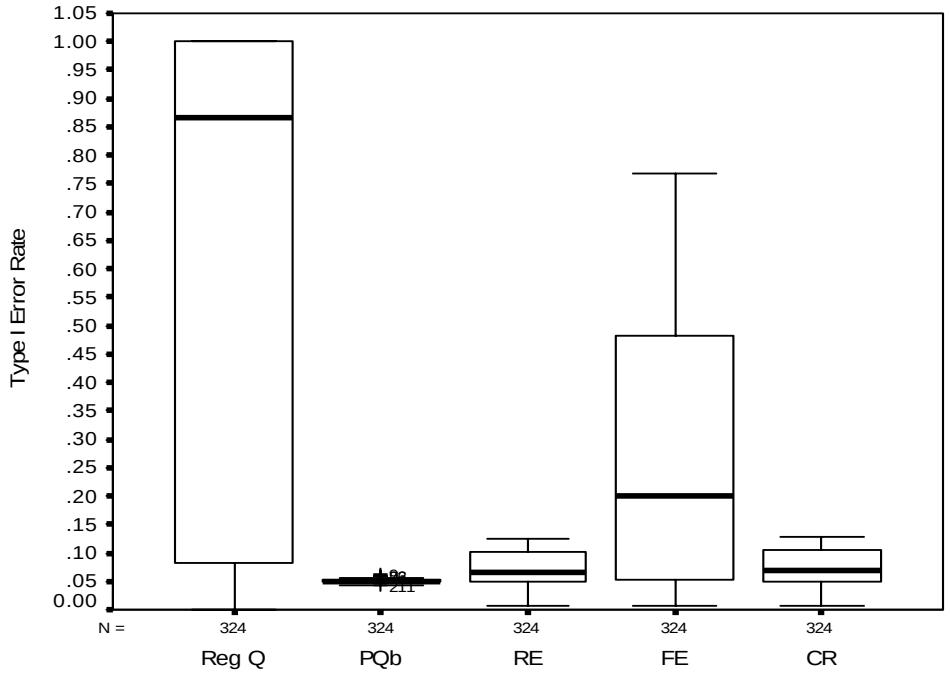
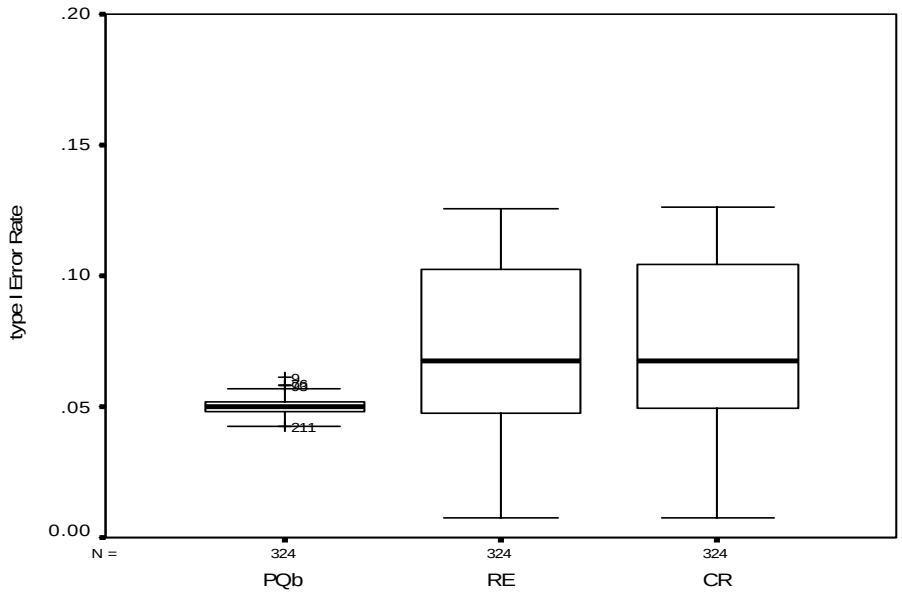


Figure 28
Magnification of Type I Error Rate Estimates Across Experimental Conditions for skewness/kurtosis = 2/6



*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Summary of Box Plots

In short, small K, increasing τ^2 and increasing primary study sample sizes resulted in increasingly inflated Type I error for all tests (except permuted Q). This trend was most pronounced for regular Q and FE, and to a lesser extent, RE and CR tests.

Changes in the population shape did not result in significant changes in the tests' performance, only in differences between tests. But these patterns of performance remained constant for each test across conditions of increasing skewness and kurtosis.

Permuted Q uniquely controlled Type I error across all changes in conditions. It was the only test to maintain Type I error rates near .05, regardless of the extremity of the isolated variable.

Proportion of Simulations Controlling Type I Error

The proportion of simulations were calculated by summing the number of conditions with marginal rejection rates meeting Bradley's criterion for acceptable Type I error control. For nominal α = .05, the range of acceptability was restricted to rejection rates no smaller than .045 and smaller or equal to .055.

The proportion tables for both τ^2 =.33 and τ^2 = 1 (Tables 5 and 6 for τ^2 =.33, and tables 7 and 8 for τ^2 =1) dovetail the results illustrated in the box plots pertaining to τ^2. Increases in τ^2 superceded all other factors in terms of producing inflated Type I error rates. Due to this effect, only the permuted Q consistently showed proportions of simulations with adequate Type I error greater than 50%. All other tests yielded either no proportions of adequate Type I error or only a few conditions with proportions less than 50% showing adequacy of Type I error control. This pattern was true for τ^2 =.33 and τ^2 = 1, regardless of whether δ =0 or .8.

It is worth noting that τ^2 = 0, δ=.8 returned more conditions with proportions of simulations with adequate Type I error control than τ^2 = 0, δ=0. Here, the FE test presented either comparable or greater proportions of Type I error control than either the RE or CR tests, particularly for primary study sample sizes of 6/4, 20/20 and 100/100. As described earlier, this influence can be explained by the model underlying the formulation of the FE test.

Table 3
Proportion of Simulations with Adequate Type I Error Control ($\tau^2 = 0$, $\delta = 0$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.222	**0.889**	0.000	0.111	0.000
4/6	4/6	0.000	**1.000**	0.000	0.000	0.000
6/4	4/6	0.000	**0.889**	0.111	0.222	0.111
20/20	4/6	0.444	0.444	0.444	**0.667**	**0.556**
16/24	4/6	0.000	**0.889**	0.111	0.111	0.111
24/16	4/6	0.111	**0.889**	0.222	0.111	0.000
100/100	4/6	**0.889**	**0.889**	**0.556**	**0.667**	**0.667**
80/120	4/6	0.222	**0.889**	0.111	0.222	0.222
120/80	4/6	0.222	**1.000**	0.444	0.333	0.333
5/5	12/18	0.222	**1.000**	0.000	0.000	0.000
4/6	12/18	0.000	**1.000**	0.000	0.000	0.000
6/4	12/18	0.000	**0.778**	0.444	0.111	0.444
20/20	12/18	0.222	**0.889**	0.222	0.333	0.333
16/24	12/18	0.000	**0.778**	0.000	0.111	0.111
24/16	12/18	0.000	**0.889**	0.444	0.000	0.222
100/100	12/18	**0.556**	**0.889**	0.222	**1.000**	**0.778**
80/120	12/18	0.222	**1.000**	0.222	0.333	0.333
120/80	12/18	0.111	**0.778**	0.333	0.222	0.222

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.259	**0.852**	0.222	0.444	0.296
sk=1.00, kr=3.00	4/6	0.222	**0.852**	0.259	0.222	0.185
sk=2.00, kr=6.00	4/6	0.222	**0.926**	0.185	0.148	0.185
sk=0.00, kr=0.00	12/18	0.148	**0.815**	0.222	0.222	0.296
sk=1.00, kr=3.00	12/18	0.148	**0.963**	0.259	0.259	0.185
sk=2.00, kr=6.00	12/18	0.148	**0.889**	0.185	0.222	0.333

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.259	**0.926**	0.222	0.444	0.296
1/2	4/6	0.185	**0.778**	0.259	0.222	0.185
1/4	4/6	0.259	**0.889**	0.185	0.148	0.185
1/1	12/18	0.148	**0.926**	0.111	0.333	0.259
1/2	12/18	0.148	**0.778**	0.222	0.185	0.296
1/4	12/18	0.148	**0.963**	0.296	0.185	0.259

Note: Proportions reflecting more than half of the conditions expressing adequate Type I error control are bolded.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

In all but one of the conditions for $\tau^2 = 0$, $\delta = 0$, Permuted Q evidenced the greatest proportion of simulations with adequate Type I error control. Neither distribution shape nor within group variances significantly alters this pattern. Furthermore, changes in K do not lend a significant influence.

Table 4

Proportion of Simulations with Adequate Type I Error Control ($\tau^2 = 0$, $\delta = .8$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.111	**1.000**	0.111	0.111	0.111
4/6	4/6	0.000	**1.000**	0.111	0.111	0.111
6/4	4/6	0.222	**1.000**	0.444	0.333	0.333
20/20	4/6	0.111	**0.889**	0.333	0.444	**0.556**
16/24	4/6	0.000	**1.000**	0.222	0.111	0.111
24/16	4/6	0.222	**1.000**	**0.556**	**0.556**	0.444
100/100	4/6	0.222	**0.889**	0.333	0.333	0.333
80/120	4/6	0.111	**1.000**	0.222	0.222	0.222
120/80	4/6	0.222	**1.000**	**0.556**	0.444	0.444
5/5	12/18	0.000	**0.889**	0.000	0.111	0.111
4/6	12/18	0.000	**0.889**	0.000	0.000	0.000
6/4	12/18	0.111	**1.000**	0.111	0.222	0.111
20/20	12/18	0.111	**0.889**	0.333	0.444	0.444
16/24	12/18	0.000	**1.000**	0.222	0.222	0.111
24/16	12/18	0.111	**1.000**	**0.556**	0.333	**0.556**
100/100	12/18	0.222	**0.889**	**0.556**	0.333	**0.556**
80/120	12/18	0.111	**0.778**	0.222	0.000	0.111
120/80	12/18	0.222	**0.889**	**0.556**	0.333	0.333

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.148	**0.926**	0.222	0.296	0.259
sk=1.00, kr=3.00	4/6	0.074	**0.926**	0.370	0.259	0.407
sk=2.00, kr=6.00	4/6	0.074	**0.926**	0.148	0.148	0.148
sk=0.00, kr=0.00	12/18	0.222	**0.889**	0.259	0.259	0.259
sk=1.00, kr=3.00	12/18	0.037	**0.852**	0.296	0.222	0.370
sk=2.00, kr=6.00	12/18	0.037	**0.963**	0.222	0.074	0.037

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.148	**0.926**	0.407	0.333	0.482
1/2	4/6	0.148	**0.889**	0.222	0.222	0.185
1/4	4/6	0.000	**0.963**	0.111	0.148	0.148
1/1	12/18	0.148	**0.889**	0.444	0.222	0.259
1/2	12/18	0.148	**0.889**	0.111	0.185	0.222
1/4	12/18	0.000	**0.926**	0.222	0.148	0.185

Note: Proportions reflecting more than half of the conditions expressing adequate Type I error control are bolded.

In 9 of the conditions utilizing Permuted Q, 100% of the simulations resulted in adequate Type I error control. For each of the other 4 tests, the proportions with adequate Type I error did not exceed 55%.

From $\tau^2=0$, $\delta=0$ (Table 3) to $\tau^2=0$, $\delta=.8$ (Table 4), the RE test yielded increases in the proportion of adequate Type I error for the equal variance conditions, whether K=10 or 30. As the variance ratios increased, proportions remained lower for this test. The CR test showed an increase in the proportion when variances were equal and K=10, only. For all other variance ratios, the CR test's proportions remained fairly constant. The FE test demonstrated decreased proportions at the equal variance condition.

Table 5

Proportion of Simulations with Adequate Type I Error Control ($\tau^2 = .33$, $\delta = 0$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	**4/6**	0.000	**0.889**	0.000	0.000	0.000
4/6	**4/6**	0.000	**0.889**	0.000	0.000	0.000
6/4	**4/6**	0.000	**1.000**	0.000	0.000	0.000
20/20	**4/6**	0.000	**1.000**	0.000	0.000	0.000
16/24	**4/6**	0.000	**0.889**	0.000	0.000	0.000
24/16	**4/6**	0.000	**1.000**	0.000	0.000	0.000
100/100	**4/6**	0.000	**1.000**	0.000	0.000	0.000
80/120	**4/6**	0.000	**1.000**	0.000	0.000	0.000
120/80	**4/6**	0.000	**1.000**	0.000	0.000	0.000
5/5	**12/18**	0.000	**0.778**	0.000	0.000	0.000
4/6	**12/18**	0.000	**1.000**	0.111	0.000	0.000
6/4	**12/18**	0.000	**1.000**	0.000	0.000	0.000
20/20	**12/18**	0.000	**0.889**	0.000	0.000	0.000
16/24	**12/18**	0.000	**1.000**	0.000	0.000	0.000
24/16	**12/18**	0.000	**1.000**	0.000	0.000	0.000
100/100	**12/18**	0.000	**0.889**	0.000	0.000	0.000
80/120	**12/18**	0.000	**0.889**	0.000	0.000	0.000
120/80	**12/18**	0.000	**1.000**	0.000	0.000	0.000

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.000	**0.963**	0.000	0.000	0.000
sk=1.00, kr=3.00	4/6	0.000	**0.926**	0.000	0.000	0.000
sk=2.00, kr=6.00	4/6	0.000	**0.889**	0.000	0.000	0.000
sk=0.00, kr=0.00	12/18	0.000	**0.963**	**0.037**	0.000	0.000
sk=1.00, kr=3.00	12/18	0.000	**0.926**	0.000	0.000	0.000
sk=2.00, kr=6.00	12/18	0.000	**0.815**	0.000	0.000	0.000

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.000	**0.926**	0.000	0.000	0.000
1/2	4/6	0.000	**0.889**	0.000	0.000	0.000
1/4	4/6	0.000	**0.963**	0.000	0.000	0.000
1/1	12/18	0.000	**0.926**	0.000	0.000	0.000
1/2	12/18	0.000	**0.926**	**0.037**	0.000	0.000
1/4	12/18	0.000	**0.852**	0.000	0.000	0.000

Note: Proportions reflecting any conditions expressing adequate Type I error control are bolded.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

These results (see Table 5) substantiate the box plots outcome previously reported. As τ^2 increased from 0 to .33, all tests other than permuted Q displayed either trace or no proportions of adequate Type I error control. Given the constant absence of Type I error control across conditions, there is evidence of the significant influence of this increase in τ^2 on Type I error control for 4 of the 5 tests.

Table 6
Proportion of Simulations with Adequate Type I Error Control ($\tau^2 = .33$, $\delta = .8$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.000	**1.000**	0.000	0.000	0.000
4/6	4/6	0.000	**0.778**	0.000	0.000	0.000
6/4	4/6	0.000	**1.000**	0.000	0.000	0.000
20/20	4/6	0.000	**0.889**	0.000	0.000	0.000
16/24	4/6	0.000	**0.889**	0.000	0.000	0.000
24/16	4/6	0.000	**1.000**	0.000	0.000	0.000
100/100	4/6	0.000	**1.000**	0.000	0.000	0.000
80/120	4/6	0.000	**0.778**	0.000	0.000	0.000
120/80	4/6	0.000	**1.000**	0.000	0.000	0.000
5/5	12/18	0.000	**0.889**	0.000	0.000	0.000
4/6	12/18	0.000	**0.889**	**0.333**	0.000	**0.333**
6/4	12/18	0.000	**1.000**	**0.111**	0.000	0.000
20/20	12/18	0.000	**0.667**	0.000	0.000	0.000
16/24	12/18	0.000	**1.000**	0.000	0.000	0.000
24/16	12/18	0.000	**1.000**	0.000	0.000	0.000
100/100	12/18	0.000	**0.889**	0.000	0.000	0.000
80/120	12/18	0.000	**0.889**	0.000	0.000	0.000
120/80	12/18	0.000	**1.000**	0.000	0.000	0.000

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.000	**0.926**	0.000	0.000	0.000
sk=1.00, kr=3.00	4/6	0.000	**0.889**	0.000	0.000	0.000
sk=2.00, kr=6.00	4/6	0.000	**0.889**	0.000	0.000	0.000
sk=0.00, kr=0.00	12/18	0.000	**0.852**	**0.037**	0.000	**0.037**
sk=1.00, kr=3.00	12/18	0.000	**0.926**	**0.037**	0.000	**0.037**
sk=2.00, kr=6.00	12/18	0.000	**0.889**	**0.037**	0.000	**0.037**

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.000	**0.926**	0.000	0.000	0.000
1/2	4/6	0.000	**0.889**	0.000	0.000	0.000
1/4	4/6	0.000	**0.889**	0.000	0.000	0.000
1/1	12/18	0.000	**0.926**	0.000	0.000	0.000
1/2	12/18	0.000	**0.852**	0.000	0.000	0.000
1/4	12/18	0.000	**0.889**	**0.111**	0.000	**0.111**

Note: Proportions reflecting any conditions expressing adequate Type I error control are bolded.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Again, with a few exceptions, only permuted Q evidenced a significant proportion of simulations with adequate Type I error control (see Table 6 above).

Table 7

Proportion of Simulations with Adequate Type I Error Control ($\tau^2 = 1$, $\delta = 0$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	**4/6**	0.000	**0.778**	0.000	0.000	0.000
4/6	**4/6**	0.000	**0.889**	0.000	0.000	0.000
6/4	**4/6**	0.000	**1.000**	0.000	0.000	0.000
20/20	**4/6**	0.000	**0.778**	0.000	0.000	0.000
16/24	**4/6**	0.000	**1.000**	0.000	0.000	0.000
24/16	**4/6**	0.000	**0.889**	0.000	0.000	0.000
100/100	**4/6**	0.000	**0.889**	0.000	0.000	0.000
80/120	**4/6**	0.000	**0.778**	0.000	0.000	0.000
120/80	**4/6**	0.000	**0.889**	0.000	0.000	0.000
5/5	**12/18**	0.000	**1.000**	0.000	0.000	0.000
4/6	**12/18**	0.000	**1.000**	0.000	0.000	0.000
6/4	**12/18**	0.000	**1.000**	0.000	0.000	0.000
20/20	**12/18**	0.000	**1.000**	0.000	0.000	0.000
16/24	**12/18**	0.000	**0.667**	0.000	0.000	0.000
24/16	**12/18**	0.000	**1.000**	0.000	0.000	0.000
100/100	**12/18**	0.000	**1.000**	0.000	0.000	0.000
80/120	**12/18**	0.000	**0.889**	0.000	0.000	0.000
120/80	**12/18**	0.000	**0.778**	0.000	0.000	0.000

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	**4/6**	0.000	**0.963**	0.000	0.000	0.000
sk=1.00, kr=3.00	**4/6**	0.000	**0.815**	0.000	0.000	0.000
sk=2.00, kr=6.00	**4/6**	0.000	**0.889**	0.000	0.000	0.000
sk=0.00, kr=0.00	**12/18**	0.000	**0.889**	0.000	0.000	0.000
sk=1.00, kr=3.00	**12/18**	0.000	**0.963**	0.000	0.000	0.000
sk=2.00, kr=6.00	**12/18**	0.000	**0.926**	0.000	0.000	0.000

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	**4/6**	0.000	**0.926**	0.000	0.000	0.000
1/2	**4/6**	0.000	**0.852**	0.000	0.000	0.000
1/4	**4/6**	0.000	**0.852**	0.000	0.000	0.000
1/1	**12/18**	0.000	**0.963**	0.000	0.000	0.000
1/2	**12/18**	0.000	**0.889**	0.000	0.000	0.000
1/4	**12/18**	0.000	**0.926**	0.000	0.000	0.000

Note: Proportions reflecting any conditions expressing adequate Type I error control are bolded.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The pattern established as τ^2 increased from 0 to .33 continued from .33 to 1.0 (see Table 7). It did not make a difference whether δ=0 or .8. Because the pattern of an absence in any proportions of adequate Type I error continued from $\tau^2 = .33$ to 1, one can conclude that increasing heterogeneity of effects had a significant influence on the effectiveness of the performance of these tests. Only permuted Q permitted the maintenance of adequate Type I error control in more than 50% of the simulations. Despite increases in heterogeneity, permuted Q maintained robustness across all changes in variance and population shape.

Table 8
Proportion of Simulations with Adequate Type I Error Control ($\tau^2 = 1$, $\delta = .8$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.000	**0.778**	0.000	0.000	0.000
4/6	4/6	0.000	**1.000**	0.000	0.000	0.000
6/4	4/6	0.000	**1.000**	0.000	0.000	0.000
20/20	4/6	0.000	**0.556**	0.000	0.000	0.000
16/24	4/6	0.000	**0.778**	0.000	0.000	0.000
24/16	4/6	0.000	**0.889**	0.000	0.000	0.000
100/100	4/6	0.000	**1.000**	0.000	0.000	0.000
80/120	4/6	0.000	**1.000**	0.000	0.000	0.000
120/80	4/6	0.000	**1.000**	0.000	0.000	0.000
5/5	12/18	0.000	**0.889**	0.000	0.000	0.000
4/6	12/18	0.000	**0.889**	0.000	0.000	0.000
6/4	12/18	0.000	**0.778**	0.000	0.000	0.000
20/20	12/18	0.000	**1.000**	0.000	0.000	0.000
16/24	12/18	0.000	**0.778**	0.000	0.000	0.000
24/16	12/18	0.000	**0.889**	0.000	0.000	0.000
100/100	12/18	0.000	**0.778**	0.000	0.000	0.000
80/120	12/18	0.000	**1.000**	0.000	0.000	0.000
120/80	12/18	0.000	**1.000**	0.000	0.000	0.000

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.000	**0.778**	0.000	0.000	0.000
sk=1.00, kr=3.00	4/6	0.000	**0.852**	0.000	0.000	0.000
sk=2.00, kr=6.00	4/6	0.000	**0.926**	0.000	0.000	0.000
sk=0.00, kr=0.00	12/18	0.000	**0.852**	0.000	0.000	0.000
sk=1.00, kr=3.00	12/18	0.000	**0.778**	0.000	0.000	0.000
sk=2.00, kr=6.00	12/18	0.000	**0.963**	0.000	0.000	0.000

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.000	**0.926**	0.000	0.000	0.000
1/2	4/6	0.000	**0.889**	0.000	0.000	0.000
1/4	4/6	0.000	**0.741**	0.000	0.000	0.000
1/1	12/18	0.000	**0.889**	0.000	0.000	0.000
1/2	12/18	0.000	**0.852**	0.000	0.000	0.000
1/4	12/18	0.000	**0.852**	0.000	0.000	0.000

Note: Proportions reflecting any conditions expressing adequate Type I error control are bolded.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The same pattern of Type I error control is evidenced regardless of the value of δ (see Table 8).

Therefore, one can conclude that increases in τ^2 negatively influenced the effect on Type I error.

Average Type I Error Rate Estimates

From the previous presentation of the proportion of simulations with adequate Type I error, it is clear that τ^2 exerted a substantial influence on the control of Type I error for all of the tests, except permuted Q. Permuted Q maintained robustness across all conditions. The following tables (Tables 9-14) will present more specific evidence of the Type I error rate for each set of conditions.

Table 9
All Average Type I Error Rates ($\tau^2 = 0$, $\delta = 0$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.042	**0.049**	0.035	0.037	0.035
4/6	4/6	0.024	**0.050**	0.026	0.028	0.027
6/4	4/6	0.081	**0.050**	0.043	**0.048**	0.044
20/20	4/6	**0.050**	**0.050**	0.044	**0.049**	**0.045**
16/24	4/6	0.026	**0.050**	0.033	0.036	0.034
24/16	4/6	0.098	**0.050**	**0.054**	0.064	0.056
100/100	4/6	**0.049**	**0.051**	**0.045**	**0.050**	**0.047**
80/120	4/6	0.028	**0.049**	0.033	0.036	0.034
120/80	4/6	0.095	**0.050**	**0.054**	0.066	0.057
5/5	12/18	0.039	**0.050**	0.033	0.034	0.033
4/6	12/18	0.013	**0.049**	0.026	0.026	0.026
6/4	12/18	0.114	**0.051**	0.041	**0.046**	0.042
20/20	12/18	**0.050**	**0.049**	0.041	**0.046**	0.043
16/24	12/18	0.019	**0.052**	0.034	0.036	0.035
24/16	12/18	0.150	**0.050**	**0.049**	0.061	**0.052**
100/100	12/18	**0.048**	**0.050**	0.044	**0.050**	**0.047**
80/120	12/18	0.020	**0.051**	0.036	0.039	0.037
120/80	12/18	0.151	**0.049**	**0.052**	0.065	0.056

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	**0.054**	**0.050**	0.040	**0.046**	0.042
sk=1.00, kr=3.00	4/6	**0.051**	**0.049**	0.040	**0.045**	0.041
sk=2.00, kr=6.00	4/6	0.059	**0.051**	0.042	**0.048**	0.043
sk=0.00, kr=0.00	12/18	0.064	**0.050**	0.039	0.044	0.040
sk=1.00, kr=3.00	12/18	0.061	**0.050**	0.039	0.044	0.041
sk=2.00, kr=6.00	12/18	0.077	**0.051**	0.041	**0.047**	0.043

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.039	**0.050**	0.039	0.043	0.040
1/2	4/6	**0.049**	**0.050**	0.040	**0.045**	0.042
1/4	4/6	0.077	**0.050**	0.043	**0.050**	0.045
1/1	12/18	0.034	**0.050**	0.037	0.041	0.039
1/2	12/18	**0.054**	**0.051**	0.040	0.045	0.042
1/4	12/18	0.113	**0.050**	0.041	**0.049**	0.043

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

This table (Table 9) illustrates that permuted Q consistently maintained rejection rates around .05. Under this condition, the FE test produced the second greatest number of adequate Type I error rates (13/30). The RE and CR tests tended to exhibit conservative Type I error rates when τ^2=0.

Table 10
All Average Type I Error Rates ($\tau^2 = 0$, $\delta = .8$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.040	**0.050**	0.032	0.035	0.033
4/6	4/6	0.025	**0.050**	0.027	0.028	0.027
6/4	4/6	0.066	**0.050**	0.042	**0.045**	0.042
20/20	4/6	**0.045**	**0.049**	0.039	0.043	0.040
16/24	4/6	0.032	**0.049**	0.032	0.035	0.033
24/16	4/6	0.074	**0.051**	**0.048**	**0.055**	**0.049**
100/100	4/6	**0.049**	**0.049**	0.041	**0.047**	0.043
80/120	4/6	0.035	**0.052**	0.034	0.038	0.035
120/80	4/6	0.076	**0.049**	**0.049**	0.057	**0.050**
5/5	12/18	0.033	**0.050**	0.031	0.032	0.031
4/6	12/18	0.020	**0.050**	0.026	0.027	0.026
6/4	12/18	0.080	**0.050**	0.037	0.040	0.038
20/20	12/18	**0.049**	**0.050**	0.039	0.043	0.041
16/24	12/18	0.033	**0.050**	0.033	0.036	0.034
24/16	12/18	0.102	**0.049**	**0.046**	**0.055**	**0.049**
100/100	12/18	0.057	**0.052**	0.041	**0.048**	0.044
80/120	12/18	0.041	**0.049**	0.032	0.036	0.034
120/80	12/18	0.110	**0.048**	**0.046**	0.056	**0.049**

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.056	**0.050**	0.041	**0.046**	0.042
sk=1.00, kr=3.00	4/6	**0.046**	**0.050**	0.038	0.042	0.039
sk=2.00, kr=6.00	4/6	**0.046**	**0.049**	0.035	0.040	0.036
sk=0.00, kr=0.00	12/18	0.069	**0.050**	0.039	**0.045**	0.041
sk=1.00, kr=3.00	12/18	**0.049**	**0.049**	0.037	0.041	0.039
sk=2.00, kr=6.00	12/18	0.057	**0.050**	0.034	0.039	0.036

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.066	**0.050**	**0.046**	**0.052**	**0.047**
1/2	4/6	0.039	**0.050**	0.037	0.040	0.038
1/4	4/6	0.042	**0.050**	0.032	0.035	0.032
1/1	12/18	0.085	**0.049**	0.044	**0.051**	**0.046**
1/2	12/18	0.038	**0.049**	0.036	0.039	0.037
1/4	12/18	**0.052**	**0.051**	0.031	0.035	0.032

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

This table (Table 10) reflects a slight decline for the regular Q and FE tests (as compared to $\tau^2 = 0$, $\delta = 0$) in the number of conditions with average Type I error rates approximating nominal α. The FE test produced adequate Type I error rates in 9 of 30 conditions, as compared with the prior 13 of 30. With the introduction of $\delta = .8$, the RE and CR tests provided a few more instances of robustness. RE and CR tests demonstrated better robustness when sample sizes in the first group were greater than the second group.

Table 11
All Average Type I Error Rates ($\tau^2 = .33$, $\delta = 0$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.340	**0.049**	0.076	0.102	0.077
4/6	4/6	0.270	**0.050**	0.072	0.092	0.074
6/4	4/6	0.403	**0.050**	0.083	0.117	0.085
20/20	4/6	0.909	**0.050**	0.114	0.309	0.115
16/24	4/6	0.875	**0.051**	0.115	0.290	0.116
24/16	4/6	0.921	**0.050**	0.115	0.324	0.116
100/100	4/6	0.999	**0.049**	0.117	0.614	0.117
80/120	4/6	0.999	**0.049**	0.117	0.599	0.117
120/80	4/6	1.000	**0.049**	0.117	0.622	0.117
5/5	12/18	0.640	**0.049**	0.063	0.100	0.065
4/6	12/18	0.504	**0.050**	0.061	0.087	0.063
6/4	12/18	0.726	**0.048**	0.062	0.110	0.064
20/20	12/18	0.999	**0.049**	0.067	0.302	0.067
16/24	12/18	0.998	**0.049**	0.068	0.281	0.068
24/16	12/18	1.000	**0.049**	0.068	0.314	0.068
100/100	12/18	1.000	**0.052**	0.071	0.611	0.071
80/120	12/18	1.000	**0.050**	0.068	0.595	0.068
120/80	12/18	1.000	**0.050**	0.069	0.615	0.069

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.723	**0.049**	0.101	0.334	0.102
sk=1.00, kr=3.00	4/6	0.742	**0.049**	0.102	0.341	0.103
sk=2.00, kr=6.00	4/6	0.773	**0.050**	0.105	0.348	0.106
sk=0.00, kr=0.00	12/18	0.842	**0.050**	0.066	0.329	0.066
sk=1.00, kr=3.00	12/18	0.869	**0.050**	0.067	0.334	0.067
sk=2.00, kr=6.00	12/18	0.911	**0.049**	0.066	0.341	0.067

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.736	**0.049**	0.102	0.338	0.103
1/2	4/6	0.742	**0.050**	0.102	0.338	0.103
1/4	4/6	0.761	**0.050**	0.104	0.346	0.105
1/1	12/18	0.862	**0.049**	0.066	0.332	0.067
1/2	12/18	0.871	**0.049**	0.066	0.336	0.066
1/4	12/18	0.890	**0.050**	0.067	0.337	0.067

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Only permuted Q maintained adequate Type I error control (see Table 11). The other 4 tests exhibited inflated Type I error consistent across all conditions. By elevating the K to 30, there appeared to be a greater degree of robustness maintained for the RE and CR tests across all conditions. The RE and CR tests outperformed the regular Q and FE tests in terms of overall effectiveness. But the increase in τ^2 from 0 to .33 promoted Type I error inflation for all of the tests, except permuted Q.

Table 12
All Average Type I Error Rates ($\tau^2 = .33$, $\delta = .8$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.312	**0.051**	0.076	0.101	0.077
4/6	4/6	0.247	**0.049**	0.069	0.088	0.070
6/4	4/6	0.375	**0.050**	0.079	0.112	0.081
20/20	4/6	0.892	**0.052**	0.115	0.302	0.116
16/24	4/6	0.856	**0.050**	0.111	0.283	0.112
24/16	4/6	0.902	**0.048**	0.112	0.313	0.113
100/100	4/6	0.999	**0.050**	0.116	0.607	0.116
80/120	4/6	0.999	**0.049**	0.117	0.595	0.117
120/80	4/6	0.999	**0.048**	0.116	0.610	0.116
5/5	12/18	0.595	**0.050**	0.062	0.098	0.064
4/6	12/18	0.460	**0.049**	0.057	0.082	0.059
6/4	12/18	0.690	**0.050**	0.063	0.106	0.064
20/20	12/18	0.999	**0.048**	0.066	0.299	0.066
16/24	12/18	0.997	**0.049**	0.068	0.275	0.069
24/16	12/18	0.999	**0.051**	0.069	0.304	0.069
100/100	12/18	1.000	**0.048**	0.068	0.604	0.068
80/120	12/18	1.000	**0.049**	0.069	0.588	0.069
120/80	12/18	1.000	**0.049**	0.067	0.608	0.067

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.716	**0.049**	0.099	0.327	0.100
sk=1.00, kr=3.00	4/6	0.728	**0.049**	0.101	0.335	0.102
sk=2.00, kr=6.00	4/6	0.750	**0.050**	0.103	0.342	0.104
sk=0.00, kr=0.00	12/18	0.837	**0.049**	0.065	0.324	0.065
sk=1.00, kr=3.00	12/18	0.855	**0.050**	0.065	0.328	0.066
sk=2.00, kr=6.00	12/18	0.888	**0.050**	0.067	0.336	0.067

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.734	**0.051**	0.103	0.337	0.104
1/2	4/6	0.727	**0.049**	0.100	0.333	0.102
1/4	4/6	0.733	**0.049**	0.100	0.333	0.101
1/1	12/18	0.864	**0.049**	0.065	0.330	0.066
1/2	12/18	0.855	**0.050**	0.066	0.328	0.067
1/4	12/18	0.861	**0.049**	0.065	0.330	0.066

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

As the average Type I error rates indicate (Table 12), only permuted Q maintained adequate robustness. The rates associated with the population shapes and variances suggest small K tended to minimize robustness to a greater extent for the RE and CR tests. These tests, though still inflating Type I error, did not perform as poorly when K=30. Regular Q and FE tests demonstrated inflated Type I to a much greater extent, than the RE or CR tests. It should be noted that the population shape did not present any more of a significant challenge to the control of Type I error for any of the tests than did the variances.

Table 13
All Average Type I Error Rates ($\tau^2 = 1$, $\delta = 0$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.768	**0.049**	0.102	0.193	0.103
4/6	4/6	0.701	**0.050**	0.098	0.179	0.099
6/4	4/6	0.798	**0.050**	0.104	0.202	0.105
20/20	4/6	0.996	**0.052**	0.115	0.475	0.116
16/24	4/6	0.994	**0.049**	0.115	0.462	0.115
24/16	4/6	0.996	**0.052**	0.117	0.481	0.117
100/100	4/6	1.000	**0.049**	0.116	0.742	0.116
80/120	4/6	1.000	**0.051**	0.117	0.730	0.117
120/80	4/6	1.000	**0.048**	0.115	0.747	0.115
5/5	12/18	0.987	**0.049**	0.067	0.185	0.067
4/6	12/18	0.973	**0.050**	0.067	0.173	0.067
6/4	12/18	0.991	**0.050**	0.066	0.195	0.066
20/20	12/18	1.000	**0.051**	0.068	0.467	0.068
16/24	12/18	1.000	**0.051**	0.069	0.452	0.069
24/16	12/18	1.000	**0.050**	0.068	0.474	0.068
100/100	12/18	1.000	**0.048**	0.066	0.734	0.066
80/120	12/18	1.000	**0.048**	0.067	0.725	0.067
120/80	12/18	1.000	**0.050**	0.069	0.740	0.069

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.899	**0.050**	0.110	0.460	0.111
sk=1.00, kr=3.00	4/6	0.916	**0.049**	0.111	0.469	0.111
sk=2.00, kr=6.00	4/6	0.937	**0.051**	0.112	0.476	0.112
sk=0.00, kr=0.00	12/18	0.991	**0.050**	0.068	0.454	0.068
sk=1.00, kr=3.00	12/18	0.995	**0.050**	0.068	0.462	0.068
sk=2.00, kr=6.00	12/18	0.998	**0.049**	0.067	0.466	0.067

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.913	**0.050**	0.111	0.465	0.111
1/2	4/6	0.916	**0.050**	0.111	0.466	0.111
1/4	4/6	0.922	**0.050**	0.111	0.473	0.112
1/1	12/18	0.995	**0.050**	0.068	0.460	0.068
1/2	12/18	0.995	**0.050**	0.068	0.459	0.068
1/4	12/18	0.994	**0.049**	0.067	0.462	0.067

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The average Type I error rates continued to escalate as heterogeneity of effects increased from .33 to 1 (see Table 12 as compared to Table 13). Again, the 4 tests other than permuted Q tended to produce inflated Type I error. The regular Q and FE tests' Type I error rates inflated to a much greater extent than those of either the RE or CR tests. These patterns were evident across changes in population shape and variance ratios, regardless of normality and equal variance. Lastly, Type I error for the FE test increased dramatically when the N increased to 200.

Table 14
All Average Type I Error Control ($\tau^2 = 1$, $\delta = .8$)

Primary Study Sample Sizes	N of studies	Reg Q	PQ_b	RE	FE	CR
5/5	4/6	0.740	**0.050**	0.104	0.196	0.106
4/6	4/6	0.682	**0.050**	0.098	0.179	0.100
6/4	4/6	0.777	**0.050**	0.104	0.205	0.105
20/20	4/6	0.995	**0.047**	0.113	0.476	0.114
16/24	4/6	0.993	**0.050**	0.114	0.458	0.114
24/16	4/6	0.995	**0.051**	0.115	0.481	0.115
100/100	4/6	1.000	**0.048**	0.114	0.739	0.114
80/120	4/6	1.000	**0.049**	0.119	0.730	0.119
120/80	4/6	1.000	**0.051**	0.115	0.740	0.115
5/5	12/18	0.983	**0.051**	0.068	0.185	0.068
4/6	12/18	0.965	**0.052**	0.068	0.171	0.068
6/4	12/18	0.988	**0.050**	0.067	0.195	0.067
20/20	12/18	1.000	**0.051**	0.068	0.465	0.068
16/24	12/18	1.000	**0.051**	0.071	0.457	0.071
24/16	12/18	1.000	**0.050**	0.069	0.478	0.069
100/100	12/18	1.000	**0.049**	0.068	0.739	0.068
80/120	12/18	1.000	**0.050**	0.069	0.728	0.069
120/80	12/18	1.000	**0.051**	0.069	0.740	0.069

Population Shape	N of studies	Reg Q	PQ_b	RE	FE	CR
sk=0.00, kr=0.00	4/6	0.893	**0.050**	0.111	0.461	0.111
sk=1.00, kr=3.00	4/6	0.906	**0.050**	0.110	0.466	0.111
sk=2.00, kr=6.00	4/6	0.926	**0.050**	0.111	0.473	0.112
sk=0.00, kr=0.00	12/18	0.989	**0.051**	0.069	0.446	0.069
sk=1.00, kr=3.00	12/18	0.993	**0.050**	0.069	0.461	0.069
sk=2.00, kr=6.00	12/18	0.997	**0.051**	0.068	0.468	0.068

Population Variances	N of studies	Reg Q	PQ_b	RE	FE	CR
1/1	4/6	0.909	**0.049**	0.110	0.464	0.111
1/2	4/6	0.908	**0.049**	0.110	0.466	0.111
1/4	4/6	0.908	**0.051**	0.112	0.470	0.112
1/1	12/18	0.994	**0.050**	0.067	0.461	0.067
1/2	12/18	0.993	**0.051**	0.069	0.463	0.069
1/4	12/18	0.992	**0.051**	0.069	0.452	0.069

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Type I error rates for each of the tests did not change dramatically from the prior condition of $\tau^2=1$, $\delta=0$ (see Table 13 relative to Table 14 above). The same patterns of inflation arose when K increased from 10 to 30 when both population shape and population variances were controlled. The degree of inflation for each test remained fairly constant from the prior condition, as well.

Summary of Results Concerning Average Type I Error Rates

With the introduction of δ=.8 at τ^2=0, the RE, CR and FE tests began to provide more instances of adequate control of Type I error, particularly when N was 40 or greater and the first group had a larger N than the second group. The equal variances condition also seemed to have contributed to the robustness of these tests. Robustness of these tests continued to decline as heterogeneity of effects increased from .33 to 1. The regular Q and FE tests' Type I error rates inflated to a much greater extent than those of either the RE or CR tests. These patterns were evident across changes in population shape and variance ratios.

Only permuted Q maintained a consistent pattern of robustness across a wide variety of conditions. Regardless of the extent of the heterogeneity of effects, population variance ratios, sample size of the primary studies, the number of studies or population shape, error rates continued to fluctuate only slightly around the nominal alpha level, .05.

Type I Error Rate Estimates of Individual Simulated Conditions

The Type I error rate estimates of all true null conditions are now presented. These tables provide evidence of each test's performance under all of the specified conditions of the study. By presenting the individual error rates of each condition, a more detailed picture of the behavior of each of the tests can be examined more closely, so that general patterns highlighted by the prior presentations can be more fully elaborated.

Table 15

Type I Error Rate Estimates ($\tau^2 = 0$, $\delta=0$) at $\alpha=.05$ for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00	0.034	0.047	0.032	0.033	0.032
5/5	1/2	4/6	sk=0.00, kr=0.00	0.040	0.051	0.033	0.035	0.033
5/5	1/4	4/6	sk=0.00, kr=0.00	0.055	0.051	0.036	0.040	0.037
5/5	1/1	4/6	sk=1.00, kr=3.00	0.029	0.049	0.039	0.041	0.039
5/5	1/2	4/6	sk=1.00, kr=3.00	0.030	0.041	0.031	0.035	0.032
5/5	1/4	4/6	sk=1.00, kr=3.00	0.048	0.049	0.035	0.038	0.035
5/5	1/1	4/6	sk=2.00, kr=6.00	0.028	0.052	0.030	0.030	0.030
5/5	1/2	4/6	sk=2.00, kr=6.00	0.033	0.052	0.035	0.038	0.036
5/5	1/4	4/6	sk=2.00, kr=6.00	0.083	0.050	0.042	0.046	0.043
4/6	1/1	4/6	sk=0.00, kr=0.00	0.035	0.049	0.032	0.034	0.033
4/6	1/2	4/6	sk=0.00, kr=0.00	0.017	0.051	0.023	0.024	0.023
4/6	1/4	4/6	sk=0.00, kr=0.00	0.018	0.055	0.022	0.023	0.022
4/6	1/1	4/6	sk=1.00, kr=3.00	0.029	0.048	0.033	0.036	0.034
4/6	1/2	4/6	sk=1.00, kr=3.00	0.021	0.045	0.024	0.024	0.024
4/6	1/4	4/6	sk=1.00, kr=3.00	0.016	0.050	0.021	0.023	0.021
4/6	1/1	4/6	sk=2.00, kr=6.00	0.024	0.048	0.028	0.030	0.028
4/6	1/2	4/6	sk=2.00, kr=6.00	0.022	0.053	0.028	0.029	0.028
4/6	1/4	4/6	sk=2.00, kr=6.00	0.037	0.048	0.028	0.030	0.029
6/4	1/1	4/6	sk=0.00, kr=0.00	0.031	0.055	0.032	0.035	0.033
6/4	1/2	4/6	sk=0.00, kr=0.00	0.065	0.046	0.041	0.046	0.041
6/4	1/4	4/6	sk=0.00, kr=0.00	0.145	0.050	0.056	0.065	0.057
6/4	1/1	4/6	sk=1.00, kr=3.00	0.030	0.049	0.032	0.034	0.033
6/4	1/2	4/6	sk=1.00, kr=3.00	0.057	0.052	0.043	0.046	0.044
6/4	1/4	4/6	sk=1.00, kr=3.00	0.127	0.047	0.051	0.060	0.052
6/4	1/1	4/6	sk=2.00, kr=6.00	0.031	0.054	0.030	0.031	0.030
6/4	1/2	4/6	sk=2.00, kr=6.00	0.061	0.049	0.039	0.043	0.041
6/4	1/4	4/6	sk=2.00, kr=6.00	0.177	0.049	0.060	0.072	0.062
20/20	1/1	4/6	sk=0.00, kr=0.00	0.044	0.047	0.046	0.051	0.048
20/20	1/2	4/6	sk=0.00, kr=0.00	0.043	0.057	0.045	0.049	0.046
20/20	1/4	4/6	sk=0.00, kr=0.00	0.052	0.048	0.041	0.046	0.042
20/20	1/1	4/6	sk=1.00, kr=3.00	0.038	0.048	0.042	0.046	0.043
20/20	1/2	4/6	sk=1.00, kr=3.00	0.046	0.043	0.041	0.045	0.042
20/20	1/4	4/6	sk=1.00, kr=3.00	0.054	0.057	0.047	0.052	0.048
20/20	1/1	4/6	sk=2.00, kr=6.00	0.039	0.044	0.040	0.044	0.041
20/20	1/2	4/6	sk=2.00, kr=6.00	0.048	0.049	0.044	0.050	0.046
20/20	1/4	4/6	sk=2.00, kr=6.00	0.082	0.056	0.051	0.058	0.052
16/24	1/1	4/6	sk=0.00, kr=0.00	0.043	0.051	0.046	0.050	0.047
16/24	1/2	4/6	sk=0.00, kr=0.00	0.018	0.052	0.029	0.031	0.030
16/24	1/4	4/6	sk=0.00, kr=0.00	0.011	0.049	0.026	0.026	0.026
16/24	1/1	4/6	sk=1.00, kr=3.00	0.040	0.046	0.038	0.042	0.039
16/24	1/2	4/6	sk=1.00, kr=3.00	0.023	0.053	0.032	0.034	0.033
16/24	1/4	4/6	sk=1.00, kr=3.00	0.010	0.043	0.022	0.023	0.022
16/24	1/1	4/6	sk=2.00, kr=6.00	0.037	0.051	0.042	0.045	0.043
16/24	1/2	4/6	sk=2.00, kr=6.00	0.027	0.046	0.034	0.038	0.035
16/24	1/4	4/6	sk=2.00, kr=6.00	0.029	0.053	0.030	0.032	0.030

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

When N was less than 40, none of the tests except permuted Q, permitted adequate control of Type I error. Error rates tended to be conservative for the RE, FE and CR tests.

Table 15 (continued)

Type I Error Rate Estimates ($\tau^2 = 0$, $\delta=0$) at $\alpha=.05$ for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00	0.048	0.051	0.041	0.048	0.043
24/16	1/2	4/6	sk=0.00, kr=0.00	0.092	0.051	0.055	0.065	0.056
24/16	1/4	4/6	sk=0.00, kr=0.00	0.151	0.047	0.063	0.080	0.067
24/16	1/1	4/6	sk=1.00, kr=3.00	0.042	0.047	0.038	0.042	0.039
24/16	1/2	4/6	sk=1.00, kr=3.00	0.085	0.052	0.055	0.063	0.057
24/16	1/4	4/6	sk=1.00, kr=3.00	0.148	0.045	0.063	0.079	0.065
24/16	1/1	4/6	sk=2.00, kr=6.00	0.040	0.050	0.041	0.044	0.042
24/16	1/2	4/6	sk=2.00, kr=6.00	0.087	0.057	0.062	0.071	0.064
24/16	1/4	4/6	sk=2.00, kr=6.00	0.188	0.050	0.069	0.088	0.072
100/100	1/1	4/6	sk=0.00, kr=0.00	0.049	0.052	0.048	0.051	0.049
100/100	1/2	4/6	sk=0.00, kr=0.00	0.050	0.055	0.051	0.056	0.053
100/100	1/4	4/6	sk=0.00, kr=0.00	0.046	0.050	0.040	0.043	0.041
100/100	1/1	4/6	sk=1.00, kr=3.00	0.044	0.050	0.042	0.047	0.044
100/100	1/2	4/6	sk=1.00, kr=3.00	0.050	0.054	0.047	0.052	0.048
100/100	1/4	4/6	sk=1.00, kr=3.00	0.054	0.046	0.046	0.050	0.047
100/100	1/1	4/6	sk=2.00, kr=6.00	0.047	0.051	0.044	0.049	0.045
100/100	1/2	4/6	sk=2.00, kr=6.00	0.048	0.050	0.041	0.046	0.043
100/100	1/4	4/6	sk=2.00, kr=6.00	0.055	0.051	0.048	0.056	0.049
80/120	1/1	4/6	sk=0.00, kr=0.00	0.044	0.045	0.040	0.044	0.040
80/120	1/2	4/6	sk=0.00, kr=0.00	0.026	0.044	0.029	0.032	0.030
80/120	1/4	4/6	sk=0.00, kr=0.00	0.010	0.050	0.025	0.027	0.026
80/120	1/1	4/6	sk=1.00, kr=3.00	0.051	0.047	0.046	0.052	0.048
80/120	1/2	4/6	sk=1.00, kr=3.00	0.022	0.052	0.035	0.037	0.035
80/120	1/4	4/6	sk=1.00, kr=3.00	0.013	0.050	0.024	0.025	0.024
80/120	1/1	4/6	sk=2.00, kr=6.00	0.046	0.055	0.043	0.049	0.045
80/120	1/2	4/6	sk=2.00, kr=6.00	0.023	0.050	0.033	0.035	0.034
80/120	1/4	4/6	sk=2.00, kr=6.00	0.013	0.050	0.026	0.027	0.026
120/80	1/1	4/6	sk=0.00, kr=0.00	0.046	0.053	0.042	0.048	0.044
120/80	1/2	4/6	sk=0.00, kr=0.00	0.094	0.046	0.054	0.066	0.057
120/80	1/4	4/6	sk=0.00, kr=0.00	0.142	0.049	0.062	0.080	0.067
120/80	1/1	4/6	sk=1.00, kr=3.00	0.044	0.050	0.045	0.051	0.047
120/80	1/2	4/6	sk=1.00, kr=3.00	0.087	0.050	0.054	0.062	0.055
120/80	1/4	4/6	sk=1.00, kr=3.00	0.147	0.050	0.064	0.079	0.066
120/80	1/1	4/6	sk=2.00, kr=6.00	0.048	0.052	0.045	0.051	0.047
120/80	1/2	4/6	sk=2.00, kr=6.00	0.093	0.049	0.055	0.066	0.058
120/80	1/4	4/6	sk=2.00, kr=6.00	0.158	0.049	0.066	0.087	0.070

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 16

Type I Error Rate Estimates ($\tau^2 = 0$, $\delta=.8$) at $\alpha=.05$ for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00	0.032	0.051	0.031	0.034	0.032
5/5	1/2	4/6	sk=0.00, kr=0.00	0.040	0.052	0.033	0.036	0.034
5/5	1/4	4/6	sk=0.00, kr=0.00	0.066	0.049	0.037	0.040	0.037
5/5	1/1	4/6	sk=1.00, kr=3.00	0.046	0.052	0.040	0.043	0.041
5/5	1/2	4/6	sk=1.00, kr=3.00	0.031	0.047	0.024	0.025	0.024
5/5	1/4	4/6	sk=1.00, kr=3.00	0.037	0.050	0.027	0.030	0.028
5/5	1/1	4/6	sk=2.00, kr=6.00	0.072	0.050	0.044	0.048	0.046
5/5	1/2	4/6	sk=2.00, kr=6.00	0.026	0.052	0.028	0.030	0.029
5/5	1/4	4/6	sk=2.00, kr=6.00	0.015	0.048	0.024	0.025	0.024
4/6	1/1	4/6	sk=0.00, kr=0.00	0.031	0.050	0.033	0.035	0.033
4/6	1/2	4/6	sk=0.00, kr=0.00	0.021	0.047	0.025	0.026	0.025
4/6	1/4	4/6	sk=0.00, kr=0.00	0.014	0.046	0.024	0.024	0.024
4/6	1/1	4/6	sk=1.00, kr=3.00	0.043	0.049	0.036	0.038	0.037
4/6	1/2	4/6	sk=1.00, kr=3.00	0.016	0.054	0.026	0.027	0.026
4/6	1/4	4/6	sk=1.00, kr=3.00	0.011	0.050	0.018	0.018	0.018
4/6	1/1	4/6	sk=2.00, kr=6.00	0.068	0.053	0.048	0.052	0.049
4/6	1/2	4/6	sk=2.00, kr=6.00	0.016	0.050	0.022	0.024	0.022
4/6	1/4	4/6	sk=2.00, kr=6.00	0.007	0.050	0.010	0.010	0.010
6/4	1/1	4/6	sk=0.00, kr=0.00	0.030	0.048	0.033	0.034	0.033
6/4	1/2	4/6	sk=0.00, kr=0.00	0.077	0.053	0.046	0.049	0.047
6/4	1/4	4/6	sk=0.00, kr=0.00	0.141	0.051	0.056	0.065	0.057
6/4	1/1	4/6	sk=1.00, kr=3.00	0.041	0.053	0.039	0.041	0.039
6/4	1/2	4/6	sk=1.00, kr=3.00	0.053	0.048	0.040	0.043	0.041
6/4	1/4	4/6	sk=1.00, kr=3.00	0.096	0.046	0.045	0.052	0.046
6/4	1/1	4/6	sk=2.00, kr=6.00	0.068	0.050	0.045	0.048	0.045
6/4	1/2	4/6	sk=2.00, kr=6.00	0.047	0.049	0.037	0.039	0.038
6/4	1/4	4/6	sk=2.00, kr=6.00	0.044	0.054	0.033	0.036	0.034
20/20	1/1	4/6	sk=0.00, kr=0.00	0.044	0.051	0.044	0.048	0.045
20/20	1/2	4/6	sk=0.00, kr=0.00	0.043	0.055	0.049	0.052	0.049
20/20	1/4	4/6	sk=0.00, kr=0.00	0.058	0.048	0.044	0.048	0.045
20/20	1/1	4/6	sk=1.00, kr=3.00	0.071	0.047	0.048	0.054	0.050
20/20	1/2	4/6	sk=1.00, kr=3.00	0.031	0.048	0.036	0.040	0.037
20/20	1/4	4/6	sk=1.00, kr=3.00	0.023	0.054	0.031	0.033	0.032
20/20	1/1	4/6	sk=2.00, kr=6.00	0.099	0.046	0.052	0.062	0.054
20/20	1/2	4/6	sk=2.00, kr=6.00	0.026	0.042	0.028	0.033	0.029
20/20	1/4	4/6	sk=2.00, kr=6.00	0.007	0.049	0.018	0.018	0.018
16/24	1/1	4/6	sk=0.00, kr=0.00	0.042	0.053	0.040	0.044	0.041
16/24	1/2	4/6	sk=0.00, kr=0.00	0.028	0.044	0.029	0.032	0.030
16/24	1/4	4/6	sk=0.00, kr=0.00	0.015	0.049	0.027	0.028	0.027
16/24	1/1	4/6	sk=1.00, kr=3.00	0.066	0.052	0.053	0.060	0.055
16/24	1/2	4/6	sk=1.00, kr=3.00	0.019	0.047	0.028	0.031	0.029
16/24	1/4	4/6	sk=1.00, kr=3.00	0.006	0.051	0.020	0.020	0.020
16/24	1/1	4/6	sk=2.00, kr=6.00	0.099	0.048	0.055	0.065	0.057
16/24	1/2	4/6	sk=2.00, kr=6.00	0.014	0.048	0.027	0.028	0.027
16/24	1/4	4/6	sk=2.00, kr=6.00	0.002	0.049	0.008	0.009	0.009

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The regular Q test's robustness succumbed to the influence of the effect size greater than zero, as did the FE test to a lesser extent (see Table 16 above). The RE and CR tests evidenced a greater degree of robustness than in the prior table (Table 15), when effect size was held to zero.

Table 16 (continued)

Type I Error Rate Estimates ($\tau^2 = 0$, δ=.8) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00	0.045	0.050	0.040	0.046	0.042
24/16	1/2	4/6	sk=0.00, kr=0.00	0.093	0.055	0.055	0.064	0.056
24/16	1/4	4/6	sk=0.00, kr=0.00	0.154	0.051	0.065	0.081	0.068
24/16	1/1	4/6	sk=1.00, kr=3.00	0.060	0.049	0.046	0.052	0.048
24/16	1/2	4/6	sk=1.00, kr=3.00	0.066	0.044	0.046	0.053	0.047
24/16	1/4	4/6	sk=1.00, kr=3.00	0.081	0.050	0.047	0.055	0.048
24/16	1/1	4/6	sk=2.00, kr=6.00	0.099	0.049	0.054	0.063	0.057
24/16	1/2	4/6	sk=2.00, kr=6.00	0.045	0.052	0.043	0.047	0.045
24/16	1/4	4/6	sk=2.00, kr=6.00	0.026	0.054	0.031	0.033	0.032
100/100	1/1	4/6	sk=0.00, kr=0.00	0.048	0.048	0.037	0.042	0.038
100/100	1/2	4/6	sk=0.00, kr=0.00	0.051	0.052	0.046	0.053	0.048
100/100	1/4	4/6	sk=0.00, kr=0.00	0.057	0.053	0.049	0.056	0.051
100/100	1/1	4/6	sk=1.00, kr=3.00	0.083	0.044	0.049	0.058	0.051
100/100	1/2	4/6	sk=1.00, kr=3.00	0.038	0.055	0.042	0.045	0.043
100/100	1/4	4/6	sk=1.00, kr=3.00	0.020	0.051	0.033	0.035	0.033
100/100	1/1	4/6	sk=2.00, kr=6.00	0.119	0.050	0.062	0.077	0.065
100/100	1/2	4/6	sk=2.00, kr=6.00	0.020	0.047	0.036	0.038	0.036
100/100	1/4	4/6	sk=2.00, kr=6.00	0.004	0.044	0.016	0.018	0.017
80/120	1/1	4/6	sk=0.00, kr=0.00	0.046	0.056	0.046	0.050	0.046
80/120	1/2	4/6	sk=0.00, kr=0.00	0.024	0.053	0.039	0.042	0.040
80/120	1/4	4/6	sk=0.00, kr=0.00	0.013	0.049	0.025	0.026	0.026
80/120	1/1	4/6	sk=1.00, kr=3.00	0.081	0.049	0.048	0.056	0.050
80/120	1/2	4/6	sk=1.00, kr=3.00	0.018	0.052	0.031	0.034	0.032
80/120	1/4	4/6	sk=1.00, kr=3.00	0.003	0.052	0.018	0.019	0.018
80/120	1/1	4/6	sk=2.00, kr=6.00	0.117	0.050	0.061	0.073	0.063
80/120	1/2	4/6	sk=2.00, kr=6.00	0.011	0.050	0.030	0.031	0.031
80/120	1/4	4/6	sk=2.00, kr=6.00	0.001	0.053	0.010	0.010	0.010
120/80	1/1	4/6	sk=0.00, kr=0.00	0.046	0.050	0.044	0.050	0.046
120/80	1/2	4/6	sk=0.00, kr=0.00	0.094	0.048	0.054	0.065	0.056
120/80	1/4	4/6	sk=0.00, kr=0.00	0.156	0.050	0.064	0.077	0.065
120/80	1/1	4/6	sk=1.00, kr=3.00	0.068	0.049	0.052	0.061	0.053
120/80	1/2	4/6	sk=1.00, kr=3.00	0.071	0.046	0.046	0.054	0.048
120/80	1/4	4/6	sk=1.00, kr=3.00	0.064	0.050	0.047	0.053	0.049
120/80	1/1	4/6	sk=2.00, kr=6.00	0.120	0.051	0.062	0.076	0.065
120/80	1/2	4/6	sk=2.00, kr=6.00	0.045	0.047	0.041	0.045	0.041
120/80	1/4	4/6	sk=2.00, kr=6.00	0.017	0.046	0.028	0.029	0.028

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The RE and CR tests held Type I error in check on a greater number of conditions when the sample size in the first group was greater than in the second (see Table 16). This pattern of response was more consistent when sample sizes increased to 40 or above. The FE test responded most effectively (in terms of robustness) when sample sizes were equal across groups. Permuted Q maintained robustness across all conditions, regardless of the sample size of either and both groups combined.

Table 17

Type I Error Rate Estimates ($\tau^2 = 0$, $\delta=0$) at $\alpha=.05$ for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00	0.023	0.049	0.028	0.029	0.028
5/5	1/2	12/18	sk=0.00, kr=0.00	0.030	0.052	0.035	0.036	0.036
5/5	1/4	12/18	sk=0.00, kr=0.00	0.052	0.051	0.035	0.037	0.036
5/5	1/1	12/18	sk=1.00, kr=3.00	0.018	0.047	0.028	0.030	0.029
5/5	1/2	12/18	sk=1.00, kr=3.00	0.023	0.050	0.032	0.033	0.032
5/5	1/4	12/18	sk=1.00, kr=3.00	0.048	0.049	0.031	0.032	0.031
5/5	1/1	12/18	sk=2.00, kr=6.00	0.017	0.054	0.029	0.029	0.029
5/5	1/2	12/18	sk=2.00, kr=6.00	0.032	0.046	0.038	0.040	0.039
5/5	1/4	12/18	sk=2.00, kr=6.00	0.110	0.050	0.041	0.044	0.042
4/6	1/1	12/18	sk=0.00, kr=0.00	0.021	0.049	0.030	0.030	0.030
4/6	1/2	12/18	sk=0.00, kr=0.00	0.008	0.048	0.023	0.023	0.023
4/6	1/4	12/18	sk=0.00, kr=0.00	0.004	0.049	0.021	0.021	0.021
4/6	1/1	12/18	sk=1.00, kr=3.00	0.018	0.045	0.025	0.025	0.025
4/6	1/2	12/18	sk=1.00, kr=3.00	0.005	0.050	0.028	0.028	0.028
4/6	1/4	12/18	sk=1.00, kr=3.00	0.003	0.051	0.022	0.023	0.023
4/6	1/1	12/18	sk=2.00, kr=6.00	0.017	0.053	0.032	0.033	0.033
4/6	1/2	12/18	sk=2.00, kr=6.00	0.013	0.048	0.023	0.024	0.024
4/6	1/4	12/18	sk=2.00, kr=6.00	0.029	0.051	0.029	0.030	0.029
6/4	1/1	12/18	sk=0.00, kr=0.00	0.022	0.045	0.024	0.024	0.024
6/4	1/2	12/18	sk=0.00, kr=0.00	0.075	0.052	0.040	0.044	0.042
6/4	1/4	12/18	sk=0.00, kr=0.00	0.230	0.055	0.053	0.062	0.055
6/4	1/1	12/18	sk=1.00, kr=3.00	0.018	0.053	0.031	0.031	0.031
6/4	1/2	12/18	sk=1.00, kr=3.00	0.071	0.051	0.040	0.042	0.041
6/4	1/4	12/18	sk=1.00, kr=3.00	0.198	0.047	0.046	0.053	0.048
6/4	1/1	12/18	sk=2.00, kr=6.00	0.013	0.051	0.031	0.031	0.031
6/4	1/2	12/18	sk=2.00, kr=6.00	0.073	0.061	0.053	0.056	0.054
6/4	1/4	12/18	sk=2.00, kr=6.00	0.320	0.045	0.054	0.067	0.055
20/20	1/1	12/18	sk=0.00, kr=0.00	0.042	0.047	0.036	0.042	0.040
20/20	1/2	12/18	sk=0.00, kr=0.00	0.041	0.045	0.039	0.041	0.040
20/20	1/4	12/18	sk=0.00, kr=0.00	0.047	0.047	0.040	0.044	0.041
20/20	1/1	12/18	sk=1.00, kr=3.00	0.036	0.045	0.036	0.041	0.039
20/20	1/2	12/18	sk=1.00, kr=3.00	0.042	0.049	0.043	0.046	0.044
20/20	1/4	12/18	sk=1.00, kr=3.00	0.057	0.054	0.045	0.051	0.048
20/20	1/1	12/18	sk=2.00, kr=6.00	0.030	0.048	0.039	0.042	0.041
20/20	1/2	12/18	sk=2.00, kr=6.00	0.054	0.053	0.044	0.049	0.046
20/20	1/4	12/18	sk=2.00, kr=6.00	0.098	0.050	0.047	0.055	0.049
16/24	1/1	12/18	sk=0.00, kr=0.00	0.039	0.057	0.045	0.048	0.046
16/24	1/2	12/18	sk=0.00, kr=0.00	0.010	0.050	0.030	0.030	0.030
16/24	1/4	12/18	sk=0.00, kr=0.00	0.004	0.050	0.027	0.028	0.027
16/24	1/1	12/18	sk=1.00, kr=3.00	0.037	0.048	0.041	0.044	0.043
16/24	1/2	12/18	sk=1.00, kr=3.00	0.009	0.055	0.035	0.036	0.036
16/24	1/4	12/18	sk=1.00, kr=3.00	0.005	0.052	0.025	0.025	0.025
16/24	1/1	12/18	sk=2.00, kr=6.00	0.040	0.047	0.041	0.044	0.043
16/24	1/2	12/18	sk=2.00, kr=6.00	0.017	0.055	0.038	0.039	0.038
16/24	1/4	12/18	sk=2.00, kr=6.00	0.011	0.049	0.029	0.030	0.030

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

None of the tests, save permuted Q, showed consistent robustness when N was less than 40. All of the tests, except regular Q and permuted Q, tended to yield conservative Type I error rates under this set of conditions. This pattern was evident across variance ratios and population shapes.

Table 17 (continued)
Type I Error Rate Estimates ($\tau^2 = 0$, $\delta=0$) at $\alpha=.05$ for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	0.042	0.046	0.038	0.042	0.040
24/16	1/2	12/18	sk=0.00, kr=0.00	0.123	0.046	0.047	0.056	0.050
24/16	1/4	12/18	sk=0.00, kr=0.00	0.266	0.046	0.054	0.077	0.059
24/16	1/1	12/18	sk=1.00, kr=3.00	0.039	0.051	0.041	0.044	0.042
24/16	1/2	12/18	sk=1.00, kr=3.00	0.113	0.059	0.058	0.067	0.062
24/16	1/4	12/18	sk=1.00, kr=3.00	0.266	0.048	0.052	0.074	0.057
24/16	1/1	12/18	sk=2.00, kr=6.00	0.037	0.051	0.040	0.044	0.041
24/16	1/2	12/18	sk=2.00, kr=6.00	0.126	0.050	0.051	0.061	0.053
24/16	1/4	12/18	sk=2.00, kr=6.00	0.338	0.051	0.060	0.085	0.066
100/100	1/1	12/18	sk=0.00, kr=0.00	0.042	0.052	0.044	0.051	0.048
100/100	1/2	12/18	sk=0.00, kr=0.00	0.047	0.056	0.044	0.050	0.046
100/100	1/4	12/18	sk=0.00, kr=0.00	0.045	0.047	0.042	0.047	0.045
100/100	1/1	12/18	sk=1.00, kr=3.00	0.047	0.048	0.041	0.047	0.043
100/100	1/2	12/18	sk=1.00, kr=3.00	0.048	0.049	0.047	0.053	0.051
100/100	1/4	12/18	sk=1.00, kr=3.00	0.050	0.054	0.047	0.053	0.050
100/100	1/1	12/18	sk=2.00, kr=6.00	0.044	0.050	0.043	0.049	0.046
100/100	1/2	12/18	sk=2.00, kr=6.00	0.053	0.048	0.042	0.048	0.045
100/100	1/4	12/18	sk=2.00, kr=6.00	0.056	0.047	0.042	0.048	0.045
80/120	1/1	12/18	sk=0.00, kr=0.00	0.047	0.054	0.049	0.054	0.050
80/120	1/2	12/18	sk=0.00, kr=0.00	0.011	0.053	0.035	0.038	0.037
80/120	1/4	12/18	sk=0.00, kr=0.00	0.003	0.050	0.026	0.026	0.026
80/120	1/1	12/18	sk=1.00, kr=3.00	0.045	0.052	0.045	0.050	0.047
80/120	1/2	12/18	sk=1.00, kr=3.00	0.013	0.053	0.035	0.036	0.035
80/120	1/4	12/18	sk=1.00, kr=3.00	0.002	0.046	0.027	0.028	0.028
80/120	1/1	12/18	sk=2.00, kr=6.00	0.045	0.050	0.045	0.051	0.047
80/120	1/2	12/18	sk=2.00, kr=6.00	0.010	0.049	0.033	0.035	0.034
80/120	1/4	12/18	sk=2.00, kr=6.00	0.005	0.054	0.030	0.031	0.030
120/80	1/1	12/18	sk=0.00, kr=0.00	0.043	0.053	0.047	0.051	0.049
120/80	1/2	12/18	sk=0.00, kr=0.00	0.137	0.043	0.047	0.060	0.052
120/80	1/4	12/18	sk=0.00, kr=0.00	0.263	0.052	0.060	0.083	0.064
120/80	1/1	12/18	sk=1.00, kr=3.00	0.043	0.046	0.038	0.044	0.042
120/80	1/2	12/18	sk=1.00, kr=3.00	0.131	0.048	0.053	0.067	0.059
120/80	1/4	12/18	sk=1.00, kr=3.00	0.259	0.052	0.063	0.083	0.066
120/80	1/1	12/18	sk=2.00, kr=6.00	0.048	0.044	0.041	0.046	0.043
120/80	1/2	12/18	sk=2.00, kr=6.00	0.147	0.055	0.055	0.067	0.060
120/80	1/4	12/18	sk=2.00, kr=6.00	0.285	0.051	0.062	0.086	0.068

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The regular Q, FE and CR tests produced a concentration of well-maintained Type I error conditions as the sample size of the primary studies increased to 200 under the equal groups condition. As unequal sample sizes were introduced, both of these tests had diminished Type I error control. The RE test was less robust than either of these other two tests when sample sizes increased to 200. The RE test demonstrated particular robustness when the sample sizes at N=10, 40 and 200 had the first group with the larger N than the second.

Table 18

Type I Error Rate Estimates ($\tau^2 = 0$, δ=.8) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00	0.027	0.053	0.030	0.031	0.031
5/5	1/2	12/18	sk=0.00, kr=0.00	0.035	0.050	0.031	0.033	0.032
5/5	1/4	12/18	sk=0.00, kr=0.00	0.059	0.054	0.038	0.040	0.039
5/5	1/1	12/18	sk=1.00, kr=3.00	0.035	0.044	0.031	0.031	0.031
5/5	1/2	12/18	sk=1.00, kr=3.00	0.019	0.051	0.031	0.033	0.032
5/5	1/4	12/18	sk=1.00, kr=3.00	0.021	0.047	0.027	0.028	0.027
5/5	1/1	12/18	sk=2.00, kr=6.00	0.083	0.051	0.043	0.047	0.044
5/5	1/2	12/18	sk=2.00, kr=6.00	0.015	0.051	0.027	0.027	0.027
5/5	1/4	12/18	sk=2.00, kr=6.00	0.005	0.049	0.019	0.020	0.019
4/6	1/1	12/18	sk=0.00, kr=0.00	0.027	0.051	0.033	0.034	0.033
4/6	1/2	12/18	sk=0.00, kr=0.00	0.012	0.046	0.021	0.022	0.021
4/6	1/4	12/18	sk=0.00, kr=0.00	0.008	0.051	0.022	0.022	0.022
4/6	1/1	12/18	sk=1.00, kr=3.00	0.038	0.057	0.040	0.041	0.040
4/6	1/2	12/18	sk=1.00, kr=3.00	0.006	0.051	0.027	0.028	0.027
4/6	1/4	12/18	sk=1.00, kr=3.00	0.003	0.049	0.020	0.020	0.020
4/6	1/1	12/18	sk=2.00, kr=6.00	0.079	0.048	0.040	0.043	0.041
4/6	1/2	12/18	sk=2.00, kr=6.00	0.008	0.048	0.022	0.022	0.022
4/6	1/4	12/18	sk=2.00, kr=6.00	0.001	0.047	0.012	0.012	0.012
6/4	1/1	12/18	sk=0.00, kr=0.00	0.022	0.048	0.027	0.027	0.027
6/4	1/2	12/18	sk=0.00, kr=0.00	0.082	0.045	0.034	0.038	0.035
6/4	1/4	12/18	sk=0.00, kr=0.00	0.244	0.052	0.051	0.061	0.053
6/4	1/1	12/18	sk=1.00, kr=3.00	0.033	0.055	0.035	0.036	0.035
6/4	1/2	12/18	sk=1.00, kr=3.00	0.053	0.047	0.036	0.039	0.037
6/4	1/4	12/18	sk=1.00, kr=3.00	0.119	0.050	0.042	0.045	0.043
6/4	1/1	12/18	sk=2.00, kr=6.00	0.096	0.049	0.041	0.044	0.042
6/4	1/2	12/18	sk=2.00, kr=6.00	0.041	0.049	0.034	0.035	0.034
6/4	1/4	12/18	sk=2.00, kr=6.00	0.034	0.053	0.033	0.034	0.034
20/20	1/1	12/18	sk=0.00, kr=0.00	0.043	0.051	0.044	0.049	0.046
20/20	1/2	12/18	sk=0.00, kr=0.00	0.046	0.052	0.043	0.046	0.045
20/20	1/4	12/18	sk=0.00, kr=0.00	0.061	0.056	0.049	0.054	0.051
20/20	1/1	12/18	sk=1.00, kr=3.00	0.077	0.048	0.047	0.055	0.051
20/20	1/2	12/18	sk=1.00, kr=3.00	0.027	0.049	0.038	0.041	0.039
20/20	1/4	12/18	sk=1.00, kr=3.00	0.013	0.048	0.030	0.031	0.030
20/20	1/1	12/18	sk=2.00, kr=6.00	0.153	0.051	0.055	0.068	0.058
20/20	1/2	12/18	sk=2.00, kr=6.00	0.017	0.046	0.027	0.029	0.028
20/20	1/4	12/18	sk=2.00, kr=6.00	0.001	0.050	0.018	0.019	0.019
16/24	1/1	12/18	sk=0.00, kr=0.00	0.042	0.047	0.041	0.046	0.043
16/24	1/2	12/18	sk=0.00, kr=0.00	0.013	0.048	0.035	0.037	0.037
16/24	1/4	12/18	sk=0.00, kr=0.00	0.004	0.049	0.028	0.029	0.029
16/24	1/1	12/18	sk=1.00, kr=3.00	0.073	0.050	0.049	0.055	0.051
16/24	1/2	12/18	sk=1.00, kr=3.00	0.008	0.044	0.026	0.027	0.026
16/24	1/4	12/18	sk=1.00, kr=3.00	0.001	0.052	0.019	0.019	0.019
16/24	1/1	12/18	sk=2.00, kr=6.00	0.149	0.048	0.054	0.066	0.058
16/24	1/2	12/18	sk=2.00, kr=6.00	0.004	0.053	0.029	0.030	0.030
16/24	1/4	12/18	sk=2.00, kr=6.00	0.000	0.055	0.012	0.012	0.012

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

As an effect size greater than 0 was introduced, the performance of the regular Q and FE test declined (see Table 18 as compared to Table 17). The regular Q had 14 conditions with adequate Type I error when the effect size was 0 (see Table 17) as compared to 8 in the present set of conditions. The FE test had less of a decline in performance from 21 adequate conditions vs. 17 in the present set.

Table 18 (continued)
Type I Error Rate Estimates ($\tau^2 = 0$, δ=.8) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	0.048	0.047	0.040	0.044	0.042
24/16	1/2	12/18	sk=0.00, kr=0.00	0.122	0.046	0.048	0.060	0.053
24/16	1/4	12/18	sk=0.00, kr=0.00	0.283	0.048	0.053	0.071	0.057
24/16	1/1	12/18	sk=1.00, kr=3.00	0.078	0.048	0.046	0.051	0.048
24/16	1/2	12/18	sk=1.00, kr=3.00	0.080	0.045	0.042	0.050	0.046
24/16	1/4	12/18	sk=1.00, kr=3.00	0.098	0.052	0.051	0.059	0.054
24/16	1/1	12/18	sk=2.00, kr=6.00	0.152	0.052	0.057	0.069	0.061
24/16	1/2	12/18	sk=2.00, kr=6.00	0.043	0.055	0.048	0.055	0.051
24/16	1/4	12/18	sk=2.00, kr=6.00	0.011	0.051	0.033	0.036	0.034
100/100	1/1	12/18	sk=0.00, kr=0.00	0.046	0.054	0.045	0.052	0.048
100/100	1/2	12/18	sk=0.00, kr=0.00	0.052	0.055	0.051	0.059	0.055
100/100	1/4	12/18	sk=0.00, kr=0.00	0.065	0.056	0.050	0.054	0.052
100/100	1/1	12/18	sk=1.00, kr=3.00	0.108	0.048	0.047	0.060	0.054
100/100	1/2	12/18	sk=1.00, kr=3.00	0.025	0.054	0.044	0.046	0.045
100/100	1/4	12/18	sk=1.00, kr=3.00	0.008	0.052	0.029	0.031	0.031
100/100	1/1	12/18	sk=2.00, kr=6.00	0.200	0.051	0.055	0.073	0.059
100/100	1/2	12/18	sk=2.00, kr=6.00	0.013	0.048	0.032	0.035	0.033
100/100	1/4	12/18	sk=2.00, kr=6.00	0.000	0.051	0.019	0.019	0.019
80/120	1/1	12/18	sk=0.00, kr=0.00	0.049	0.042	0.037	0.043	0.040
80/120	1/2	12/18	sk=0.00, kr=0.00	0.015	0.047	0.032	0.035	0.034
80/120	1/4	12/18	sk=0.00, kr=0.00	0.006	0.049	0.025	0.026	0.026
80/120	1/1	12/18	sk=1.00, kr=3.00	0.104	0.048	0.052	0.062	0.056
80/120	1/2	12/18	sk=1.00, kr=3.00	0.008	0.056	0.035	0.036	0.036
80/120	1/4	12/18	sk=1.00, kr=3.00	0.000	0.048	0.019	0.019	0.019
80/120	1/1	12/18	sk=2.00, kr=6.00	0.180	0.048	0.052	0.071	0.057
80/120	1/2	12/18	sk=2.00, kr=6.00	0.003	0.053	0.026	0.028	0.027
80/120	1/4	12/18	sk=2.00, kr=6.00	0.000	0.049	0.007	0.007	0.007
120/80	1/1	12/18	sk=0.00, kr=0.00	0.053	0.047	0.039	0.046	0.042
120/80	1/2	12/18	sk=0.00, kr=0.00	0.133	0.051	0.055	0.066	0.058
120/80	1/4	12/18	sk=0.00, kr=0.00	0.274	0.049	0.060	0.084	0.066
120/80	1/1	12/18	sk=1.00, kr=3.00	0.102	0.046	0.046	0.057	0.050
120/80	1/2	12/18	sk=1.00, kr=3.00	0.086	0.046	0.045	0.054	0.049
120/80	1/4	12/18	sk=1.00, kr=3.00	0.088	0.048	0.046	0.054	0.050
120/80	1/1	12/18	sk=2.00, kr=6.00	0.195	0.047	0.052	0.068	0.057
120/80	1/2	12/18	sk=2.00, kr=6.00	0.051	0.043	0.039	0.044	0.040
120/80	1/4	12/18	sk=2.00, kr=6.00	0.007	0.052	0.032	0.033	0.033

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The RE and CR tests generated more conservative Type I error rates than any of the other 3 tests, but the greatest frequency (other than permuted Q) of conditions with adequately controlled Type I error. The RE test performed best when the first group had the larger sample size.

Table 19

Type I Error Rate Estimates (τ^2 =.33, δ=0) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00	0.262	0.046	0.070	0.092	0.072
5/5	1/2	4/6	sk=0.00, kr=0.00	0.280	0.051	0.070	0.091	0.072
5/5	1/4	4/6	sk=0.00, kr=0.00	0.310	0.046	0.070	0.092	0.072
5/5	1/1	4/6	sk=1.00, kr=3.00	0.296	0.044	0.070	0.091	0.072
5/5	1/2	4/6	sk=1.00, kr=3.00	0.311	0.048	0.073	0.097	0.076
5/5	1/4	4/6	sk=1.00, kr=3.00	0.358	0.055	0.082	0.111	0.084
5/5	1/1	4/6	sk=2.00, kr=6.00	0.375	0.048	0.079	0.106	0.081
5/5	1/2	4/6	sk=2.00, kr=6.00	0.408	0.049	0.079	0.113	0.081
5/5	1/4	4/6	sk=2.00, kr=6.00	0.458	0.053	0.085	0.125	0.087
4/6	1/1	4/6	sk=0.00, kr=0.00	0.251	0.046	0.069	0.085	0.070
4/6	1/2	4/6	sk=0.00, kr=0.00	0.202	0.052	0.066	0.081	0.068
4/6	1/4	4/6	sk=0.00, kr=0.00	0.179	0.045	0.061	0.073	0.062
4/6	1/1	4/6	sk=1.00, kr=3.00	0.296	0.048	0.070	0.094	0.071
4/6	1/2	4/6	sk=1.00, kr=3.00	0.237	0.059	0.080	0.098	0.080
4/6	1/4	4/6	sk=1.00, kr=3.00	0.229	0.048	0.064	0.083	0.066
4/6	1/1	4/6	sk=2.00, kr=6.00	0.373	0.046	0.084	0.110	0.086
4/6	1/2	4/6	sk=2.00, kr=6.00	0.318	0.051	0.077	0.102	0.080
4/6	1/4	4/6	sk=2.00, kr=6.00	0.347	0.052	0.077	0.103	0.078
6/4	1/1	4/6	sk=0.00, kr=0.00	0.247	0.050	0.067	0.081	0.067
6/4	1/2	4/6	sk=0.00, kr=0.00	0.342	0.052	0.085	0.112	0.087
6/4	1/4	4/6	sk=0.00, kr=0.00	0.443	0.049	0.090	0.127	0.092
6/4	1/1	4/6	sk=1.00, kr=3.00	0.303	0.051	0.076	0.099	0.078
6/4	1/2	4/6	sk=1.00, kr=3.00	0.381	0.049	0.079	0.111	0.081
6/4	1/4	4/6	sk=1.00, kr=3.00	0.513	0.049	0.094	0.143	0.096
6/4	1/1	4/6	sk=2.00, kr=6.00	0.355	0.050	0.081	0.109	0.083
6/4	1/2	4/6	sk=2.00, kr=6.00	0.451	0.048	0.081	0.121	0.084
6/4	1/4	4/6	sk=2.00, kr=6.00	0.594	0.051	0.093	0.148	0.094
20/20	1/1	4/6	sk=0.00, kr=0.00	0.902	0.045	0.110	0.296	0.112
20/20	1/2	4/6	sk=0.00, kr=0.00	0.894	0.050	0.113	0.301	0.115
20/20	1/4	4/6	sk=0.00, kr=0.00	0.894	0.052	0.119	0.302	0.119
20/20	1/1	4/6	sk=1.00, kr=3.00	0.913	0.051	0.118	0.322	0.119
20/20	1/2	4/6	sk=1.00, kr=3.00	0.911	0.052	0.113	0.302	0.114
20/20	1/4	4/6	sk=1.00, kr=3.00	0.910	0.046	0.110	0.309	0.111
20/20	1/1	4/6	sk=2.00, kr=6.00	0.922	0.053	0.116	0.317	0.118
20/20	1/2	4/6	sk=2.00, kr=6.00	0.912	0.051	0.116	0.312	0.117
20/20	1/4	4/6	sk=2.00, kr=6.00	0.920	0.045	0.112	0.323	0.113
16/24	1/1	4/6	sk=0.00, kr=0.00	0.884	0.048	0.113	0.290	0.114
16/24	1/2	4/6	sk=0.00, kr=0.00	0.862	0.052	0.116	0.277	0.117
16/24	1/4	4/6	sk=0.00, kr=0.00	0.840	0.053	0.117	0.280	0.118
16/24	1/1	4/6	sk=1.00, kr=3.00	0.901	0.049	0.111	0.303	0.111
16/24	1/2	4/6	sk=1.00, kr=3.00	0.868	0.047	0.113	0.293	0.115
16/24	1/4	4/6	sk=1.00, kr=3.00	0.853	0.049	0.108	0.274	0.110
16/24	1/1	4/6	sk=2.00, kr=6.00	0.910	0.055	0.114	0.312	0.115
16/24	1/2	4/6	sk=2.00, kr=6.00	0.888	0.049	0.114	0.283	0.115
16/24	1/4	4/6	sk=2.00, kr=6.00	0.869	0.056	0.125	0.295	0.126

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

There was a dramatic increase in the number of conditions with inflated Type I error for all tests, except permuted Q. As sample size increased to 40, the Type I error for each of the 4 tests exhibited another notable increase. This pattern resulted in a total absence of error control by the aforementioned tests.

Table 19 (continued)
Type I Error Rate Estimates (τ^2 =.33, δ=0) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00	0.888	0.050	0.119	0.311	0.121
24/16	1/2	4/6	sk=0.00, kr=0.00	0.920	0.052	0.115	0.311	0.115
24/16	1/4	4/6	sk=0.00, kr=0.00	0.928	0.049	0.115	0.341	0.116
24/16	1/1	4/6	sk=1.00, kr=3.00	0.903	0.049	0.109	0.303	0.110
24/16	1/2	4/6	sk=1.00, kr=3.00	0.919	0.048	0.111	0.315	0.112
24/16	1/4	4/6	sk=1.00, kr=3.00	0.947	0.050	0.118	0.347	0.119
24/16	1/1	4/6	sk=2.00, kr=6.00	0.906	0.055	0.123	0.316	0.124
24/16	1/2	4/6	sk=2.00, kr=6.00	0.931	0.044	0.110	0.323	0.111
24/16	1/4	4/6	sk=2.00, kr=6.00	0.947	0.052	0.113	0.345	0.114
100/100	1/1	4/6	sk=0.00, kr=0.00	0.999	0.050	0.112	0.610	0.112
100/100	1/2	4/6	sk=0.00, kr=0.00	1.000	0.043	0.112	0.609	0.112
100/100	1/4	4/6	sk=0.00, kr=0.00	1.000	0.055	0.113	0.609	0.113
100/100	1/1	4/6	sk=1.00, kr=3.00	0.999	0.050	0.118	0.620	0.118
100/100	1/2	4/6	sk=1.00, kr=3.00	0.999	0.046	0.120	0.615	0.120
100/100	1/4	4/6	sk=1.00, kr=3.00	0.999	0.052	0.120	0.613	0.120
100/100	1/1	4/6	sk=2.00, kr=6.00	0.999	0.046	0.121	0.618	0.122
100/100	1/2	4/6	sk=2.00, kr=6.00	0.999	0.050	0.114	0.615	0.114
100/100	1/4	4/6	sk=2.00, kr=6.00	1.000	0.050	0.118	0.619	0.118
80/120	1/1	4/6	sk=0.00, kr=0.00	0.999	0.048	0.118	0.605	0.118
80/120	1/2	4/6	sk=0.00, kr=0.00	0.999	0.048	0.118	0.591	0.118
80/120	1/4	4/6	sk=0.00, kr=0.00	0.999	0.048	0.119	0.587	0.119
80/120	1/1	4/6	sk=1.00, kr=3.00	0.999	0.053	0.120	0.614	0.120
80/120	1/2	4/6	sk=1.00, kr=3.00	1.000	0.047	0.108	0.594	0.108
80/120	1/4	4/6	sk=1.00, kr=3.00	0.999	0.048	0.115	0.598	0.115
80/120	1/1	4/6	sk=2.00, kr=6.00	0.999	0.048	0.114	0.605	0.114
80/120	1/2	4/6	sk=2.00, kr=6.00	0.999	0.049	0.120	0.596	0.120
80/120	1/4	4/6	sk=2.00, kr=6.00	0.999	0.053	0.119	0.597	0.119
120/80	1/1	4/6	sk=0.00, kr=0.00	0.999	0.046	0.116	0.613	0.116
120/80	1/2	4/6	sk=0.00, kr=0.00	1.000	0.051	0.116	0.617	0.116
120/80	1/4	4/6	sk=0.00, kr=0.00	1.000	0.051	0.116	0.629	0.116
120/80	1/1	4/6	sk=1.00, kr=3.00	0.999	0.049	0.120	0.606	0.120
120/80	1/2	4/6	sk=1.00, kr=3.00	1.000	0.050	0.119	0.621	0.119
120/80	1/4	4/6	sk=1.00, kr=3.00	1.000	0.047	0.117	0.636	0.117
120/80	1/1	4/6	sk=2.00, kr=6.00	0.999	0.049	0.118	0.598	0.118
120/80	1/2	4/6	sk=2.00, kr=6.00	1.000	0.050	0.122	0.637	0.122
120/80	1/4	4/6	sk=2.00, kr=6.00	1.000	0.047	0.111	0.637	0.111

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

As the sample size increased to 40 and greater, permuted Q continued to evidence adequate Type I error control across all conditions. All other tests did not maintain robustness under these conditions.

Table 20

Type I Error Rate Estimates (τ^2 =.33, δ=.8) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00	0.248	0.052	0.068	0.089	0.069
5/5	1/2	4/6	sk=0.00, kr=0.00	0.264	0.049	0.068	0.093	0.070
5/5	1/4	4/6	sk=0.00, kr=0.00	0.304	0.050	0.079	0.100	0.080
5/5	1/1	4/6	sk=1.00, kr=3.00	0.296	0.052	0.078	0.102	0.080
5/5	1/2	4/6	sk=1.00, kr=3.00	0.276	0.050	0.069	0.089	0.072
5/5	1/4	4/6	sk=1.00, kr=3.00	0.309	0.050	0.078	0.102	0.080
5/5	1/1	4/6	sk=2.00, kr=6.00	0.405	0.052	0.087	0.120	0.089
5/5	1/2	4/6	sk=2.00, kr=6.00	0.350	0.052	0.077	0.107	0.078
5/5	1/4	4/6	sk=2.00, kr=6.00	0.353	0.049	0.077	0.109	0.078
4/6	1/1	4/6	sk=0.00, kr=0.00	0.238	0.047	0.067	0.089	0.070
4/6	1/2	4/6	sk=0.00, kr=0.00	0.209	0.050	0.062	0.077	0.063
4/6	1/4	4/6	sk=0.00, kr=0.00	0.178	0.047	0.058	0.072	0.058
4/6	1/1	4/6	sk=1.00, kr=3.00	0.289	0.051	0.078	0.102	0.080
4/6	1/2	4/6	sk=1.00, kr=3.00	0.223	0.048	0.065	0.079	0.067
4/6	1/4	4/6	sk=1.00, kr=3.00	0.189	0.057	0.072	0.088	0.075
4/6	1/1	4/6	sk=2.00, kr=6.00	0.385	0.050	0.083	0.115	0.086
4/6	1/2	4/6	sk=2.00, kr=6.00	0.281	0.043	0.066	0.088	0.068
4/6	1/4	4/6	sk=2.00, kr=6.00	0.226	0.049	0.065	0.083	0.067
6/4	1/1	4/6	sk=0.00, kr=0.00	0.241	0.053	0.068	0.089	0.070
6/4	1/2	4/6	sk=0.00, kr=0.00	0.329	0.049	0.076	0.103	0.079
6/4	1/4	4/6	sk=0.00, kr=0.00	0.452	0.047	0.079	0.119	0.081
6/4	1/1	4/6	sk=1.00, kr=3.00	0.299	0.047	0.073	0.098	0.076
6/4	1/2	4/6	sk=1.00, kr=3.00	0.358	0.055	0.084	0.115	0.086
6/4	1/4	4/6	sk=1.00, kr=3.00	0.440	0.048	0.081	0.120	0.083
6/4	1/1	4/6	sk=2.00, kr=6.00	0.389	0.053	0.084	0.117	0.086
6/4	1/2	4/6	sk=2.00, kr=6.00	0.407	0.046	0.085	0.123	0.088
6/4	1/4	4/6	sk=2.00, kr=6.00	0.455	0.049	0.083	0.123	0.084
20/20	1/1	4/6	sk=0.00, kr=0.00	0.881	0.057	0.122	0.296	0.124
20/20	1/2	4/6	sk=0.00, kr=0.00	0.884	0.048	0.110	0.280	0.111
20/20	1/4	4/6	sk=0.00, kr=0.00	0.883	0.054	0.115	0.292	0.116
20/20	1/1	4/6	sk=1.00, kr=3.00	0.906	0.050	0.119	0.319	0.120
20/20	1/2	4/6	sk=1.00, kr=3.00	0.892	0.048	0.111	0.301	0.111
20/20	1/4	4/6	sk=1.00, kr=3.00	0.890	0.052	0.106	0.300	0.106
20/20	1/1	4/6	sk=2.00, kr=6.00	0.917	0.054	0.122	0.321	0.122
20/20	1/2	4/6	sk=2.00, kr=6.00	0.890	0.052	0.117	0.308	0.117
20/20	1/4	4/6	sk=2.00, kr=6.00	0.881	0.054	0.116	0.301	0.116
16/24	1/1	4/6	sk=0.00, kr=0.00	0.872	0.049	0.115	0.293	0.116
16/24	1/2	4/6	sk=0.00, kr=0.00	0.846	0.048	0.103	0.265	0.105
16/24	1/4	4/6	sk=0.00, kr=0.00	0.813	0.049	0.108	0.261	0.111
16/24	1/1	4/6	sk=1.00, kr=3.00	0.891	0.054	0.123	0.309	0.124
16/24	1/2	4/6	sk=1.00, kr=3.00	0.847	0.048	0.111	0.273	0.113
16/24	1/4	4/6	sk=1.00, kr=3.00	0.836	0.042	0.106	0.265	0.108
16/24	1/1	4/6	sk=2.00, kr=6.00	0.905	0.055	0.118	0.318	0.119
16/24	1/2	4/6	sk=2.00, kr=6.00	0.863	0.053	0.111	0.293	0.112
16/24	1/4	4/6	sk=2.00, kr=6.00	0.827	0.050	0.102	0.268	0.104

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The same pattern of results continued (see Table 20) as displayed in the prior table (Table 19) and the permuted Q retained its robustness in terms of Type I error control. Again, none of the other tests maintained robustness under increasing heterogeneity of effects.

Table 20 (continued)
Type I Error Rate Estimates (τ^2 =.33, δ=.8) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00	0.872	0.048	0.114	0.283	0.115
24/16	1/2	4/6	sk=0.00, kr=0.00	0.900	0.045	0.109	0.307	0.110
24/16	1/4	4/6	sk=0.00, kr=0.00	0.917	0.047	0.117	0.335	0.118
24/16	1/1	4/6	sk=1.00, kr=3.00	0.887	0.050	0.107	0.293	0.108
24/16	1/2	4/6	sk=1.00, kr=3.00	0.905	0.045	0.108	0.311	0.110
24/16	1/4	4/6	sk=1.00, kr=3.00	0.916	0.049	0.112	0.328	0.113
24/16	1/1	4/6	sk=2.00, kr=6.00	0.899	0.048	0.109	0.308	0.110
24/16	1/2	4/6	sk=2.00, kr=6.00	0.910	0.051	0.117	0.321	0.118
24/16	1/4	4/6	sk=2.00, kr=6.00	0.917	0.049	0.110	0.330	0.111
100/100	1/1	4/6	sk=0.00, kr=0.00	0.999	0.047	0.114	0.603	0.114
100/100	1/2	4/6	sk=0.00, kr=0.00	0.999	0.053	0.127	0.608	0.127
100/100	1/4	4/6	sk=0.00, kr=0.00	0.999	0.047	0.113	0.599	0.113
100/100	1/1	4/6	sk=1.00, kr=3.00	0.998	0.049	0.109	0.609	0.109
100/100	1/2	4/6	sk=1.00, kr=3.00	1.000	0.051	0.118	0.611	0.118
100/100	1/4	4/6	sk=1.00, kr=3.00	1.000	0.047	0.116	0.602	0.116
100/100	1/1	4/6	sk=2.00, kr=6.00	0.998	0.054	0.118	0.607	0.118
100/100	1/2	4/6	sk=2.00, kr=6.00	0.999	0.050	0.113	0.618	0.113
100/100	1/4	4/6	sk=2.00, kr=6.00	1.000	0.049	0.112	0.603	0.112
80/120	1/1	4/6	sk=0.00, kr=0.00	0.999	0.050	0.117	0.599	0.117
80/120	1/2	4/6	sk=0.00, kr=0.00	0.999	0.043	0.116	0.572	0.116
80/120	1/4	4/6	sk=0.00, kr=0.00	0.999	0.050	0.116	0.587	0.116
80/120	1/1	4/6	sk=1.00, kr=3.00	0.999	0.051	0.120	0.611	0.120
80/120	1/2	4/6	sk=1.00, kr=3.00	1.000	0.044	0.109	0.599	0.109
80/120	1/4	4/6	sk=1.00, kr=3.00	0.999	0.046	0.110	0.583	0.110
80/120	1/1	4/6	sk=2.00, kr=6.00	0.999	0.049	0.122	0.618	0.123
80/120	1/2	4/6	sk=2.00, kr=6.00	0.999	0.054	0.119	0.599	0.119
80/120	1/4	4/6	sk=2.00, kr=6.00	0.999	0.050	0.118	0.584	0.118
120/80	1/1	4/6	sk=0.00, kr=0.00	1.000	0.048	0.111	0.592	0.111
120/80	1/2	4/6	sk=0.00, kr=0.00	0.999	0.052	0.119	0.617	0.119
120/80	1/4	4/6	sk=0.00, kr=0.00	1.000	0.046	0.110	0.612	0.110
120/80	1/1	4/6	sk=1.00, kr=3.00	0.999	0.050	0.117	0.596	0.117
120/80	1/2	4/6	sk=1.00, kr=3.00	0.999	0.049	0.118	0.612	0.118
120/80	1/4	4/6	sk=1.00, kr=3.00	0.999	0.049	0.118	0.614	0.118
120/80	1/1	4/6	sk=2.00, kr=6.00	1.000	0.044	0.113	0.608	0.113
120/80	1/2	4/6	sk=2.00, kr=6.00	0.999	0.051	0.123	0.626	0.123
120/80	1/4	4/6	sk=2.00, kr=6.00	1.000	0.045	0.117	0.617	0.117

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Permuted Q continued to maintain Type I error control across all conditions, as sample sizes increased. According to Bradley's criterion, all other tests did not maintain Type I error control, as reflected by the overly inflated Type I error rates. In fact, Type I error rates for all other tests became increasingly inflated as the primary sample sizes increased.

Table 21

Type I Error Rate Estimates (τ^2 =.33, δ=.8) at α=.10 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00	0.364	0.103	0.127	0.151	0.131
5/5	1/2	4/6	sk=0.00, kr=0.00	0.374	0.099	0.128	0.163	0.134
5/5	1/4	4/6	sk=0.00, kr=0.00	0.422	0.104	0.134	0.166	0.137
5/5	1/1	4/6	sk=1.00, kr=3.00	0.428	0.101	0.135	0.169	0.139
5/5	1/2	4/6	sk=1.00, kr=3.00	0.397	0.097	0.125	0.155	0.129
5/5	1/4	4/6	sk=1.00, kr=3.00	0.427	0.104	0.137	0.168	0.141
5/5	1/1	4/6	sk=2.00, kr=6.00	0.536	0.107	0.140	0.190	0.145
5/5	1/2	4/6	sk=2.00, kr=6.00	0.467	0.097	0.133	0.176	0.137
5/5	1/4	4/6	sk=2.00, kr=6.00	0.470	0.102	0.137	0.177	0.141
4/6	1/1	4/6	sk=0.00, kr=0.00	0.349	0.098	0.127	0.159	0.132
4/6	1/2	4/6	sk=0.00, kr=0.00	0.312	0.099	0.114	0.141	0.119
4/6	1/4	4/6	sk=0.00, kr=0.00	0.284	0.101	0.113	0.130	0.118
4/6	1/1	4/6	sk=1.00, kr=3.00	0.411	0.109	0.139	0.175	0.146
4/6	1/2	4/6	sk=1.00, kr=3.00	0.332	0.093	0.114	0.137	0.117
4/6	1/4	4/6	sk=1.00, kr=3.00	0.292	0.112	0.122	0.145	0.126
4/6	1/1	4/6	sk=2.00, kr=6.00	0.509	0.106	0.141	0.186	0.148
4/6	1/2	4/6	sk=2.00, kr=6.00	0.386	0.098	0.128	0.159	0.133
4/6	1/4	4/6	sk=2.00, kr=6.00	0.322	0.103	0.116	0.144	0.122
6/4	1/1	4/6	sk=0.00, kr=0.00	0.353	0.108	0.132	0.159	0.136
6/4	1/2	4/6	sk=0.00, kr=0.00	0.459	0.101	0.138	0.174	0.143
6/4	1/4	4/6	sk=0.00, kr=0.00	0.571	0.100	0.136	0.189	0.139
6/4	1/1	4/6	sk=1.00, kr=3.00	0.428	0.097	0.135	0.172	0.141
6/4	1/2	4/6	sk=1.00, kr=3.00	0.484	0.106	0.143	0.180	0.146
6/4	1/4	4/6	sk=1.00, kr=3.00	0.561	0.104	0.142	0.198	0.149
6/4	1/1	4/6	sk=2.00, kr=6.00	0.525	0.102	0.143	0.186	0.147
6/4	1/2	4/6	sk=2.00, kr=6.00	0.532	0.104	0.144	0.195	0.150
6/4	1/4	4/6	sk=2.00, kr=6.00	0.571	0.099	0.141	0.194	0.146
20/20	1/1	4/6	sk=0.00, kr=0.00	0.919	0.112	0.180	0.374	0.182
20/20	1/2	4/6	sk=0.00, kr=0.00	0.923	0.098	0.167	0.367	0.170
20/20	1/4	4/6	sk=0.00, kr=0.00	0.930	0.106	0.178	0.378	0.180
20/20	1/1	4/6	sk=1.00, kr=3.00	0.944	0.112	0.179	0.402	0.181
20/20	1/2	4/6	sk=1.00, kr=3.00	0.928	0.104	0.171	0.383	0.172
20/20	1/4	4/6	sk=1.00, kr=3.00	0.928	0.098	0.167	0.386	0.170
20/20	1/1	4/6	sk=2.00, kr=6.00	0.946	0.110	0.182	0.405	0.184
20/20	1/2	4/6	sk=2.00, kr=6.00	0.926	0.106	0.182	0.400	0.184
20/20	1/4	4/6	sk=2.00, kr=6.00	0.919	0.107	0.177	0.383	0.180
16/24	1/1	4/6	sk=0.00, kr=0.00	0.916	0.098	0.174	0.374	0.176
16/24	1/2	4/6	sk=0.00, kr=0.00	0.897	0.096	0.167	0.346	0.170
16/24	1/4	4/6	sk=0.00, kr=0.00	0.878	0.101	0.166	0.341	0.169
16/24	1/1	4/6	sk=1.00, kr=3.00	0.930	0.112	0.184	0.391	0.186
16/24	1/2	4/6	sk=1.00, kr=3.00	0.900	0.099	0.179	0.358	0.183
16/24	1/4	4/6	sk=1.00, kr=3.00	0.885	0.096	0.169	0.349	0.171
16/24	1/1	4/6	sk=2.00, kr=6.00	0.939	0.106	0.182	0.400	0.185
16/24	1/2	4/6	sk=2.00, kr=6.00	0.913	0.103	0.172	0.379	0.175
16/24	1/4	4/6	sk=2.00, kr=6.00	0.886	0.098	0.166	0.352	0.170

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 21 (continued)
Type I Error Rate Estimates (τ^2 =.33, δ=.8) at α=.10 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00	0.919	0.105	0.178	0.374	0.182
24/16	1/2	4/6	sk=0.00, kr=0.00	0.933	0.099	0.174	0.386	0.176
24/16	1/4	4/6	sk=0.00, kr=0.00	0.946	0.105	0.185	0.417	0.187
24/16	1/1	4/6	sk=1.00, kr=3.00	0.925	0.101	0.171	0.383	0.172
24/16	1/2	4/6	sk=1.00, kr=3.00	0.938	0.097	0.174	0.392	0.175
24/16	1/4	4/6	sk=1.00, kr=3.00	0.948	0.103	0.177	0.409	0.178
24/16	1/1	4/6	sk=2.00, kr=6.00	0.931	0.099	0.174	0.390	0.175
24/16	1/2	4/6	sk=2.00, kr=6.00	0.944	0.110	0.179	0.391	0.181
24/16	1/4	4/6	sk=2.00, kr=6.00	0.949	0.100	0.176	0.421	0.177
100/100	1/1	4/6	sk=0.00, kr=0.00	0.999	0.099	0.174	0.663	0.174
100/100	1/2	4/6	sk=0.00, kr=0.00	0.999	0.110	0.190	0.662	0.190
100/100	1/4	4/6	sk=0.00, kr=0.00	1.000	0.100	0.174	0.661	0.174
100/100	1/1	4/6	sk=1.00, kr=3.00	0.999	0.096	0.175	0.667	0.175
100/100	1/2	4/6	sk=1.00, kr=3.00	1.000	0.105	0.179	0.669	0.179
100/100	1/4	4/6	sk=1.00, kr=3.00	1.000	0.100	0.175	0.662	0.175
100/100	1/1	4/6	sk=2.00, kr=6.00	0.999	0.104	0.181	0.663	0.181
100/100	1/2	4/6	sk=2.00, kr=6.00	0.999	0.098	0.170	0.671	0.170
100/100	1/4	4/6	sk=2.00, kr=6.00	1.000	0.099	0.179	0.657	0.179
80/120	1/1	4/6	sk=0.00, kr=0.00	0.999	0.102	0.180	0.657	0.180
80/120	1/2	4/6	sk=0.00, kr=0.00	1.000	0.101	0.181	0.634	0.181
80/120	1/4	4/6	sk=0.00, kr=0.00	0.999	0.101	0.179	0.647	0.179
80/120	1/1	4/6	sk=1.00, kr=3.00	1.000	0.101	0.175	0.665	0.175
80/120	1/2	4/6	sk=1.00, kr=3.00	1.000	0.095	0.171	0.656	0.171
80/120	1/4	4/6	sk=1.00, kr=3.00	0.999	0.094	0.171	0.649	0.171
80/120	1/1	4/6	sk=2.00, kr=6.00	1.000	0.109	0.191	0.678	0.191
80/120	1/2	4/6	sk=2.00, kr=6.00	1.000	0.108	0.182	0.660	0.182
80/120	1/4	4/6	sk=2.00, kr=6.00	1.000	0.100	0.171	0.645	0.171
120/80	1/1	4/6	sk=0.00, kr=0.00	1.000	0.100	0.175	0.660	0.175
120/80	1/2	4/6	sk=0.00, kr=0.00	1.000	0.105	0.182	0.673	0.182
120/80	1/4	4/6	sk=0.00, kr=0.00	1.000	0.093	0.171	0.671	0.171
120/80	1/1	4/6	sk=1.00, kr=3.00	1.000	0.102	0.182	0.650	0.182
120/80	1/2	4/6	sk=1.00, kr=3.00	1.000	0.101	0.173	0.666	0.173
120/80	1/4	4/6	sk=1.00, kr=3.00	0.999	0.102	0.175	0.671	0.175
120/80	1/1	4/6	sk=2.00, kr=6.00	1.000	0.100	0.171	0.665	0.171
120/80	1/2	4/6	sk=2.00, kr=6.00	1.000	0.108	0.183	0.682	0.183
120/80	1/4	4/6	sk=2.00, kr=6.00	1.000	0.102	0.175	0.675	0.175

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Increasing the nominal alpha level from .05 to .10 (compare Tables 20 and 21) did not enhance the performance of any of the five tests being investigated, when K=10. The permuted Q still maintained adequate Type I error control. No other test maintained adequate Type I error control. When the primary study sample sizes were 40 or greater, both the regular Q and the FE tests still did not constrain Type I error. Therefore, when there was true equality between groups i.e. no difference, these tests incorrectly determined that a treatment had an effect.

Table 22

Type I Error Rate Estimates (τ^2 =.33, δ=0) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00	0.500	0.056	0.068	0.094	0.070
5/5	1/2	12/18	sk=0.00, kr=0.00	0.527	0.053	0.064	0.090	0.066
5/5	1/4	12/18	sk=0.00, kr=0.00	0.597	0.051	0.062	0.089	0.063
5/5	1/1	12/18	sk=1.00, kr=3.00	0.590	0.048	0.061	0.095	0.063
5/5	1/2	12/18	sk=1.00, kr=3.00	0.618	0.048	0.060	0.099	0.063
5/5	1/4	12/18	sk=1.00, kr=3.00	0.680	0.046	0.060	0.098	0.062
5/5	1/1	12/18	sk=2.00, kr=6.00	0.696	0.048	0.062	0.106	0.063
5/5	1/2	12/18	sk=2.00, kr=6.00	0.734	0.044	0.060	0.106	0.061
5/5	1/4	12/18	sk=2.00, kr=6.00	0.822	0.050	0.066	0.120	0.068
4/6	1/1	12/18	sk=0.00, kr=0.00	0.480	0.054	0.066	0.087	0.067
4/6	1/2	12/18	sk=0.00, kr=0.00	0.362	0.050	0.055	0.074	0.057
4/6	1/4	12/18	sk=0.00, kr=0.00	0.319	0.047	0.056	0.073	0.059
4/6	1/1	12/18	sk=1.00, kr=3.00	0.573	0.049	0.061	0.093	0.063
4/6	1/2	12/18	sk=1.00, kr=3.00	0.465	0.049	0.060	0.085	0.063
4/6	1/4	12/18	sk=1.00, kr=3.00	0.405	0.054	0.064	0.083	0.067
4/6	1/1	12/18	sk=2.00, kr=6.00	0.695	0.049	0.061	0.100	0.062
4/6	1/2	12/18	sk=2.00, kr=6.00	0.622	0.051	0.062	0.096	0.063
4/6	1/4	12/18	sk=2.00, kr=6.00	0.617	0.050	0.061	0.092	0.063
6/4	1/1	12/18	sk=0.00, kr=0.00	0.482	0.049	0.060	0.086	0.062
6/4	1/2	12/18	sk=0.00, kr=0.00	0.664	0.048	0.065	0.106	0.066
6/4	1/4	12/18	sk=0.00, kr=0.00	0.816	0.047	0.060	0.115	0.061
6/4	1/1	12/18	sk=1.00, kr=3.00	0.568	0.050	0.060	0.089	0.062
6/4	1/2	12/18	sk=1.00, kr=3.00	0.720	0.046	0.061	0.106	0.062
6/4	1/4	12/18	sk=1.00, kr=3.00	0.855	0.052	0.069	0.134	0.069
6/4	1/1	12/18	sk=2.00, kr=6.00	0.683	0.047	0.064	0.101	0.066
6/4	1/2	12/18	sk=2.00, kr=6.00	0.814	0.045	0.064	0.118	0.065
6/4	1/4	12/18	sk=2.00, kr=6.00	0.928	0.046	0.060	0.132	0.060
20/20	1/1	12/18	sk=0.00, kr=0.00	0.999	0.051	0.068	0.294	0.068
20/20	1/2	12/18	sk=0.00, kr=0.00	0.999	0.047	0.065	0.293	0.065
20/20	1/4	12/18	sk=0.00, kr=0.00	0.999	0.055	0.070	0.303	0.070
20/20	1/1	12/18	sk=1.00, kr=3.00	0.999	0.051	0.068	0.303	0.068
20/20	1/2	12/18	sk=1.00, kr=3.00	0.999	0.052	0.067	0.309	0.067
20/20	1/4	12/18	sk=1.00, kr=3.00	1.000	0.051	0.070	0.304	0.070
20/20	1/1	12/18	sk=2.00, kr=6.00	1.000	0.050	0.070	0.314	0.070
20/20	1/2	12/18	sk=2.00, kr=6.00	1.000	0.044	0.061	0.299	0.061
20/20	1/4	12/18	sk=2.00, kr=6.00	0.999	0.043	0.059	0.300	0.059
16/24	1/1	12/18	sk=0.00, kr=0.00	0.999	0.048	0.070	0.293	0.070
16/24	1/2	12/18	sk=0.00, kr=0.00	0.997	0.049	0.068	0.272	0.068
16/24	1/4	12/18	sk=0.00, kr=0.00	0.997	0.053	0.070	0.258	0.070
16/24	1/1	12/18	sk=1.00, kr=3.00	1.000	0.047	0.065	0.280	0.065
16/24	1/2	12/18	sk=1.00, kr=3.00	0.999	0.048	0.069	0.285	0.069
16/24	1/4	12/18	sk=1.00, kr=3.00	0.997	0.051	0.068	0.266	0.068
16/24	1/1	12/18	sk=2.00, kr=6.00	1.000	0.046	0.063	0.306	0.063
16/24	1/2	12/18	sk=2.00, kr=6.00	0.999	0.051	0.067	0.292	0.067
16/24	1/4	12/18	sk=2.00, kr=6.00	0.997	0.049	0.070	0.275	0.070

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Again, permuted Q maintained its robustness across most all conditions. All other tests resulted in inflated Type I error. Despite the increase in the total study sample, K, the RE and CR tests did not regain robustness.

Table 22 (continued)
Type I Error Rate Estimates (τ^2 =.33, δ=0) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	0.999	0.050	0.070	0.286	0.070
24/16	1/2	12/18	sk=0.00, kr=0.00	1.000	0.048	0.066	0.313	0.066
24/16	1/4	12/18	sk=0.00, kr=0.00	1.000	0.051	0.067	0.319	0.067
24/16	1/1	12/18	sk=1.00, kr=3.00	0.999	0.050	0.069	0.293	0.069
24/16	1/2	12/18	sk=1.00, kr=3.00	1.000	0.055	0.075	0.322	0.075
24/16	1/4	12/18	sk=1.00, kr=3.00	1.000	0.045	0.065	0.334	0.065
24/16	1/1	12/18	sk=2.00, kr=6.00	0.999	0.047	0.062	0.305	0.062
24/16	1/2	12/18	sk=2.00, kr=6.00	0.999	0.051	0.068	0.328	0.068
24/16	1/4	12/18	sk=2.00, kr=6.00	1.000	0.047	0.067	0.327	0.067
100/100	1/1	12/18	sk=0.00, kr=0.00	1.000	0.048	0.067	0.605	0.067
100/100	1/2	12/18	sk=0.00, kr=0.00	1.000	0.052	0.071	0.618	0.071
100/100	1/4	12/18	sk=0.00, kr=0.00	1.000	0.048	0.066	0.608	0.066
100/100	1/1	12/18	sk=1.00, kr=3.00	1.000	0.048	0.066	0.606	0.066
100/100	1/2	12/18	sk=1.00, kr=3.00	1.000	0.054	0.076	0.610	0.076
100/100	1/4	12/18	sk=1.00, kr=3.00	1.000	0.051	0.068	0.610	0.068
100/100	1/1	12/18	sk=2.00, kr=6.00	1.000	0.053	0.076	0.601	0.076
100/100	1/2	12/18	sk=2.00, kr=6.00	1.000	0.052	0.071	0.616	0.071
100/100	1/4	12/18	sk=2.00, kr=6.00	1.000	0.058	0.079	0.624	0.079
80/120	1/1	12/18	sk=0.00, kr=0.00	1.000	0.047	0.064	0.608	0.064
80/120	1/2	12/18	sk=0.00, kr=0.00	1.000	0.046	0.062	0.599	0.062
80/120	1/4	12/18	sk=0.00, kr=0.00	1.000	0.051	0.065	0.584	0.065
80/120	1/1	12/18	sk=1.00, kr=3.00	1.000	0.043	0.065	0.592	0.065
80/120	1/2	12/18	sk=1.00, kr=3.00	1.000	0.050	0.067	0.593	0.067
80/120	1/4	12/18	sk=1.00, kr=3.00	1.000	0.052	0.075	0.582	0.075
80/120	1/1	12/18	sk=2.00, kr=6.00	1.000	0.050	0.068	0.614	0.068
80/120	1/2	12/18	sk=2.00, kr=6.00	1.000	0.052	0.071	0.585	0.071
80/120	1/4	12/18	sk=2.00, kr=6.00	1.000	0.056	0.074	0.598	0.074
120/80	1/1	12/18	sk=0.00, kr=0.00	1.000	0.052	0.073	0.601	0.073
120/80	1/2	12/18	sk=0.00, kr=0.00	1.000	0.048	0.068	0.610	0.068
120/80	1/4	12/18	sk=0.00, kr=0.00	1.000	0.047	0.067	0.613	0.067
120/80	1/1	12/18	sk=1.00, kr=3.00	1.000	0.051	0.070	0.608	0.070
120/80	1/2	12/18	sk=1.00, kr=3.00	1.000	0.047	0.064	0.621	0.064
120/80	1/4	12/18	sk=1.00, kr=3.00	1.000	0.052	0.072	0.619	0.072
120/80	1/1	12/18	sk=2.00, kr=6.00	1.000	0.050	0.070	0.598	0.070
120/80	1/2	12/18	sk=2.00, kr=6.00	1.000	0.052	0.070	0.626	0.070
120/80	1/4	12/18	sk=2.00, kr=6.00	1.000	0.047	0.067	0.636	0.067

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Increasing the sample size to 40 and above did not facilitate enhanced robustness for the regular Q, RE, FE and CR tests.

Table 23

Type I Error Rate Estimates (τ^2 =.33, δ=.8) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00	0.482	0.052	0.063	0.090	0.065
5/5	1/2	12/18	sk=0.00, kr=0.00	0.515	0.056	0.068	0.092	0.070
5/5	1/4	12/18	sk=0.00, kr=0.00	0.589	0.051	0.064	0.095	0.066
5/5	1/1	12/18	sk=1.00, kr=3.00	0.595	0.047	0.060	0.094	0.062
5/5	1/2	12/18	sk=1.00, kr=3.00	0.558	0.049	0.061	0.093	0.064
5/5	1/4	12/18	sk=1.00, kr=3.00	0.582	0.048	0.059	0.099	0.061
5/5	1/1	12/18	sk=2.00, kr=6.00	0.741	0.048	0.062	0.113	0.063
5/5	1/2	12/18	sk=2.00, kr=6.00	0.652	0.053	0.065	0.105	0.066
5/5	1/4	12/18	sk=2.00, kr=6.00	0.639	0.047	0.059	0.099	0.062
4/6	1/1	12/18	sk=0.00, kr=0.00	0.463	0.053	0.061	0.088	0.063
4/6	1/2	12/18	sk=0.00, kr=0.00	0.363	0.046	0.057	0.073	0.059
4/6	1/4	12/18	sk=0.00, kr=0.00	0.317	0.050	0.052	0.068	0.055
4/6	1/1	12/18	sk=1.00, kr=3.00	0.572	0.051	0.060	0.091	0.061
4/6	1/2	12/18	sk=1.00, kr=3.00	0.419	0.052	0.058	0.081	0.060
4/6	1/4	12/18	sk=1.00, kr=3.00	0.337	0.050	0.054	0.070	0.055
4/6	1/1	12/18	sk=2.00, kr=6.00	0.728	0.048	0.061	0.105	0.062
4/6	1/2	12/18	sk=2.00, kr=6.00	0.521	0.050	0.063	0.090	0.066
4/6	1/4	12/18	sk=2.00, kr=6.00	0.417	0.044	0.050	0.074	0.053
6/4	1/1	12/18	sk=0.00, kr=0.00	0.466	0.046	0.056	0.081	0.058
6/4	1/2	12/18	sk=0.00, kr=0.00	0.624	0.047	0.059	0.094	0.060
6/4	1/4	12/18	sk=0.00, kr=0.00	0.796	0.046	0.060	0.115	0.061
6/4	1/1	12/18	sk=1.00, kr=3.00	0.573	0.055	0.068	0.099	0.069
6/4	1/2	12/18	sk=1.00, kr=3.00	0.686	0.050	0.062	0.102	0.064
6/4	1/4	12/18	sk=1.00, kr=3.00	0.786	0.050	0.067	0.126	0.068
6/4	1/1	12/18	sk=2.00, kr=6.00	0.729	0.046	0.059	0.101	0.060
6/4	1/2	12/18	sk=2.00, kr=6.00	0.749	0.052	0.066	0.111	0.067
6/4	1/4	12/18	sk=2.00, kr=6.00	0.798	0.055	0.071	0.126	0.073
20/20	1/1	12/18	sk=0.00, kr=0.00	0.999	0.048	0.066	0.295	0.066
20/20	1/2	12/18	sk=0.00, kr=0.00	0.999	0.044	0.064	0.295	0.064
20/20	1/4	12/18	sk=0.00, kr=0.00	0.999	0.044	0.063	0.288	0.063
20/20	1/1	12/18	sk=1.00, kr=3.00	1.000	0.049	0.070	0.302	0.070
20/20	1/2	12/18	sk=1.00, kr=3.00	0.999	0.049	0.066	0.291	0.066
20/20	1/4	12/18	sk=1.00, kr=3.00	0.999	0.047	0.067	0.289	0.067
20/20	1/1	12/18	sk=2.00, kr=6.00	1.000	0.043	0.060	0.312	0.060
20/20	1/2	12/18	sk=2.00, kr=6.00	1.000	0.058	0.073	0.308	0.073
20/20	1/4	12/18	sk=2.00, kr=6.00	0.999	0.051	0.067	0.307	0.067
16/24	1/1	12/18	sk=0.00, kr=0.00	0.996	0.047	0.066	0.279	0.066
16/24	1/2	12/18	sk=0.00, kr=0.00	0.996	0.049	0.069	0.268	0.069
16/24	1/4	12/18	sk=0.00, kr=0.00	0.996	0.049	0.068	0.260	0.069
16/24	1/1	12/18	sk=1.00, kr=3.00	1.000	0.049	0.069	0.285	0.069
16/24	1/2	12/18	sk=1.00, kr=3.00	0.996	0.046	0.065	0.270	0.065
16/24	1/4	12/18	sk=1.00, kr=3.00	0.995	0.051	0.069	0.257	0.069
16/24	1/1	12/18	sk=2.00, kr=6.00	1.000	0.051	0.068	0.314	0.068
16/24	1/2	12/18	sk=2.00, kr=6.00	0.998	0.049	0.067	0.274	0.067
16/24	1/4	12/18	sk=2.00, kr=6.00	0.996	0.053	0.074	0.266	0.074

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

At δ=.8, (see Table 23), permuted Q demonstrated continued robustness with the increase of K to 30 and optimal effectiveness upon the increase in sample size to 40. All other tests failed to provide adequate robustness under these conditions. The RE and CR tests showed minimal improvement to robustness with the increase in effect size to .8, particularly at sample sizes below 40.

Table 23 (continued)
Type I Error Rate Estimates (τ^2 =.33, δ=.8) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	0.997	0.050	0.070	0.278	0.070
24/16	1/2	12/18	sk=0.00, kr=0.00	0.999	0.049	0.067	0.295	0.067
24/16	1/4	12/18	sk=0.00, kr=0.00	0.999	0.044	0.062	0.316	0.062
24/16	1/1	12/18	sk=1.00, kr=3.00	0.999	0.051	0.068	0.290	0.068
24/16	1/2	12/18	sk=1.00, kr=3.00	0.999	0.048	0.064	0.303	0.064
24/16	1/4	12/18	sk=1.00, kr=3.00	1.000	0.055	0.071	0.327	0.071
24/16	1/1	12/18	sk=2.00, kr=6.00	1.000	0.054	0.071	0.304	0.071
24/16	1/2	12/18	sk=2.00, kr=6.00	1.000	0.053	0.072	0.301	0.072
24/16	1/4	12/18	sk=2.00, kr=6.00	1.000	0.055	0.075	0.325	0.075
100/100	1/1	12/18	sk=0.00, kr=0.00	1.000	0.048	0.070	0.592	0.070
100/100	1/2	12/18	sk=0.00, kr=0.00	1.000	0.049	0.072	0.603	0.072
100/100	1/4	12/18	sk=0.00, kr=0.00	1.000	0.050	0.065	0.601	0.065
100/100	1/1	12/18	sk=1.00, kr=3.00	1.000	0.052	0.070	0.600	0.070
100/100	1/2	12/18	sk=1.00, kr=3.00	1.000	0.044	0.063	0.604	0.063
100/100	1/4	12/18	sk=1.00, kr=3.00	1.000	0.047	0.067	0.600	0.067
100/100	1/1	12/18	sk=2.00, kr=6.00	1.000	0.046	0.065	0.622	0.065
100/100	1/2	12/18	sk=2.00, kr=6.00	1.000	0.049	0.068	0.604	0.068
100/100	1/4	12/18	sk=2.00, kr=6.00	1.000	0.046	0.069	0.611	0.069
80/120	1/1	12/18	sk=0.00, kr=0.00	1.000	0.046	0.067	0.589	0.067
80/120	1/2	12/18	sk=0.00, kr=0.00	1.000	0.053	0.075	0.590	0.075
80/120	1/4	12/18	sk=0.00, kr=0.00	1.000	0.049	0.068	0.582	0.068
80/120	1/1	12/18	sk=1.00, kr=3.00	1.000	0.042	0.061	0.597	0.061
80/120	1/2	12/18	sk=1.00, kr=3.00	1.000	0.051	0.070	0.580	0.070
80/120	1/4	12/18	sk=1.00, kr=3.00	1.000	0.049	0.070	0.577	0.070
80/120	1/1	12/18	sk=2.00, kr=6.00	1.000	0.052	0.072	0.606	0.072
80/120	1/2	12/18	sk=2.00, kr=6.00	1.000	0.049	0.071	0.591	0.071
80/120	1/4	12/18	sk=2.00, kr=6.00	1.000	0.050	0.070	0.581	0.070
120/80	1/1	12/18	sk=0.00, kr=0.00	1.000	0.050	0.068	0.596	0.068
120/80	1/2	12/18	sk=0.00, kr=0.00	1.000	0.048	0.064	0.607	0.064
120/80	1/4	12/18	sk=0.00, kr=0.00	1.000	0.046	0.062	0.618	0.062
120/80	1/1	12/18	sk=1.00, kr=3.00	1.000	0.048	0.065	0.598	0.065
120/80	1/2	12/18	sk=1.00, kr=3.00	1.000	0.051	0.071	0.612	0.071
120/80	1/4	12/18	sk=1.00, kr=3.00	1.000	0.054	0.068	0.618	0.068
120/80	1/1	12/18	sk=2.00, kr=6.00	1.000	0.049	0.068	0.600	0.068
120/80	1/2	12/18	sk=2.00, kr=6.00	1.000	0.047	0.066	0.609	0.066
120/80	1/4	12/18	sk=2.00, kr=6.00	1.000	0.049	0.066	0.614	0.066

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

When N was elevated to 40 and above, permuted Q evidenced adequate Type I error control. All other tests manifested an absence of robustness.

Table 24

Type I Error Rate Estimates (τ^2 =.33, δ=.8) at α=.10 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00	0.618	0.105	0.117	0.156	0.123
5/5	1/2	12/18	sk=0.00, kr=0.00	0.645	0.105	0.119	0.156	0.123
5/5	1/4	12/18	sk=0.00, kr=0.00	0.703	0.102	0.122	0.167	0.126
5/5	1/1	12/18	sk=1.00, kr=3.00	0.714	0.100	0.114	0.162	0.119
5/5	1/2	12/18	sk=1.00, kr=3.00	0.683	0.102	0.115	0.158	0.119
5/5	1/4	12/18	sk=1.00, kr=3.00	0.697	0.099	0.115	0.163	0.119
5/5	1/1	12/18	sk=2.00, kr=6.00	0.827	0.102	0.123	0.184	0.125
5/5	1/2	12/18	sk=2.00, kr=6.00	0.754	0.102	0.119	0.172	0.122
5/5	1/4	12/18	sk=2.00, kr=6.00	0.742	0.099	0.115	0.166	0.118
4/6	1/1	12/18	sk=0.00, kr=0.00	0.591	0.103	0.113	0.152	0.119
4/6	1/2	12/18	sk=0.00, kr=0.00	0.489	0.093	0.102	0.125	0.106
4/6	1/4	12/18	sk=0.00, kr=0.00	0.438	0.099	0.106	0.126	0.110
4/6	1/1	12/18	sk=1.00, kr=3.00	0.696	0.094	0.109	0.158	0.114
4/6	1/2	12/18	sk=1.00, kr=3.00	0.552	0.096	0.103	0.137	0.110
4/6	1/4	12/18	sk=1.00, kr=3.00	0.467	0.099	0.106	0.136	0.112
4/6	1/1	12/18	sk=2.00, kr=6.00	0.817	0.100	0.117	0.177	0.119
4/6	1/2	12/18	sk=2.00, kr=6.00	0.642	0.104	0.113	0.155	0.116
4/6	1/4	12/18	sk=2.00, kr=6.00	0.537	0.095	0.103	0.138	0.108
6/4	1/1	12/18	sk=0.00, kr=0.00	0.596	0.099	0.110	0.144	0.115
6/4	1/2	12/18	sk=0.00, kr=0.00	0.737	0.094	0.110	0.165	0.116
6/4	1/4	12/18	sk=0.00, kr=0.00	0.876	0.094	0.112	0.184	0.113
6/4	1/1	12/18	sk=1.00, kr=3.00	0.706	0.105	0.120	0.166	0.125
6/4	1/2	12/18	sk=1.00, kr=3.00	0.788	0.096	0.114	0.172	0.117
6/4	1/4	12/18	sk=1.00, kr=3.00	0.867	0.104	0.125	0.197	0.127
6/4	1/1	12/18	sk=2.00, kr=6.00	0.826	0.094	0.113	0.171	0.115
6/4	1/2	12/18	sk=2.00, kr=6.00	0.842	0.101	0.120	0.183	0.122
6/4	1/4	12/18	sk=2.00, kr=6.00	0.867	0.108	0.129	0.202	0.131
20/20	1/1	12/18	sk=0.00, kr=0.00	0.999	0.099	0.121	0.376	0.121
20/20	1/2	12/18	sk=0.00, kr=0.00	1.000	0.102	0.123	0.377	0.123
20/20	1/4	12/18	sk=0.00, kr=0.00	1.000	0.098	0.121	0.369	0.121
20/20	1/1	12/18	sk=1.00, kr=3.00	1.000	0.105	0.128	0.383	0.128
20/20	1/2	12/18	sk=1.00, kr=3.00	0.999	0.102	0.122	0.377	0.122
20/20	1/4	12/18	sk=1.00, kr=3.00	1.000	0.098	0.119	0.362	0.119
20/20	1/1	12/18	sk=2.00, kr=6.00	1.000	0.091	0.114	0.393	0.114
20/20	1/2	12/18	sk=2.00, kr=6.00	1.000	0.104	0.126	0.392	0.126
20/20	1/4	12/18	sk=2.00, kr=6.00	1.000	0.098	0.122	0.388	0.122
16/24	1/1	12/18	sk=0.00, kr=0.00	0.999	0.098	0.118	0.366	0.118
16/24	1/2	12/18	sk=0.00, kr=0.00	0.999	0.100	0.122	0.351	0.122
16/24	1/4	12/18	sk=0.00, kr=0.00	0.998	0.101	0.124	0.345	0.125
16/24	1/1	12/18	sk=1.00, kr=3.00	1.000	0.099	0.121	0.366	0.121
16/24	1/2	12/18	sk=1.00, kr=3.00	0.999	0.097	0.116	0.362	0.116
16/24	1/4	12/18	sk=1.00, kr=3.00	0.998	0.099	0.117	0.347	0.117
16/24	1/1	12/18	sk=2.00, kr=6.00	1.000	0.105	0.124	0.400	0.124
16/24	1/2	12/18	sk=2.00, kr=6.00	0.999	0.099	0.119	0.360	0.119
16/24	1/4	12/18	sk=2.00, kr=6.00	0.998	0.109	0.133	0.343	0.133

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Increasing the nominal alpha level from .05 to .10 (see Table 24 above) resulted in a minimal improvement in robustness for the RE and CR tests when K=30 and for sample sizes smaller than 40. For the RE and CR tests, Type I error was a central aspect supported by the increase in nominal alpha.

Table 24 (continued)
Type I Error Rate Estimates (τ^2 =.33, δ=.8) at α=.10 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	0.998	0.098	0.121	0.359	0.121
24/16	1/2	12/18	sk=0.00, kr=0.00	0.999	0.100	0.118	0.379	0.118
24/16	1/4	12/18	sk=0.00, kr=0.00	1.000	0.094	0.117	0.399	0.117
24/16	1/1	12/18	sk=1.00, kr=3.00	1.000	0.100	0.123	0.375	0.123
24/16	1/2	12/18	sk=1.00, kr=3.00	1.000	0.096	0.114	0.386	0.114
24/16	1/4	12/18	sk=1.00, kr=3.00	1.000	0.102	0.123	0.406	0.123
24/16	1/1	12/18	sk=2.00, kr=6.00	1.000	0.102	0.119	0.393	0.119
24/16	1/2	12/18	sk=2.00, kr=6.00	1.000	0.103	0.120	0.391	0.120
24/16	1/4	12/18	sk=2.00, kr=6.00	1.000	0.105	0.127	0.410	0.127
100/100	1/1	12/18	sk=0.00, kr=0.00	1.000	0.103	0.122	0.648	0.122
100/100	1/2	12/18	sk=0.00, kr=0.00	1.000	0.099	0.124	0.662	0.124
100/100	1/4	12/18	sk=0.00, kr=0.00	1.000	0.099	0.120	0.660	0.120
100/100	1/1	12/18	sk=1.00, kr=3.00	1.000	0.105	0.126	0.662	0.126
100/100	1/2	12/18	sk=1.00, kr=3.00	1.000	0.096	0.118	0.659	0.118
100/100	1/4	12/18	sk=1.00, kr=3.00	1.000	0.097	0.121	0.662	0.121
100/100	1/1	12/18	sk=2.00, kr=6.00	1.000	0.095	0.117	0.680	0.117
100/100	1/2	12/18	sk=2.00, kr=6.00	1.000	0.099	0.124	0.663	0.124
100/100	1/4	12/18	sk=2.00, kr=6.00	1.000	0.100	0.121	0.672	0.121
80/120	1/1	12/18	sk=0.00, kr=0.00	1.000	0.101	0.122	0.649	0.122
80/120	1/2	12/18	sk=0.00, kr=0.00	1.000	0.104	0.126	0.648	0.126
80/120	1/4	12/18	sk=0.00, kr=0.00	1.000	0.096	0.113	0.643	0.113
80/120	1/1	12/18	sk=1.00, kr=3.00	1.000	0.096	0.119	0.663	0.119
80/120	1/2	12/18	sk=1.00, kr=3.00	1.000	0.100	0.122	0.641	0.122
80/120	1/4	12/18	sk=1.00, kr=3.00	1.000	0.102	0.122	0.636	0.122
80/120	1/1	12/18	sk=2.00, kr=6.00	1.000	0.103	0.126	0.664	0.126
80/120	1/2	12/18	sk=2.00, kr=6.00	1.000	0.103	0.127	0.651	0.127
80/120	1/4	12/18	sk=2.00, kr=6.00	1.000	0.105	0.128	0.647	0.128
120/80	1/1	12/18	sk=0.00, kr=0.00	1.000	0.099	0.121	0.657	0.121
120/80	1/2	12/18	sk=0.00, kr=0.00	1.000	0.095	0.118	0.666	0.118
120/80	1/4	12/18	sk=0.00, kr=0.00	1.000	0.098	0.120	0.674	0.120
120/80	1/1	12/18	sk=1.00, kr=3.00	1.000	0.101	0.123	0.666	0.123
120/80	1/2	12/18	sk=1.00, kr=3.00	1.000	0.104	0.132	0.669	0.132
120/80	1/4	12/18	sk=1.00, kr=3.00	1.000	0.102	0.123	0.678	0.123
120/80	1/1	12/18	sk=2.00, kr=6.00	1.000	0.097	0.123	0.656	0.123
120/80	1/2	12/18	sk=2.00, kr=6.00	1.000	0.101	0.123	0.668	0.123
120/80	1/4	12/18	sk=2.00, kr=6.00	1.000	0.097	0.122	0.674	0.122

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded areas signify those conditions in which the error rate fell within Bradley's criterion.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

As the sample size increases (see Table 24 above), the RE and CR tests produced greater inflation of Type I error rates. Therefore, permuted Q is the only test effective under these conditions.

Table 25

Type I Error Rate Estimates (τ^2 =1, δ=0) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00	0.697	0.052	0.101	0.176	0.102
5/5	1/2	4/6	sk=0.00, kr=0.00	0.710	0.048	0.097	0.175	0.098
5/5	1/4	4/6	sk=0.00, kr=0.00	0.744	0.049	0.102	0.184	0.103
5/5	1/1	4/6	sk=1.00, kr=3.00	0.753	0.050	0.105	0.190	0.107
5/5	1/2	4/6	sk=1.00, kr=3.00	0.753	0.051	0.106	0.188	0.107
5/5	1/4	4/6	sk=1.00, kr=3.00	0.778	0.041	0.094	0.195	0.095
5/5	1/1	4/6	sk=2.00, kr=6.00	0.813	0.044	0.097	0.203	0.097
5/5	1/2	4/6	sk=2.00, kr=6.00	0.826	0.052	0.106	0.207	0.106
5/5	1/4	4/6	sk=2.00, kr=6.00	0.835	0.052	0.107	0.224	0.107
4/6	1/1	4/6	sk=0.00, kr=0.00	0.679	0.046	0.095	0.169	0.097
4/6	1/2	4/6	sk=0.00, kr=0.00	0.632	0.048	0.091	0.157	0.093
4/6	1/4	4/6	sk=0.00, kr=0.00	0.600	0.054	0.097	0.161	0.098
4/6	1/1	4/6	sk=1.00, kr=3.00	0.729	0.049	0.095	0.183	0.096
4/6	1/2	4/6	sk=1.00, kr=3.00	0.690	0.048	0.097	0.176	0.099
4/6	1/4	4/6	sk=1.00, kr=3.00	0.664	0.049	0.098	0.173	0.099
4/6	1/1	4/6	sk=2.00, kr=6.00	0.800	0.049	0.105	0.209	0.106
4/6	1/2	4/6	sk=2.00, kr=6.00	0.765	0.052	0.100	0.195	0.101
4/6	1/4	4/6	sk=2.00, kr=6.00	0.750	0.056	0.106	0.190	0.107
6/4	1/1	4/6	sk=0.00, kr=0.00	0.681	0.050	0.101	0.173	0.102
6/4	1/2	4/6	sk=0.00, kr=0.00	0.758	0.045	0.098	0.182	0.099
6/4	1/4	4/6	sk=0.00, kr=0.00	0.817	0.047	0.100	0.198	0.102
6/4	1/1	4/6	sk=1.00, kr=3.00	0.736	0.053	0.104	0.189	0.106
6/4	1/2	4/6	sk=1.00, kr=3.00	0.800	0.047	0.099	0.194	0.099
6/4	1/4	4/6	sk=1.00, kr=3.00	0.861	0.053	0.111	0.230	0.112
6/4	1/1	4/6	sk=2.00, kr=6.00	0.796	0.052	0.101	0.196	0.102
6/4	1/2	4/6	sk=2.00, kr=6.00	0.848	0.049	0.110	0.216	0.111
6/4	1/4	4/6	sk=2.00, kr=6.00	0.888	0.050	0.107	0.242	0.107
20/20	1/1	4/6	sk=0.00, kr=0.00	0.994	0.054	0.122	0.473	0.122
20/20	1/2	4/6	sk=0.00, kr=0.00	0.995	0.056	0.115	0.464	0.115
20/20	1/4	4/6	sk=0.00, kr=0.00	0.997	0.050	0.113	0.470	0.114
20/20	1/1	4/6	sk=1.00, kr=3.00	0.997	0.049	0.112	0.460	0.112
20/20	1/2	4/6	sk=1.00, kr=3.00	0.997	0.056	0.117	0.478	0.117
20/20	1/4	4/6	sk=1.00, kr=3.00	0.996	0.052	0.122	0.484	0.122
20/20	1/1	4/6	sk=2.00, kr=6.00	0.997	0.051	0.111	0.479	0.111
20/20	1/2	4/6	sk=2.00, kr=6.00	0.996	0.048	0.110	0.487	0.110
20/20	1/4	4/6	sk=2.00, kr=6.00	0.997	0.051	0.116	0.482	0.116
16/24	1/1	4/6	sk=0.00, kr=0.00	0.994	0.052	0.120	0.473	0.120
16/24	1/2	4/6	sk=0.00, kr=0.00	0.993	0.049	0.112	0.452	0.112
16/24	1/4	4/6	sk=0.00, kr=0.00	0.991	0.048	0.114	0.441	0.114
16/24	1/1	4/6	sk=1.00, kr=3.00	0.996	0.046	0.116	0.467	0.116
16/24	1/2	4/6	sk=1.00, kr=3.00	0.996	0.050	0.117	0.463	0.117
16/24	1/4	4/6	sk=1.00, kr=3.00	0.994	0.049	0.114	0.474	0.114
16/24	1/1	4/6	sk=2.00, kr=6.00	0.995	0.051	0.109	0.468	0.109
16/24	1/2	4/6	sk=2.00, kr=6.00	0.996	0.054	0.122	0.482	0.122
16/24	1/4	4/6	sk=2.00, kr=6.00	0.994	0.050	0.119	0.469	0.119

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

As heterogeneity of effects increased from .33 to 1, Type I error rates for all tests, but permuted Q, rose steadily, rendering these tests lacking in robustness.

Table 25 (continued)

Type I Error Rate Estimates (τ^2 =1, δ=0) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00	0.996	0.052	0.120	0.457	0.120
24/16	1/2	4/6	sk=0.00, kr=0.00	0.995	0.051	0.114	0.474	0.114
24/16	1/4	4/6	sk=0.00, kr=0.00	0.998	0.049	0.117	0.484	0.117
24/16	1/1	4/6	sk=1.00, kr=3.00	0.995	0.050	0.113	0.462	0.113
24/16	1/2	4/6	sk=1.00, kr=3.00	0.996	0.052	0.117	0.491	0.117
24/16	1/4	4/6	sk=1.00, kr=3.00	0.997	0.054	0.119	0.495	0.119
24/16	1/1	4/6	sk=2.00, kr=6.00	0.995	0.050	0.114	0.467	0.114
24/16	1/2	4/6	sk=2.00, kr=6.00	0.997	0.054	0.114	0.489	0.114
24/16	1/4	4/6	sk=2.00, kr=6.00	0.997	0.057	0.124	0.509	0.125
100/100	1/1	4/6	sk=0.00, kr=0.00	1.000	0.050	0.117	0.736	0.117
100/100	1/2	4/6	sk=0.00, kr=0.00	1.000	0.049	0.117	0.738	0.117
100/100	1/4	4/6	sk=0.00, kr=0.00	1.000	0.048	0.110	0.752	0.110
100/100	1/1	4/6	sk=1.00, kr=3.00	1.000	0.052	0.122	0.745	0.122
100/100	1/2	4/6	sk=1.00, kr=3.00	1.000	0.044	0.118	0.734	0.118
100/100	1/4	4/6	sk=1.00, kr=3.00	1.000	0.048	0.115	0.744	0.116
100/100	1/1	4/6	sk=2.00, kr=6.00	1.000	0.049	0.116	0.752	0.116
100/100	1/2	4/6	sk=2.00, kr=6.00	1.000	0.050	0.116	0.741	0.116
100/100	1/4	4/6	sk=2.00, kr=6.00	1.000	0.047	0.114	0.732	0.114
80/120	1/1	4/6	sk=0.00, kr=0.00	1.000	0.054	0.120	0.735	0.120
80/120	1/2	4/6	sk=0.00, kr=0.00	1.000	0.053	0.116	0.725	0.116
80/120	1/4	4/6	sk=0.00, kr=0.00	1.000	0.049	0.114	0.725	0.114
80/120	1/1	4/6	sk=1.00, kr=3.00	1.000	0.043	0.107	0.733	0.107
80/120	1/2	4/6	sk=1.00, kr=3.00	1.000	0.049	0.124	0.731	0.124
80/120	1/4	4/6	sk=1.00, kr=3.00	1.000	0.051	0.119	0.728	0.119
80/120	1/1	4/6	sk=2.00, kr=6.00	1.000	0.048	0.120	0.739	0.120
80/120	1/2	4/6	sk=2.00, kr=6.00	1.000	0.052	0.114	0.730	0.114
80/120	1/4	4/6	sk=2.00, kr=6.00	1.000	0.055	0.117	0.727	0.117
120/80	1/1	4/6	sk=0.00, kr=0.00	1.000	0.050	0.113	0.735	0.113
120/80	1/2	4/6	sk=0.00, kr=0.00	1.000	0.051	0.121	0.745	0.121
120/80	1/4	4/6	sk=0.00, kr=0.00	1.000	0.048	0.113	0.755	0.113
120/80	1/1	4/6	sk=1.00, kr=3.00	1.000	0.048	0.114	0.742	0.114
120/80	1/2	4/6	sk=1.00, kr=3.00	1.000	0.041	0.107	0.754	0.107
120/80	1/4	4/6	sk=1.00, kr=3.00	1.000	0.046	0.113	0.746	0.113
120/80	1/1	4/6	sk=2.00, kr=6.00	1.000	0.054	0.118	0.738	0.118
120/80	1/2	4/6	sk=2.00, kr=6.00	1.000	0.051	0.118	0.742	0.118
120/80	1/4	4/6	sk=2.00, kr=6.00	1.000	0.046	0.116	0.766	0.116

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

As N increased to 40 and above (see Table 25 above), all tests, but permuted Q, continued to show inflated

Type I error rates. Permuted Q maintained robustness at higher sample sizes with K=10.

Table 26

Type I Error Rate Estimates (τ^2=1, δ=.8) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00	0.670	0.048	0.097	0.175	0.098
5/5	1/2	4/6	sk=0.00, kr=0.00	0.692	0.044	0.099	0.182	0.101
5/5	1/4	4/6	sk=0.00, kr=0.00	0.712	0.059	0.115	0.203	0.117
5/5	1/1	4/6	sk=1.00, kr=3.00	0.735	0.050	0.104	0.190	0.104
5/5	1/2	4/6	sk=1.00, kr=3.00	0.737	0.052	0.101	0.194	0.102
5/5	1/4	4/6	sk=1.00, kr=3.00	0.756	0.051	0.105	0.198	0.106
5/5	1/1	4/6	sk=2.00, kr=6.00	0.796	0.048	0.106	0.207	0.107
5/5	1/2	4/6	sk=2.00, kr=6.00	0.784	0.049	0.108	0.208	0.109
5/5	1/4	4/6	sk=2.00, kr=6.00	0.776	0.052	0.105	0.207	0.106
4/6	1/1	4/6	sk=0.00, kr=0.00	0.665	0.048	0.099	0.181	0.101
4/6	1/2	4/6	sk=0.00, kr=0.00	0.615	0.048	0.095	0.161	0.097
4/6	1/4	4/6	sk=0.00, kr=0.00	0.595	0.053	0.097	0.160	0.099
4/6	1/1	4/6	sk=1.00, kr=3.00	0.728	0.052	0.099	0.185	0.101
4/6	1/2	4/6	sk=1.00, kr=3.00	0.669	0.050	0.098	0.177	0.100
4/6	1/4	4/6	sk=1.00, kr=3.00	0.638	0.051	0.098	0.169	0.099
4/6	1/1	4/6	sk=2.00, kr=6.00	0.785	0.052	0.103	0.197	0.105
4/6	1/2	4/6	sk=2.00, kr=6.00	0.741	0.048	0.094	0.184	0.095
4/6	1/4	4/6	sk=2.00, kr=6.00	0.702	0.051	0.100	0.195	0.100
6/4	1/1	4/6	sk=0.00, kr=0.00	0.667	0.048	0.100	0.177	0.102
6/4	1/2	4/6	sk=0.00, kr=0.00	0.736	0.050	0.102	0.190	0.103
6/4	1/4	4/6	sk=0.00, kr=0.00	0.800	0.056	0.113	0.222	0.115
6/4	1/1	4/6	sk=1.00, kr=3.00	0.735	0.049	0.103	0.190	0.105
6/4	1/2	4/6	sk=1.00, kr=3.00	0.773	0.054	0.108	0.209	0.110
6/4	1/4	4/6	sk=1.00, kr=3.00	0.824	0.051	0.106	0.221	0.107
6/4	1/1	4/6	sk=2.00, kr=6.00	0.794	0.050	0.105	0.207	0.106
6/4	1/2	4/6	sk=2.00, kr=6.00	0.816	0.049	0.100	0.208	0.101
6/4	1/4	4/6	sk=2.00, kr=6.00	0.846	0.046	0.094	0.221	0.095
20/20	1/1	4/6	sk=0.00, kr=0.00	0.996	0.045	0.109	0.466	0.109
20/20	1/2	4/6	sk=0.00, kr=0.00	0.995	0.043	0.112	0.465	0.112
20/20	1/4	4/6	sk=0.00, kr=0.00	0.996	0.043	0.112	0.473	0.112
20/20	1/1	4/6	sk=1.00, kr=3.00	0.996	0.040	0.105	0.474	0.105
20/20	1/2	4/6	sk=1.00, kr=3.00	0.994	0.043	0.110	0.485	0.110
20/20	1/4	4/6	sk=1.00, kr=3.00	0.995	0.053	0.118	0.483	0.118
20/20	1/1	4/6	sk=2.00, kr=6.00	0.995	0.051	0.120	0.481	0.120
20/20	1/2	4/6	sk=2.00, kr=6.00	0.995	0.052	0.117	0.481	0.117
20/20	1/4	4/6	sk=2.00, kr=6.00	0.996	0.056	0.118	0.480	0.118
16/24	1/1	4/6	sk=0.00, kr=0.00	0.994	0.051	0.112	0.460	0.112
16/24	1/2	4/6	sk=0.00, kr=0.00	0.992	0.050	0.118	0.456	0.118
16/24	1/4	4/6	sk=0.00, kr=0.00	0.991	0.049	0.113	0.435	0.113
16/24	1/1	4/6	sk=1.00, kr=3.00	0.995	0.051	0.113	0.478	0.113
16/24	1/2	4/6	sk=1.00, kr=3.00	0.991	0.052	0.114	0.446	0.115
16/24	1/4	4/6	sk=1.00, kr=3.00	0.990	0.057	0.112	0.454	0.112
16/24	1/1	4/6	sk=2.00, kr=6.00	0.996	0.044	0.108	0.471	0.109
16/24	1/2	4/6	sk=2.00, kr=6.00	0.994	0.049	0.114	0.456	0.115
16/24	1/4	4/6	sk=2.00, kr=6.00	0.991	0.052	0.121	0.464	0.121

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Though limited in overall effectiveness, permuted Q maintained adequate robustness over a majority of the conditions (see Table 26). All other tests did not maintain Type I error control, making those tests ineffective as well.

Table 26 (continued)
Type I Error Rate Estimates (τ^2 =1, δ=.8) at α=.05 for K=10, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00	0.995	0.048	0.110	0.466	0.110
24/16	1/2	4/6	sk=0.00, kr=0.00	0.994	0.055	0.116	0.475	0.116
24/16	1/4	4/6	sk=0.00, kr=0.00	0.997	0.048	0.119	0.490	0.119
24/16	1/1	4/6	sk=1.00, kr=3.00	0.995	0.053	0.119	0.468	0.119
24/16	1/2	4/6	sk=1.00, kr=3.00	0.995	0.046	0.108	0.475	0.108
24/16	1/4	4/6	sk=1.00, kr=3.00	0.996	0.048	0.113	0.485	0.114
24/16	1/1	4/6	sk=2.00, kr=6.00	0.995	0.053	0.121	0.480	0.121
24/16	1/2	4/6	sk=2.00, kr=6.00	0.995	0.053	0.118	0.494	0.118
24/16	1/4	4/6	sk=2.00, kr=6.00	0.997	0.049	0.113	0.503	0.113
100/100	1/1	4/6	sk=0.00, kr=0.00	1.000	0.048	0.121	0.738	0.121
100/100	1/2	4/6	sk=0.00, kr=0.00	1.000	0.049	0.114	0.741	0.114
100/100	1/4	4/6	sk=0.00, kr=0.00	1.000	0.048	0.112	0.745	0.112
100/100	1/1	4/6	sk=1.00, kr=3.00	1.000	0.048	0.113	0.743	0.113
100/100	1/2	4/6	sk=1.00, kr=3.00	1.000	0.051	0.118	0.738	0.118
100/100	1/4	4/6	sk=1.00, kr=3.00	1.000	0.046	0.115	0.736	0.115
100/100	1/1	4/6	sk=2.00, kr=6.00	1.000	0.046	0.107	0.727	0.107
100/100	1/2	4/6	sk=2.00, kr=6.00	1.000	0.050	0.110	0.738	0.110
100/100	1/4	4/6	sk=2.00, kr=6.00	1.000	0.048	0.117	0.743	0.117
80/120	1/1	4/6	sk=0.00, kr=0.00	1.000	0.053	0.117	0.721	0.117
80/120	1/2	4/6	sk=0.00, kr=0.00	1.000	0.049	0.118	0.727	0.118
80/120	1/4	4/6	sk=0.00, kr=0.00	1.000	0.049	0.125	0.730	0.125
80/120	1/1	4/6	sk=1.00, kr=3.00	1.000	0.050	0.121	0.730	0.121
80/120	1/2	4/6	sk=1.00, kr=3.00	1.000	0.048	0.116	0.736	0.116
80/120	1/4	4/6	sk=1.00, kr=3.00	1.000	0.046	0.117	0.723	0.117
80/120	1/1	4/6	sk=2.00, kr=6.00	1.000	0.046	0.117	0.733	0.117
80/120	1/2	4/6	sk=2.00, kr=6.00	1.000	0.048	0.118	0.733	0.118
80/120	1/4	4/6	sk=2.00, kr=6.00	1.000	0.053	0.119	0.731	0.119
120/80	1/1	4/6	sk=0.00, kr=0.00	1.000	0.049	0.111	0.728	0.111
120/80	1/2	4/6	sk=0.00, kr=0.00	1.000	0.053	0.114	0.743	0.114
120/80	1/4	4/6	sk=0.00, kr=0.00	1.000	0.055	0.123	0.749	0.123
120/80	1/1	4/6	sk=1.00, kr=3.00	1.000	0.052	0.113	0.732	0.113
120/80	1/2	4/6	sk=1.00, kr=3.00	1.000	0.053	0.115	0.736	0.115
120/80	1/4	4/6	sk=1.00, kr=3.00	1.000	0.050	0.111	0.741	0.111
120/80	1/1	4/6	sk=2.00, kr=6.00	1.000	0.054	0.119	0.735	0.119
120/80	1/2	4/6	sk=2.00, kr=6.00	1.000	0.050	0.121	0.747	0.121
120/80	1/4	4/6	sk=2.00, kr=6.00	1.000	0.045	0.109	0.752	0.109

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Permuted Q's robustness remained intact with increasing sample sizes (see Table 26 above). Therefore, permuted Q appeared to be robust to multiple violations of assumptions. All other tests projected inflated Type I error rates.

Table 27

Type I Error Rate Estimates (τ^2 =1, δ=0) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00	0.977	0.051	0.067	0.164	0.067
5/5	1/2	12/18	sk=0.00, kr=0.00	0.979	0.045	0.061	0.165	0.061
5/5	1/4	12/18	sk=0.00, kr=0.00	0.982	0.046	0.062	0.171	0.062
5/5	1/1	12/18	sk=1.00, kr=3.00	0.986	0.052	0.070	0.183	0.070
5/5	1/2	12/18	sk=1.00, kr=3.00	0.987	0.049	0.068	0.188	0.068
5/5	1/4	12/18	sk=1.00, kr=3.00	0.990	0.049	0.067	0.182	0.067
5/5	1/1	12/18	sk=2.00, kr=6.00	0.994	0.052	0.071	0.198	0.071
5/5	1/2	12/18	sk=2.00, kr=6.00	0.995	0.048	0.065	0.201	0.065
5/5	1/4	12/18	sk=2.00, kr=6.00	0.996	0.052	0.068	0.212	0.068
4/6	1/1	12/18	sk=0.00, kr=0.00	0.970	0.046	0.062	0.163	0.062
4/6	1/2	12/18	sk=0.00, kr=0.00	0.954	0.048	0.068	0.152	0.068
4/6	1/4	12/18	sk=0.00, kr=0.00	0.935	0.050	0.068	0.154	0.068
4/6	1/1	12/18	sk=1.00, kr=3.00	0.985	0.054	0.070	0.181	0.070
4/6	1/2	12/18	sk=1.00, kr=3.00	0.974	0.052	0.070	0.173	0.070
4/6	1/4	12/18	sk=1.00, kr=3.00	0.968	0.047	0.061	0.163	0.061
4/6	1/1	12/18	sk=2.00, kr=6.00	0.993	0.054	0.073	0.196	0.073
4/6	1/2	12/18	sk=2.00, kr=6.00	0.989	0.047	0.065	0.185	0.066
4/6	1/4	12/18	sk=2.00, kr=6.00	0.985	0.049	0.068	0.186	0.068
6/4	1/1	12/18	sk=0.00, kr=0.00	0.972	0.052	0.069	0.168	0.069
6/4	1/2	12/18	sk=0.00, kr=0.00	0.986	0.047	0.064	0.176	0.064
6/4	1/4	12/18	sk=0.00, kr=0.00	0.995	0.048	0.063	0.200	0.063
6/4	1/1	12/18	sk=1.00, kr=3.00	0.983	0.049	0.064	0.180	0.064
6/4	1/2	12/18	sk=1.00, kr=3.00	0.993	0.052	0.070	0.194	0.070
6/4	1/4	12/18	sk=1.00, kr=3.00	0.998	0.051	0.067	0.213	0.067
6/4	1/1	12/18	sk=2.00, kr=6.00	0.995	0.050	0.070	0.193	0.070
6/4	1/2	12/18	sk=2.00, kr=6.00	0.999	0.050	0.064	0.206	0.064
6/4	1/4	12/18	sk=2.00, kr=6.00	1.000	0.051	0.066	0.226	0.066
20/20	1/1	12/18	sk=0.00, kr=0.00	1.000	0.052	0.071	0.478	0.071
20/20	1/2	12/18	sk=0.00, kr=0.00	1.000	0.054	0.071	0.455	0.071
20/20	1/4	12/18	sk=0.00, kr=0.00	1.000	0.049	0.064	0.465	0.064
20/20	1/1	12/18	sk=1.00, kr=3.00	1.000	0.048	0.068	0.461	0.068
20/20	1/2	12/18	sk=1.00, kr=3.00	1.000	0.052	0.070	0.465	0.070
20/20	1/4	12/18	sk=1.00, kr=3.00	1.000	0.050	0.066	0.481	0.066
20/20	1/1	12/18	sk=2.00, kr=6.00	1.000	0.049	0.068	0.463	0.068
20/20	1/2	12/18	sk=2.00, kr=6.00	1.000	0.052	0.073	0.470	0.073
20/20	1/4	12/18	sk=2.00, kr=6.00	1.000	0.048	0.065	0.468	0.065
16/24	1/1	12/18	sk=0.00, kr=0.00	1.000	0.059	0.080	0.456	0.080
16/24	1/2	12/18	sk=0.00, kr=0.00	1.000	0.053	0.068	0.443	0.068
16/24	1/4	12/18	sk=0.00, kr=0.00	1.000	0.050	0.068	0.436	0.068
16/24	1/1	12/18	sk=1.00, kr=3.00	1.000	0.049	0.065	0.465	0.065
16/24	1/2	12/18	sk=1.00, kr=3.00	1.000	0.047	0.067	0.449	0.067
16/24	1/4	12/18	sk=1.00, kr=3.00	1.000	0.056	0.073	0.449	0.073
16/24	1/1	12/18	sk=2.00, kr=6.00	1.000	0.046	0.065	0.470	0.065
16/24	1/2	12/18	sk=2.00, kr=6.00	1.000	0.044	0.061	0.443	0.061
16/24	1/4	12/18	sk=2.00, kr=6.00	1.000	0.051	0.070	0.454	0.070

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Only permuted Q maintained adequate Type I error control under increasing heterogeneity of effects. Again, permuted Q's robustness was demonstrated across most all conditions under investigation. Although the RE and CR tests' error rates were not as inflated as those of regular Q and the FE test, the former two tests still presented error rates exceeding nominal alpha.

Table 27 (continued)
Type I Error Rate Estimates (τ^2 =1, δ=0) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	1.000	0.051	0.069	0.458	0.069
24/16	1/2	12/18	sk=0.00, kr=0.00	1.000	0.053	0.073	0.482	0.073
24/16	1/4	12/18	sk=0.00, kr=0.00	1.000	0.050	0.067	0.470	0.067
24/16	1/1	12/18	sk=1.00, kr=3.00	1.000	0.046	0.062	0.462	0.062
24/16	1/2	12/18	sk=1.00, kr=3.00	1.000	0.051	0.070	0.481	0.070
24/16	1/4	12/18	sk=1.00, kr=3.00	1.000	0.050	0.066	0.479	0.066
24/16	1/1	12/18	sk=2.00, kr=6.00	1.000	0.049	0.065	0.469	0.065
24/16	1/2	12/18	sk=2.00, kr=6.00	1.000	0.053	0.071	0.478	0.071
24/16	1/4	12/18	sk=2.00, kr=6.00	1.000	0.047	0.068	0.487	0.068
100/100	1/1	12/18	sk=0.00, kr=0.00	1.000	0.047	0.067	0.737	0.067
100/100	1/2	12/18	sk=0.00, kr=0.00	1.000	0.052	0.069	0.729	0.069
100/100	1/4	12/18	sk=0.00, kr=0.00	1.000	0.046	0.064	0.738	0.064
100/100	1/1	12/18	sk=1.00, kr=3.00	1.000	0.047	0.063	0.734	0.063
100/100	1/2	12/18	sk=1.00, kr=3.00	1.000	0.047	0.066	0.743	0.066
100/100	1/4	12/18	sk=1.00, kr=3.00	1.000	0.050	0.066	0.728	0.066
100/100	1/1	12/18	sk=2.00, kr=6.00	1.000	0.047	0.067	0.734	0.067
100/100	1/2	12/18	sk=2.00, kr=6.00	1.000	0.047	0.062	0.733	0.062
100/100	1/4	12/18	sk=2.00, kr=6.00	1.000	0.049	0.068	0.733	0.068
80/120	1/1	12/18	sk=0.00, kr=0.00	1.000	0.052	0.071	0.727	0.071
80/120	1/2	12/18	sk=0.00, kr=0.00	1.000	0.043	0.064	0.727	0.064
80/120	1/4	12/18	sk=0.00, kr=0.00	1.000	0.050	0.066	0.717	0.066
80/120	1/1	12/18	sk=1.00, kr=3.00	1.000	0.047	0.065	0.728	0.065
80/120	1/2	12/18	sk=1.00, kr=3.00	1.000	0.049	0.072	0.731	0.072
80/120	1/4	12/18	sk=1.00, kr=3.00	1.000	0.046	0.066	0.722	0.066
80/120	1/1	12/18	sk=2.00, kr=6.00	1.000	0.048	0.065	0.741	0.065
80/120	1/2	12/18	sk=2.00, kr=6.00	1.000	0.052	0.069	0.725	0.069
80/120	1/4	12/18	sk=2.00, kr=6.00	1.000	0.045	0.064	0.704	0.064
120/80	1/1	12/18	sk=0.00, kr=0.00	1.000	0.051	0.069	0.734	0.069
120/80	1/2	12/18	sk=0.00, kr=0.00	1.000	0.052	0.066	0.735	0.066
120/80	1/4	12/18	sk=0.00, kr=0.00	1.000	0.057	0.075	0.752	0.075
120/80	1/1	12/18	sk=1.00, kr=3.00	1.000	0.050	0.067	0.743	0.067
120/80	1/2	12/18	sk=1.00, kr=3.00	1.000	0.054	0.078	0.733	0.078
120/80	1/4	12/18	sk=1.00, kr=3.00	1.000	0.050	0.069	0.749	0.069
120/80	1/1	12/18	sk=2.00, kr=6.00	1.000	0.048	0.068	0.735	0.068
120/80	1/2	12/18	sk=2.00, kr=6.00	1.000	0.043	0.063	0.732	0.063
120/80	1/4	12/18	sk=2.00, kr=6.00	1.000	0.048	0.064	0.744	0.064

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The Type I error control of the regular Q, RE, FE and CR tests did not improve as sample sizes increased to 40 and greater (see Table 27 above). Permuted Q maintained adequate Type I error control in these conditions.

Table 28
Type I Error Rate Estimates (τ^2 =1, δ=.8) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00	0.967	0.055	0.073	0.174	0.074
5/5	1/2	12/18	sk=0.00, kr=0.00	0.970	0.053	0.072	0.177	0.073
5/5	1/4	12/18	sk=0.00, kr=0.00	0.978	0.054	0.072	0.181	0.073
5/5	1/1	12/18	sk=1.00, kr=3.00	0.983	0.045	0.061	0.179	0.062
5/5	1/2	12/18	sk=1.00, kr=3.00	0.981	0.049	0.069	0.178	0.069
5/5	1/4	12/18	sk=1.00, kr=3.00	0.987	0.051	0.066	0.191	0.066
5/5	1/1	12/18	sk=2.00, kr=6.00	0.994	0.050	0.064	0.191	0.064
5/5	1/2	12/18	sk=2.00, kr=6.00	0.991	0.052	0.066	0.197	0.066
5/5	1/4	12/18	sk=2.00, kr=6.00	0.993	0.049	0.063	0.197	0.063
4/6	1/1	12/18	sk=0.00, kr=0.00	0.969	0.048	0.063	0.170	0.064
4/6	1/2	12/18	sk=0.00, kr=0.00	0.943	0.051	0.066	0.156	0.067
4/6	1/4	12/18	sk=0.00, kr=0.00	0.932	0.052	0.071	0.158	0.071
4/6	1/1	12/18	sk=1.00, kr=3.00	0.980	0.053	0.068	0.184	0.068
4/6	1/2	12/18	sk=1.00, kr=3.00	0.962	0.058	0.076	0.174	0.076
4/6	1/4	12/18	sk=1.00, kr=3.00	0.953	0.047	0.064	0.156	0.064
4/6	1/1	12/18	sk=2.00, kr=6.00	0.992	0.054	0.068	0.195	0.068
4/6	1/2	12/18	sk=2.00, kr=6.00	0.980	0.047	0.065	0.175	0.065
4/6	1/4	12/18	sk=2.00, kr=6.00	0.971	0.055	0.070	0.171	0.070
6/4	1/1	12/18	sk=0.00, kr=0.00	0.964	0.047	0.066	0.165	0.066
6/4	1/2	12/18	sk=0.00, kr=0.00	0.985	0.055	0.071	0.196	0.071
6/4	1/4	12/18	sk=0.00, kr=0.00	0.993	0.049	0.068	0.205	0.068
6/4	1/1	12/18	sk=1.00, kr=3.00	0.980	0.045	0.061	0.171	0.062
6/4	1/2	12/18	sk=1.00, kr=3.00	0.990	0.049	0.069	0.190	0.069
6/4	1/4	12/18	sk=1.00, kr=3.00	0.996	0.049	0.064	0.194	0.064
6/4	1/1	12/18	sk=2.00, kr=6.00	0.994	0.049	0.067	0.201	0.067
6/4	1/2	12/18	sk=2.00, kr=6.00	0.998	0.050	0.066	0.209	0.066
6/4	1/4	12/18	sk=2.00, kr=6.00	0.997	0.057	0.072	0.226	0.072
20/20	1/1	12/18	sk=0.00, kr=0.00	1.000	0.050	0.069	0.465	0.069
20/20	1/2	12/18	sk=0.00, kr=0.00	1.000	0.047	0.065	0.465	0.065
20/20	1/4	12/18	sk=0.00, kr=0.00	1.000	0.051	0.067	0.447	0.067
20/20	1/1	12/18	sk=1.00, kr=3.00	1.000	0.054	0.067	0.466	0.067
20/20	1/2	12/18	sk=1.00, kr=3.00	1.000	0.051	0.068	0.455	0.068
20/20	1/4	12/18	sk=1.00, kr=3.00	1.000	0.049	0.068	0.458	0.068
20/20	1/1	12/18	sk=2.00, kr=6.00	1.000	0.050	0.068	0.485	0.068
20/20	1/2	12/18	sk=2.00, kr=6.00	1.000	0.053	0.068	0.481	0.068
20/20	1/4	12/18	sk=2.00, kr=6.00	1.000	0.050	0.069	0.463	0.069
16/24	1/1	12/18	sk=0.00, kr=0.00	1.000	0.051	0.071	0.454	0.071
16/24	1/2	12/18	sk=0.00, kr=0.00	1.000	0.051	0.069	0.459	0.069
16/24	1/4	12/18	sk=0.00, kr=0.00	1.000	0.058	0.077	0.433	0.077
16/24	1/1	12/18	sk=1.00, kr=3.00	1.000	0.054	0.076	0.475	0.076
16/24	1/2	12/18	sk=1.00, kr=3.00	1.000	0.044	0.067	0.462	0.067
16/24	1/4	12/18	sk=1.00, kr=3.00	1.000	0.047	0.069	0.436	0.069
16/24	1/1	12/18	sk=2.00, kr=6.00	1.000	0.050	0.069	0.478	0.069
16/24	1/2	12/18	sk=2.00, kr=6.00	1.000	0.050	0.070	0.462	0.070
16/24	1/4	12/18	sk=2.00, kr=6.00	1.000	0.053	0.067	0.454	0.067

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Again, permuted Q maintained adequate Type I error control. However, none of the tests offered any degree of effectiveness as heterogeneity of effects increased to 1 (see Table 28 above). Type I error rates of the other tests remained elevated.

Table 28 (continued)
Type I Error Rate Estimates (τ^2 =1, δ=.8) at α=.05 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	1.000	0.046	0.063	0.454	0.063
24/16	1/2	12/18	sk=0.00, kr=0.00	1.000	0.058	0.077	0.491	0.077
24/16	1/4	12/18	sk=0.00, kr=0.00	1.000	0.046	0.061	0.478	0.061
24/16	1/1	12/18	sk=1.00, kr=3.00	1.000	0.054	0.074	0.461	0.074
24/16	1/2	12/18	sk=1.00, kr=3.00	1.000	0.050	0.070	0.474	0.070
24/16	1/4	12/18	sk=1.00, kr=3.00	1.000	0.055	0.076	0.499	0.076
24/16	1/1	12/18	sk=2.00, kr=6.00	1.000	0.049	0.067	0.471	0.067
24/16	1/2	12/18	sk=2.00, kr=6.00	1.000	0.048	0.066	0.489	0.066
24/16	1/4	12/18	sk=2.00, kr=6.00	1.000	0.048	0.069	0.484	0.069
100/100	1/1	12/18	sk=0.00, kr=0.00	1.000	0.048	0.065	0.736	0.065
100/100	1/2	12/18	sk=0.00, kr=0.00	1.000	0.054	0.069	0.744	0.069
100/100	1/4	12/18	sk=0.00, kr=0.00	1.000	0.049	0.067	0.738	0.067
100/100	1/1	12/18	sk=1.00, kr=3.00	1.000	0.044	0.063	0.746	0.063
100/100	1/2	12/18	sk=1.00, kr=3.00	1.000	0.053	0.076	0.742	0.076
100/100	1/4	12/18	sk=1.00, kr=3.00	1.000	0.047	0.066	0.742	0.066
100/100	1/1	12/18	sk=2.00, kr=6.00	1.000	0.048	0.068	0.726	0.068
100/100	1/2	12/18	sk=2.00, kr=6.00	1.000	0.047	0.066	0.741	0.066
100/100	1/4	12/18	sk=2.00, kr=6.00	1.000	0.052	0.069	0.734	0.069
80/120	1/1	12/18	sk=0.00, kr=0.00	1.000	0.050	0.072	0.720	0.072
80/120	1/2	12/18	sk=0.00, kr=0.00	1.000	0.049	0.068	0.725	0.068
80/120	1/4	12/18	sk=0.00, kr=0.00	1.000	0.050	0.071	0.720	0.071
80/120	1/1	12/18	sk=1.00, kr=3.00	1.000	0.049	0.064	0.744	0.064
80/120	1/2	12/18	sk=1.00, kr=3.00	1.000	0.052	0.071	0.731	0.071
80/120	1/4	12/18	sk=1.00, kr=3.00	1.000	0.052	0.072	0.726	0.072
80/120	1/1	12/18	sk=2.00, kr=6.00	1.000	0.045	0.064	0.742	0.064
80/120	1/2	12/18	sk=2.00, kr=6.00	1.000	0.051	0.069	0.720	0.069
80/120	1/4	12/18	sk=2.00, kr=6.00	1.000	0.052	0.071	0.726	0.071
120/80	1/1	12/18	sk=0.00, kr=0.00	1.000	0.050	0.066	0.726	0.066
120/80	1/2	12/18	sk=0.00, kr=0.00	1.000	0.051	0.070	0.745	0.070
120/80	1/4	12/18	sk=0.00, kr=0.00	1.000	0.053	0.070	0.746	0.070
120/80	1/1	12/18	sk=1.00, kr=3.00	1.000	0.048	0.066	0.742	0.066
120/80	1/2	12/18	sk=1.00, kr=3.00	1.000	0.051	0.070	0.726	0.070
120/80	1/4	12/18	sk=1.00, kr=3.00	1.000	0.053	0.072	0.757	0.072
120/80	1/1	12/18	sk=2.00, kr=6.00	1.000	0.052	0.069	0.720	0.069
120/80	1/2	12/18	sk=2.00, kr=6.00	1.000	0.052	0.071	0.744	0.071
120/80	1/4	12/18	sk=2.00, kr=6.00	1.000	0.051	0.070	0.757	0.070

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Increases in sample size to 40 and above did not improve the Type I error control of any of the tests, aside from permuted Q (see Table 28 above). Due to the continued inflation of Type I error rates, it can be concluded that under increasing heterogeneity of effects, application of these tests would not be warranted.

Table 29

Type I Error Rate Estimates (τ^2 =1, δ=.8) at α=.10 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00	0.984	0.113	0.131	0.254	0.131
5/5	1/2	12/18	sk=0.00, kr=0.00	0.985	0.105	0.127	0.256	0.128
5/5	1/4	12/18	sk=0.00, kr=0.00	0.989	0.106	0.131	0.263	0.131
5/5	1/1	12/18	sk=1.00, kr=3.00	0.992	0.097	0.114	0.260	0.114
5/5	1/2	12/18	sk=1.00, kr=3.00	0.991	0.099	0.119	0.252	0.120
5/5	1/4	12/18	sk=1.00, kr=3.00	0.994	0.101	0.123	0.272	0.124
5/5	1/1	12/18	sk=2.00, kr=6.00	0.998	0.098	0.118	0.269	0.118
5/5	1/2	12/18	sk=2.00, kr=6.00	0.996	0.103	0.121	0.273	0.122
5/5	1/4	12/18	sk=2.00, kr=6.00	0.997	0.101	0.119	0.283	0.119
4/6	1/1	12/18	sk=0.00, kr=0.00	0.986	0.104	0.127	0.249	0.127
4/6	1/2	12/18	sk=0.00, kr=0.00	0.971	0.101	0.121	0.227	0.122
4/6	1/4	12/18	sk=0.00, kr=0.00	0.964	0.107	0.128	0.241	0.129
4/6	1/1	12/18	sk=1.00, kr=3.00	0.991	0.099	0.119	0.259	0.120
4/6	1/2	12/18	sk=1.00, kr=3.00	0.983	0.106	0.126	0.250	0.127
4/6	1/4	12/18	sk=1.00, kr=3.00	0.976	0.096	0.113	0.236	0.114
4/6	1/1	12/18	sk=2.00, kr=6.00	0.997	0.101	0.121	0.283	0.122
4/6	1/2	12/18	sk=2.00, kr=6.00	0.990	0.097	0.120	0.252	0.120
4/6	1/4	12/18	sk=2.00, kr=6.00	0.983	0.103	0.119	0.247	0.119
6/4	1/1	12/18	sk=0.00, kr=0.00	0.982	0.099	0.119	0.251	0.120
6/4	1/2	12/18	sk=0.00, kr=0.00	0.993	0.111	0.128	0.273	0.128
6/4	1/4	12/18	sk=0.00, kr=0.00	0.997	0.099	0.125	0.283	0.125
6/4	1/1	12/18	sk=1.00, kr=3.00	0.990	0.090	0.110	0.250	0.110
6/4	1/2	12/18	sk=1.00, kr=3.00	0.996	0.100	0.123	0.276	0.123
6/4	1/4	12/18	sk=1.00, kr=3.00	0.998	0.095	0.112	0.278	0.112
6/4	1/1	12/18	sk=2.00, kr=6.00	0.998	0.101	0.123	0.280	0.123
6/4	1/2	12/18	sk=2.00, kr=6.00	0.999	0.102	0.120	0.291	0.120
6/4	1/4	12/18	sk=2.00, kr=6.00	0.998	0.111	0.125	0.300	0.125
20/20	1/1	12/18	sk=0.00, kr=0.00	1.000	0.105	0.124	0.534	0.124
20/20	1/2	12/18	sk=0.00, kr=0.00	1.000	0.098	0.119	0.538	0.119
20/20	1/4	12/18	sk=0.00, kr=0.00	1.000	0.099	0.115	0.521	0.115
20/20	1/1	12/18	sk=1.00, kr=3.00	1.000	0.102	0.124	0.544	0.124
20/20	1/2	12/18	sk=1.00, kr=3.00	1.000	0.103	0.126	0.531	0.126
20/20	1/4	12/18	sk=1.00, kr=3.00	1.000	0.101	0.124	0.532	0.124
20/20	1/1	12/18	sk=2.00, kr=6.00	1.000	0.098	0.125	0.560	0.125
20/20	1/2	12/18	sk=2.00, kr=6.00	1.000	0.094	0.123	0.555	0.123
20/20	1/4	12/18	sk=2.00, kr=6.00	1.000	0.102	0.122	0.540	0.122
16/24	1/1	12/18	sk=0.00, kr=0.00	1.000	0.101	0.125	0.528	0.125
16/24	1/2	12/18	sk=0.00, kr=0.00	1.000	0.099	0.125	0.540	0.125
16/24	1/4	12/18	sk=0.00, kr=0.00	1.000	0.108	0.127	0.510	0.127
16/24	1/1	12/18	sk=1.00, kr=3.00	1.000	0.110	0.129	0.553	0.129
16/24	1/2	12/18	sk=1.00, kr=3.00	1.000	0.095	0.123	0.537	0.123
16/24	1/4	12/18	sk=1.00, kr=3.00	1.000	0.099	0.118	0.512	0.118
16/24	1/1	12/18	sk=2.00, kr=6.00	1.000	0.103	0.122	0.549	0.122
16/24	1/2	12/18	sk=2.00, kr=6.00	1.000	0.104	0.126	0.537	0.126
16/24	1/4	12/18	sk=2.00, kr=6.00	1.000	0.102	0.123	0.528	0.123

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Again, increasing the nominal alpha from .05 to .10 did not provide a substantive improvement to the performance of any of the tests, including permuted Q. And Type I error control was poor for all tests, other than permuted Q (see Table 29 above).

Table 29 (continued)
Type I Error Rate Estimates (τ^2 =1, δ=.8) at α=.10 for K=30, N=5000

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	1.000	0.093	0.113	0.526	0.113
24/16	1/2	12/18	sk=0.00, kr=0.00	1.000	0.110	0.135	0.562	0.135
24/16	1/4	12/18	sk=0.00, kr=0.00	1.000	0.097	0.118	0.551	0.118
24/16	1/1	12/18	sk=1.00, kr=3.00	1.000	0.106	0.127	0.543	0.127
24/16	1/2	12/18	sk=1.00, kr=3.00	1.000	0.107	0.128	0.553	0.128
24/16	1/4	12/18	sk=1.00, kr=3.00	1.000	0.105	0.129	0.572	0.129
24/16	1/1	12/18	sk=2.00, kr=6.00	1.000	0.100	0.121	0.551	0.121
24/16	1/2	12/18	sk=2.00, kr=6.00	1.000	0.100	0.122	0.562	0.122
24/16	1/4	12/18	sk=2.00, kr=6.00	1.000	0.103	0.124	0.555	0.124
100/100	1/1	12/18	sk=0.00, kr=0.00	1.000	0.100	0.123	0.778	0.123
100/100	1/2	12/18	sk=0.00, kr=0.00	1.000	0.104	0.130	0.783	0.130
100/100	1/4	12/18	sk=0.00, kr=0.00	1.000	0.099	0.123	0.776	0.123
100/100	1/1	12/18	sk=1.00, kr=3.00	1.000	0.093	0.115	0.786	0.115
100/100	1/2	12/18	sk=1.00, kr=3.00	1.000	0.111	0.130	0.779	0.130
100/100	1/4	12/18	sk=1.00, kr=3.00	1.000	0.097	0.119	0.780	0.119
100/100	1/1	12/18	sk=2.00, kr=6.00	1.000	0.105	0.128	0.770	0.128
100/100	1/2	12/18	sk=2.00, kr=6.00	1.000	0.098	0.120	0.781	0.120
100/100	1/4	12/18	sk=2.00, kr=6.00	1.000	0.102	0.126	0.777	0.126
80/120	1/1	12/18	sk=0.00, kr=0.00	1.000	0.100	0.121	0.764	0.121
80/120	1/2	12/18	sk=0.00, kr=0.00	1.000	0.101	0.120	0.770	0.120
80/120	1/4	12/18	sk=0.00, kr=0.00	1.000	0.099	0.124	0.762	0.124
80/120	1/1	12/18	sk=1.00, kr=3.00	1.000	0.101	0.124	0.783	0.124
80/120	1/2	12/18	sk=1.00, kr=3.00	1.000	0.100	0.122	0.775	0.122
80/120	1/4	12/18	sk=1.00, kr=3.00	1.000	0.102	0.125	0.769	0.125
80/120	1/1	12/18	sk=2.00, kr=6.00	1.000	0.098	0.120	0.786	0.120
80/120	1/2	12/18	sk=2.00, kr=6.00	1.000	0.104	0.129	0.761	0.129
80/120	1/4	12/18	sk=2.00, kr=6.00	1.000	0.099	0.117	0.773	0.117
120/80	1/1	12/18	sk=0.00, kr=0.00	1.000	0.104	0.127	0.766	0.127
120/80	1/2	12/18	sk=0.00, kr=0.00	1.000	0.102	0.124	0.780	0.124
120/80	1/4	12/18	sk=0.00, kr=0.00	1.000	0.100	0.125	0.789	0.125
120/80	1/1	12/18	sk=1.00, kr=3.00	1.000	0.097	0.118	0.787	0.118
120/80	1/2	12/18	sk=1.00, kr=3.00	1.000	0.103	0.127	0.770	0.127
120/80	1/4	12/18	sk=1.00, kr=3.00	1.000	0.099	0.122	0.793	0.122
120/80	1/1	12/18	sk=2.00, kr=6.00	1.000	0.105	0.120	0.761	0.120
120/80	1/2	12/18	sk=2.00, kr=6.00	1.000	0.100	0.124	0.783	0.124
120/80	1/4	12/18	sk=2.00, kr=6.00	1.000	0.104	0.125	0.796	0.125

Note: All unshaded areas for each of the tests reflect either inflated or conservative Type I error. Shaded cells signify those conditions in which the error rate fell within Bradley's criterion.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Increasing nominal alpha to .10 did not alter the effectiveness of permuted Q. But this test did remain robust under these conditions (see Table 29 above). Despite the increase in sample size, the Type I error rates of all other tests continued to exceed the nominal alpha level.

Summary of Type I Error Rate Estimates

All of the tests offered some degree of effectiveness when τ^2=0, whether nominal alpha was .05 or .10. Robustness was inhibited when K=10 and sample sizes were small.

Once K was increased to 30, at τ^2=0, δ=0, all tests evidenced robust performance for several (13 and 19) conditions. The regular Q (14 conditions of robustness), FE (21 conditions of robustness) and CR (19 conditions of robustness) tests produced a concentration of well-maintained Type I error conditions as the sample size of the primary studies increased to 200 under the equal groups condition. As unequal sample sizes were introduced, both of these tests had diminished Type I error control. The RE test was less robust than either of these other two tests when sample sizes increased to 200. The RE test (18 conditions of robustness) demonstrated particular robustness when the sample sizes at N=10, 40 and 200 had the first group with the larger N than the second.

As an effect size greater than 0 was introduced, the performance of the regular Q and FE test diminished. The regular Q had 14 conditions with adequate Type I error when the effect size was 0 (see Table 17) as compared to 8 in the present set of conditions (Table 18). The FE test had less of a decline in performance from 21 adequate conditions vs. 17 in the present set. The RE and CR tests generated more conservative Type I error rates than any of the other 3 tests, but the greatest frequency (other than permuted Q) of conditions with adequately controlled Type I error. Again, the RE test performed best when the first group had the larger sample size.

There was a dramatic increase in the number of conditions with inflated Type I error for all tests, except permuted Q, once τ^2 increased to .33 (see Tables 19 and 20). As sample size increased to 40, the Type I error for each of the 4 tests exhibited another notable increase. This pattern resulted in a total absence of error control by the aforementioned tests.

The same pattern of results continued as displayed in Table 19 and the permuted Q retained its robustness in terms of Type I error control, when τ^2 = .33 and δ=.8, K=10 (Table 20). Again, none of the other tests maintained robustness under increasing heterogeneity of effects.

At δ=.8, τ^2 = .33, permuted Q remained robust and effective with the increase of K to 30 and upon the increase in sample size to 40 (see Table 23). All other tests failed to provide adequate robustness under

these conditions. The RE and CR tests showed minimal improvement to robustness with the increase in effect size to .8. However, at sample sizes below 40 effective use of these tests in the few robust conditions would not be warranted (to be discussed further in the summary on power estimates).

Increasing the nominal alpha level from .05 to .10 did not enhance the performance of any of the five tests being investigated, when K=10 (see Table 21). The permuted Q still maintained adequate Type I error control, though it was not truly effective. No other test maintained adequate Type I error control. As K increased to 30 (see Table 24), the robustness of permuted Q again became integral to its utility.

Only permuted Q maintained adequate Type I error control under increasing heterogeneity of effects at τ^2=1(see Tables 26 and 27). However, none of the tests offered any degree of effectiveness as heterogeneity of effects increased to 1 (power was lacking). Type I error rates of the other tests remained elevated. Though the RE and CR tests' error rates were not as inflated as regular Q and the FE test, increasing heterogeneity contributed to the erosion of robustness for both tests.

Conditions Wherein Test Simulations Demonstrate Both Adequate Type I Error and Good Power

It is not enough for a test to maintain adequate Type I error control. Without possessing good power under given conditions, a test will fail to detect true differences when they occur. Failure to detect these true differences renders a test ineffective in the ultimate task of identifying true variance between groups. For example, it would be important for a school administrator to know the effectiveness of a reading program for various groups of students in order to evaluate the usefulness of that program before investing future resources.

In order to provide a more comprehensive analysis of the utility of the five tests being investigated, two sets of power tables are presented. Both lend a picture of the degree of specific effectiveness for each test. The focus of each together will facilitate the practitioner in his/her selection of the statistics to be applied for meta-analytic research.

Power Estimates for Conditions Indicating Adequate Type I Error Control

The following presentation provides an explanation for both the conditions under which each test extends good power to detect true differences, as well as situations for which the tests possessed inadequate power. The first set of power estimate tables (Tables 30-38) originated from power estimates for all conditions under investigation.

Power estimates were removed for those conditions under which tests did not permit adequate Type I error control. All estimates remaining reflected both good and insufficient power values for those conditions where tests proved to be robust. By retaining all power values for all conditions signifying adequate Type I error control, two pieces of information arise. First, this presentation can illuminate those conditions under which a test performed either effectively or not (effective conditions were highlighted in green). Specifically, this data presentation allows the reader to determine when a test achieved Type I error control and whether or not good power accompanied the robustness of the test. Secondly, one can determine to what extent a test failed to provide good power under given conditions. Good power was identified as any power estimate of .795 or higher.

The reason for presenting the results in this manner is that most practitioners first consider whether a test demonstrates adequate Type I error control. If Type I error is not well-maintained, then considerations of power are deemed irrelevant. Typically, the practitioner wants to determine first whether a particular null hypothesis is to be rejected or maintained. If a given test does not provide an adequate degree of robustness, the test, for that given set of conditions, is rendered invalid. The flip side of this argument is that the researcher is no longer taking a disconfirmatory approach to hypothesis testing. Furthermore, alternative hypotheses are not being investigated. But that concern involves a deeper epistemological argument, beyond the purview of the present study.

Table 30

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 0$, $\delta=.8$) at $\alpha=.05$ for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00		0.376			
5/5	1/2	4/6	sk=0.00, kr=0.00		0.368			
5/5	1/4	4/6	sk=0.00, kr=0.00		0.370			
5/5	1/1	4/6	sk=1.00, kr=3.00	0.143	0.413			
5/5	1/2	4/6	sk=1.00, kr=3.00		0.421			
5/5	1/4	4/6	sk=1.00, kr=3.00		0.439			
5/5	1/1	4/6	sk=2.00, kr=6.00		0.455		0.498	0.489
5/5	1/2	4/6	sk=2.00, kr=6.00		0.500			
5/5	1/4	4/6	sk=2.00, kr=6.00		0.513			
4/6	1/1	4/6	sk=0.00, kr=0.00		0.364			
4/6	1/2	4/6	sk=0.00, kr=0.00		0.383			
4/6	1/4	4/6	sk=0.00, kr=0.00		0.411			
4/6	1/1	4/6	sk=1.00, kr=3.00		0.403			
4/6	1/2	4/6	sk=1.00, kr=3.00		0.441			
4/6	1/4	4/6	sk=1.00, kr=3.00		0.480			
4/6	1/1	4/6	sk=2.00, kr=6.00		0.430	0.448	0.460	0.451
4/6	1/2	4/6	sk=2.00, kr=6.00		0.503			
4/6	1/4	4/6	sk=2.00, kr=6.00		0.544			
6/4	1/1	4/6	sk=0.00, kr=0.00		0.363			
6/4	1/2	4/6	sk=0.00, kr=0.00		0.344	0.380	0.400	0.381
6/4	1/4	4/6	sk=0.00, kr=0.00		0.322			
6/4	1/1	4/6	sk=1.00, kr=3.00		0.401			
6/4	1/2	4/6	sk=1.00, kr=3.00	0.212	0.393			
6/4	1/4	4/6	sk=1.00, kr=3.00		0.384	0.456	0.487	0.459
6/4	1/1	4/6	sk=2.00, kr=6.00		0.435	0.466	0.478	0.470
6/4	1/2	4/6	sk=2.00, kr=6.00	0.249	0.448			
6/4	1/4	4/6	sk=2.00, kr=6.00		0.463			
20/20	1/1	4/6	sk=0.00, kr=0.00		0.917		0.966	0.962
20/20	1/2	4/6	sk=0.00, kr=0.00		0.917	0.957	0.965	0.958
20/20	1/4	4/6	sk=0.00, kr=0.00		0.904		0.969	0.961
20/20	1/1	4/6	sk=1.00, kr=3.00		0.912	0.962	0.970	0.963
20/20	1/2	4/6	sk=1.00, kr=3.00		0.925			
20/20	1/4	4/6	sk=1.00, kr=3.00		0.945			
20/20	1/1	4/6	sk=2.00, kr=6.00		0.917	0.960		0.961
20/20	1/2	4/6	sk=2.00, kr=6.00					
20/20	1/4	4/6	sk=2.00, kr=6.00		0.970			
16/24	1/1	4/6	sk=0.00, kr=0.00		0.904			
16/24	1/2	4/6	sk=0.00, kr=0.00					
16/24	1/4	4/6	sk=0.00, kr=0.00		0.937			
16/24	1/1	4/6	sk=1.00, kr=3.00		0.908	0.956		0.957
16/24	1/2	4/6	sk=1.00, kr=3.00		0.937			
16/24	1/4	4/6	sk=1.00, kr=3.00		0.959			
16/24	1/1	4/6	sk=2.00, kr=6.00		0.911	0.952		
16/24	1/2	4/6	sk=2.00, kr=6.00		0.956			
16/24	1/4	4/6	sk=2.00, kr=6.00		0.976			

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Power was low for all test conditions with adequate Type I error, when sample sizes were small (see Table 30). Because power was low, none of the tests would prove effective under these conditions, until sample sizes increase to 40 and above.

Table 30 (continued)

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 0$, $\delta=.8$) at $\alpha=.05$ for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00		0.907		0.961	
24/16	1/2	4/6	sk=0.00, kr=0.00		0.885	0.948		
24/16	1/4	4/6	sk=0.00, kr=0.00		0.870			
24/16	1/1	4/6	sk=1.00, kr=3.00		0.913	0.956	0.966	0.957
24/16	1/2	4/6	sk=1.00, kr=3.00			0.965	0.973	0.966
24/16	1/4	4/6	sk=1.00, kr=3.00		0.905	0.960	0.974	0.962
24/16	1/1	4/6	sk=2.00, kr=6.00		0.913	0.959		
24/16	1/2	4/6	sk=2.00, kr=6.00	0.828	0.935		0.979	0.974
24/16	1/4	4/6	sk=2.00, kr=6.00		0.948			
100/100	1/1	4/6	sk=0.00, kr=0.00	1.000	1.000			
100/100	1/2	4/6	sk=0.00, kr=0.00	1.000	1.000	1.000	1.000	1.000
100/100	1/4	4/6	sk=0.00, kr=0.00		1.000	1.000		1.000
100/100	1/1	4/6	sk=1.00, kr=3.00			1.000		1.000
100/100	1/2	4/6	sk=1.00, kr=3.00		1.000		1.000	
100/100	1/4	4/6	sk=1.00, kr=3.00		1.000			
100/100	1/1	4/6	sk=2.00, kr=6.00		1.000			
100/100	1/2	4/6	sk=2.00, kr=6.00		1.000			
100/100	1/4	4/6	sk=2.00, kr=6.00					
80/120	1/1	4/6	sk=0.00, kr=0.00	1.000		1.000	1.000	1.000
80/120	1/2	4/6	sk=0.00, kr=0.00		1.000			
80/120	1/4	4/6	sk=0.00, kr=0.00		1.000			
80/120	1/1	4/6	sk=1.00, kr=3.00		1.000	1.000		1.000
80/120	1/2	4/6	sk=1.00, kr=3.00		1.000			
80/120	1/4	4/6	sk=1.00, kr=3.00		1.000			
80/120	1/1	4/6	sk=2.00, kr=6.00		1.000			
80/120	1/2	4/6	sk=2.00, kr=6.00		1.000			
80/120	1/4	4/6	sk=2.00, kr=6.00		1.000			
120/80	1/1	4/6	sk=0.00, kr=0.00	1.000	1.000		1.000	1.000
120/80	1/2	4/6	sk=0.00, kr=0.00		1.000	1.000		
120/80	1/4	4/6	sk=0.00, kr=0.00		1.000			
120/80	1/1	4/6	sk=1.00, kr=3.00		1.000	1.000		1.000
120/80	1/2	4/6	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
120/80	1/4	4/6	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
120/80	1/1	4/6	sk=2.00, kr=6.00		1.000			
120/80	1/2	4/6	sk=2.00, kr=6.00	1.000	1.000		1.000	
120/80	1/4	4/6	sk=2.00, kr=6.00		1.000			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

As sample sizes increased to 40 and greater, there were a greater number of conditions with both robustness and good power for all of the tests (see Table 30). After permuted Q, the RE and CR tests provided the greatest number of conditions with robustness and good power (19 conditions each). Regular Q yielded 6 effective conditions, while the FE test demonstrated 16. Permuted Q achieved the majority of effective conditions, once sample sizes increased to 40 and greater.

Table 31

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 0$, δ=.8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00		0.869			
5/5	1/2	12/18	sk=0.00, kr=0.00		0.877			
5/5	1/4	12/18	sk=0.00, kr=0.00		0.872			
5/5	1/1	12/18	sk=1.00, kr=3.00					
5/5	1/2	12/18	sk=1.00, kr=3.00		0.916			
5/5	1/4	12/18	sk=1.00, kr=3.00		0.931			
5/5	1/1	12/18	sk=2.00, kr=6.00		0.930		0.928	
5/5	1/2	12/18	sk=2.00, kr=6.00		0.959			
5/5	1/4	12/18	sk=2.00, kr=6.00		0.967			
4/6	1/1	12/18	sk=0.00, kr=0.00		0.865			
4/6	1/2	12/18	sk=0.00, kr=0.00		0.880			
4/6	1/4	12/18	sk=0.00, kr=0.00		0.891			
4/6	1/1	12/18	sk=1.00, kr=3.00					
4/6	1/2	12/18	sk=1.00, kr=3.00		0.930			
4/6	1/4	12/18	sk=1.00, kr=3.00		0.942			
4/6	1/1	12/18	sk=2.00, kr=6.00		0.923			
4/6	1/2	12/18	sk=2.00, kr=6.00		0.962			
4/6	1/4	12/18	sk=2.00, kr=6.00		0.977			
6/4	1/1	12/18	sk=0.00, kr=0.00		0.870			
6/4	1/2	12/18	sk=0.00, kr=0.00		0.836			
6/4	1/4	12/18	sk=0.00, kr=0.00		0.814	0.836		0.839
6/4	1/1	12/18	sk=1.00, kr=3.00		0.902			
6/4	1/2	12/18	sk=1.00, kr=3.00	0.375	0.904			
6/4	1/4	12/18	sk=1.00, kr=3.00		0.907		0.926	
6/4	1/1	12/18	sk=2.00, kr=6.00		0.924			
6/4	1/2	12/18	sk=2.00, kr=6.00		0.949			
6/4	1/4	12/18	sk=2.00, kr=6.00		0.953			
20/20	1/1	12/18	sk=0.00, kr=0.00		1.000		1.000	1.000
20/20	1/2	12/18	sk=0.00, kr=0.00	0.983	1.000		1.000	1.000
20/20	1/4	12/18	sk=0.00, kr=0.00			1.000	1.000	1.000
20/20	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
20/20	1/2	12/18	sk=1.00, kr=3.00		1.000			
20/20	1/4	12/18	sk=1.00, kr=3.00		1.000			
20/20	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
20/20	1/2	12/18	sk=2.00, kr=6.00		1.000			
20/20	1/4	12/18	sk=2.00, kr=6.00		1.000			
16/24	1/1	12/18	sk=0.00, kr=0.00		1.000		1.000	
16/24	1/2	12/18	sk=0.00, kr=0.00		1.000			
16/24	1/4	12/18	sk=0.00, kr=0.00		1.000			
16/24	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
16/24	1/2	12/18	sk=1.00, kr=3.00					
16/24	1/4	12/18	sk=1.00, kr=3.00		1.000			
16/24	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
16/24	1/2	12/18	sk=2.00, kr=6.00		1.000			
16/24	1/4	12/18	sk=2.00, kr=6.00		1.000			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

With the increase in K to 30 (Table 31), the FE test offered the same number of effective, powerful conditions as with K=10 (see Table 30). But the RE test manifested an increased number of effective conditions for use (24 as compared to 19 in the previous K=10 conditions), primarily occurring in the equal and first group with greater sample size conditions. The CR test presented no change in the number of effective conditions and these were concentrated in the equal sample sizes conditions.

Table 31 (continued)

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 0$, δ=.8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00	0.978	1.000			
24/16	1/2	12/18	sk=0.00, kr=0.00		1.000	1.000		1.000
24/16	1/4	12/18	sk=0.00, kr=0.00		1.000	1.000		
24/16	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
24/16	1/2	12/18	sk=1.00, kr=3.00		1.000		1.000	1.000
24/16	1/4	12/18	sk=1.00, kr=3.00		1.000	1.000		1.000
24/16	1/1	12/18	sk=2.00, kr=6.00		1.000			
24/16	1/2	12/18	sk=2.00, kr=6.00		1.000	1.000	1.000	1.000
24/16	1/4	12/18	sk=2.00, kr=6.00		1.000			
100/100	1/1	12/18	sk=0.00, kr=0.00	1.000	1.000	1.000	1.000	1.000
100/100	1/2	12/18	sk=0.00, kr=0.00	1.000	1.000	1.000		1.000
100/100	1/4	12/18	sk=0.00, kr=0.00			1.000	1.000	1.000
100/100	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000		1.000
100/100	1/2	12/18	sk=1.00, kr=3.00		1.000		1.000	1.000
100/100	1/4	12/18	sk=1.00, kr=3.00		1.000			
100/100	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
100/100	1/2	12/18	sk=2.00, kr=6.00		1.000			
100/100	1/4	12/18	sk=2.00, kr=6.00		1.000			
80/120	1/1	12/18	sk=0.00, kr=0.00	1.000				
80/120	1/2	12/18	sk=0.00, kr=0.00		1.000			
80/120	1/4	12/18	sk=0.00, kr=0.00		1.000			
80/120	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000		
80/120	1/2	12/18	sk=1.00, kr=3.00					
80/120	1/4	12/18	sk=1.00, kr=3.00		1.000			
80/120	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
80/120	1/2	12/18	sk=2.00, kr=6.00		1.000			
80/120	1/4	12/18	sk=2.00, kr=6.00		1.000			
120/80	1/1	12/18	sk=0.00, kr=0.00	1.000	1.000		1.000	
120/80	1/2	12/18	sk=0.00, kr=0.00		1.000	1.000		
120/80	1/4	12/18	sk=0.00, kr=0.00		1.000			
120/80	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000		1.000
120/80	1/2	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
120/80	1/4	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
120/80	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
120/80	1/2	12/18	sk=2.00, kr=6.00	1.000				
120/80	1/4	12/18	sk=2.00, kr=6.00		1.000			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Permuted Q had an increased number of effective conditions both with the conditions with smaller and larger sample size (compare Table 30 to Table 31). These effective conditions for use were spread across all conditions for permuted Q.

Table 32

Power Estimates for Conditions Indicating Adequate Type I Error Control (τ^2 = .33, δ=.8) at α=.05 for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00		0.228			
5/5	1/2	4/6	sk=0.00, kr=0.00		0.229			
5/5	1/4	4/6	sk=0.00, kr=0.00		0.234			
5/5	1/1	4/6	sk=1.00, kr=3.00		0.247			
5/5	1/2	4/6	sk=1.00, kr=3.00		0.238			
5/5	1/4	4/6	sk=1.00, kr=3.00		0.245			
5/5	1/1	4/6	sk=2.00, kr=6.00		0.248			
5/5	1/2	4/6	sk=2.00, kr=6.00		0.246			
5/5	1/4	4/6	sk=2.00, kr=6.00		0.269			
4/6	1/1	4/6	sk=0.00, kr=0.00		0.231			
4/6	1/2	4/6	sk=0.00, kr=0.00		0.227			
4/6	1/4	4/6	sk=0.00, kr=0.00		0.247			
4/6	1/1	4/6	sk=1.00, kr=3.00		0.232			
4/6	1/2	4/6	sk=1.00, kr=3.00		0.248			
4/6	1/4	4/6	sk=1.00, kr=3.00					
4/6	1/1	4/6	sk=2.00, kr=6.00		0.248			
4/6	1/2	4/6	sk=2.00, kr=6.00					
4/6	1/4	4/6	sk=2.00, kr=6.00		0.281			
6/4	1/1	4/6	sk=0.00, kr=0.00		0.229			
6/4	1/2	4/6	sk=0.00, kr=0.00		0.219			
6/4	1/4	4/6	sk=0.00, kr=0.00		0.215			
6/4	1/1	4/6	sk=1.00, kr=3.00		0.230			
6/4	1/2	4/6	sk=1.00, kr=3.00		0.242			
6/4	1/4	4/6	sk=1.00, kr=3.00		0.232			
6/4	1/1	4/6	sk=2.00, kr=6.00		0.241			
6/4	1/2	4/6	sk=2.00, kr=6.00		0.245			
6/4	1/4	4/6	sk=2.00, kr=6.00		0.251			
20/20	1/1	4/6	sk=0.00, kr=0.00					
20/20	1/2	4/6	sk=0.00, kr=0.00		0.366			
20/20	1/4	4/6	sk=0.00, kr=0.00		0.371			
20/20	1/1	4/6	sk=1.00, kr=3.00		0.374			
20/20	1/2	4/6	sk=1.00, kr=3.00		0.367			
20/20	1/4	4/6	sk=1.00, kr=3.00		0.394			
20/20	1/1	4/6	sk=2.00, kr=6.00		0.376			
20/20	1/2	4/6	sk=2.00, kr=6.00		0.381			
20/20	1/4	4/6	sk=2.00, kr=6.00		0.381			
16/24	1/1	4/6	sk=0.00, kr=0.00		0.366			
16/24	1/2	4/6	sk=0.00, kr=0.00		0.372			
16/24	1/4	4/6	sk=0.00, kr=0.00		0.373			
16/24	1/1	4/6	sk=1.00, kr=3.00		0.369			
16/24	1/2	4/6	sk=1.00, kr=3.00		0.392			
16/24	1/4	4/6	sk=1.00, kr=3.00					
16/24	1/1	4/6	sk=2.00, kr=6.00		0.379			
16/24	1/2	4/6	sk=2.00, kr=6.00		0.388			
16/24	1/4	4/6	sk=2.00, kr=6.00		0.394			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

With the increase in τ^2 to .33 at K=10 for nominal alpha, .05, sufficient power was absent for all tests, including permuted Q (see Table 32). Power remained limited for permuted Q, regardless of increases in sample size, as long as K=10.

Table 32 (continued)

Power Estimates for Conditions Indicating Adequate Type I Error Control (τ^2 = .33, δ=.8) at α=.05 for =10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00		0.371			
24/16	1/2	4/6	sk=0.00, kr=0.00		0.365			
24/16	1/4	4/6	sk=0.00, kr=0.00		0.359			
24/16	1/1	4/6	sk=1.00, kr=3.00		0.371			
24/16	1/2	4/6	sk=1.00, kr=3.00		0.367			
24/16	1/4	4/6	sk=1.00, kr=3.00		0.356			
24/16	1/1	4/6	sk=2.00, kr=6.00		0.372			
24/16	1/2	4/6	sk=2.00, kr=6.00		0.381			
24/16	1/4	4/6	sk=2.00, kr=6.00		0.370			
100/100	1/1	4/6	sk=0.00, kr=0.00		0.436			
100/100	1/2	4/6	sk=0.00, kr=0.00		0.447			
100/100	1/4	4/6	sk=0.00, kr=0.00		0.447			
100/100	1/1	4/6	sk=1.00, kr=3.00		0.445			
100/100	1/2	4/6	sk=1.00, kr=3.00		0.441			
100/100	1/4	4/6	sk=1.00, kr=3.00		0.452			
100/100	1/1	4/6	sk=2.00, kr=6.00		0.447			
100/100	1/2	4/6	sk=2.00, kr=6.00		0.446			
100/100	1/4	4/6	sk=2.00, kr=6.00		0.444			
80/120	1/1	4/6	sk=0.00, kr=0.00		0.455			
80/120	1/2	4/6	sk=0.00, kr=0.00					
80/120	1/4	4/6	sk=0.00, kr=0.00		0.444			
80/120	1/1	4/6	sk=1.00, kr=3.00		0.446			
80/120	1/2	4/6	sk=1.00, kr=3.00					
80/120	1/4	4/6	sk=1.00, kr=3.00		0.446			
80/120	1/1	4/6	sk=2.00, kr=6.00		0.447			
80/120	1/2	4/6	sk=2.00, kr=6.00		0.443			
80/120	1/4	4/6	sk=2.00, kr=6.00		0.452			
120/80	1/1	4/6	sk=0.00, kr=0.00		0.450			
120/80	1/2	4/6	sk=0.00, kr=0.00		0.447			
120/80	1/4	4/6	sk=0.00, kr=0.00		0.450			
120/80	1/1	4/6	sk=1.00, kr=3.00		0.441			
120/80	1/2	4/6	sk=1.00, kr=3.00		0.437			
120/80	1/4	4/6	sk=1.00, kr=3.00		0.452			
120/80	1/1	4/6	sk=2.00, kr=6.00					
120/80	1/2	4/6	sk=2.00, kr=6.00		0.441			
120/80	1/4	4/6	sk=2.00, kr=6.00		0.439			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 33

Power Estimates for Conditions Indicating Adequate Type I Error Control (τ^2 = .33, δ=.8) at α=.10 for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00		0.360			
5/5	1/2	4/6	sk=0.00, kr=0.00		0.359			
5/5	1/4	4/6	sk=0.00, kr=0.00		0.365			
5/5	1/1	4/6	sk=1.00, kr=3.00		0.375			
5/5	1/2	4/6	sk=1.00, kr=3.00		0.372			
5/5	1/4	4/6	sk=1.00, kr=3.00		0.387			
5/5	1/1	4/6	sk=2.00, kr=6.00		0.373			
5/5	1/2	4/6	sk=2.00, kr=6.00		0.384			
5/5	1/4	4/6	sk=2.00, kr=6.00		0.409			
4/6	1/1	4/6	sk=0.00, kr=0.00		0.364			
4/6	1/2	4/6	sk=0.00, kr=0.00		0.352			
4/6	1/4	4/6	sk=0.00, kr=0.00		0.377			
4/6	1/1	4/6	sk=1.00, kr=3.00		0.366			
4/6	1/2	4/6	sk=1.00, kr=3.00		0.388			
4/6	1/4	4/6	sk=1.00, kr=3.00					
4/6	1/1	4/6	sk=2.00, kr=6.00		0.379			
4/6	1/2	4/6	sk=2.00, kr=6.00		0.407			
4/6	1/4	4/6	sk=2.00, kr=6.00		0.424			
6/4	1/1	4/6	sk=0.00, kr=0.00		0.362			
6/4	1/2	4/6	sk=0.00, kr=0.00		0.351			
6/4	1/4	4/6	sk=0.00, kr=0.00		0.334			
6/4	1/1	4/6	sk=1.00, kr=3.00		0.353			
6/4	1/2	4/6	sk=1.00, kr=3.00		0.377			
6/4	1/4	4/6	sk=1.00, kr=3.00		0.362			
6/4	1/1	4/6	sk=2.00, kr=6.00		0.363			
6/4	1/2	4/6	sk=2.00, kr=6.00		0.383			
6/4	1/4	4/6	sk=2.00, kr=6.00		0.390			
20/20	1/1	4/6	sk=0.00, kr=0.00					
20/20	1/2	4/6	sk=0.00, kr=0.00		0.527			
20/20	1/4	4/6	sk=0.00, kr=0.00		0.527			
20/20	1/1	4/6	sk=1.00, kr=3.00					
20/20	1/2	4/6	sk=1.00, kr=3.00		0.531			
20/20	1/4	4/6	sk=1.00, kr=3.00		0.546			
20/20	1/1	4/6	sk=2.00, kr=6.00		0.531			
20/20	1/2	4/6	sk=2.00, kr=6.00		0.545			
20/20	1/4	4/6	sk=2.00, kr=6.00		0.539			
16/24	1/1	4/6	sk=0.00, kr=0.00		0.527			
16/24	1/2	4/6	sk=0.00, kr=0.00		0.528			
16/24	1/4	4/6	sk=0.00, kr=0.00		0.533			
16/24	1/1	4/6	sk=1.00, kr=3.00					
16/24	1/2	4/6	sk=1.00, kr=3.00		0.548			
16/24	1/4	4/6	sk=1.00, kr=3.00		0.543			
16/24	1/1	4/6	sk=2.00, kr=6.00		0.527			
16/24	1/2	4/6	sk=2.00, kr=6.00		0.537			
16/24	1/4	4/6	sk=2.00, kr=6.00		0.548			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The increase in nominal alpha to .10 for the τ^2=.33 condition at K=10 (see Table 33) provided a small, but inadequate, improvement in power for all of the tests, including permuted Q. The combined increase in heterogeneity and small K suppressed power for all of the tests, making them all ineffective. This pattern was evident as heterogeneity increased to 1 for both nominal alpha .05 and .10 when K=10.

Table 33 (continued)

Power Estimates for Conditions Indicating Adequate Type I Error Control (τ^2 = .33, δ=.8) at α=.10 for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00		0.520			
24/16	1/2	4/6	sk=0.00, kr=0.00		0.517			
24/16	1/4	4/6	sk=0.00, kr=0.00		0.521			
24/16	1/1	4/6	sk=1.00, kr=3.00		0.531			
24/16	1/2	4/6	sk=1.00, kr=3.00		0.524			
24/16	1/4	4/6	sk=1.00, kr=3.00		0.507			
24/16	1/1	4/6	sk=2.00, kr=6.00		0.522			
24/16	1/2	4/6	sk=2.00, kr=6.00		0.534			
24/16	1/4	4/6	sk=2.00, kr=6.00		0.523			
100/100	1/1	4/6	sk=0.00, kr=0.00		0.600			
100/100	1/2	4/6	sk=0.00, kr=0.00		0.606			
100/100	1/4	4/6	sk=0.00, kr=0.00		0.604			
100/100	1/1	4/6	sk=1.00, kr=3.00		0.616			
100/100	1/2	4/6	sk=1.00, kr=3.00		0.600			
100/100	1/4	4/6	sk=1.00, kr=3.00		0.611			
100/100	1/1	4/6	sk=2.00, kr=6.00		0.606			
100/100	1/2	4/6	sk=2.00, kr=6.00		0.608			
100/100	1/4	4/6	sk=2.00, kr=6.00		0.606			
80/120	1/1	4/6	sk=0.00, kr=0.00		0.609			
80/120	1/2	4/6	sk=0.00, kr=0.00		0.595			
80/120	1/4	4/6	sk=0.00, kr=0.00		0.606			
80/120	1/1	4/6	sk=1.00, kr=3.00		0.608			
80/120	1/2	4/6	sk=1.00, kr=3.00		0.616			
80/120	1/4	4/6	sk=1.00, kr=3.00		0.612			
80/120	1/1	4/6	sk=2.00, kr=6.00		0.607			
80/120	1/2	4/6	sk=2.00, kr=6.00		0.600			
80/120	1/4	4/6	sk=2.00, kr=6.00		0.604			
120/80	1/1	4/6	sk=0.00, kr=0.00		0.605			
120/80	1/2	4/6	sk=0.00, kr=0.00		0.605			
120/80	1/4	4/6	sk=0.00, kr=0.00		0.607			
120/80	1/1	4/6	sk=1.00, kr=3.00		0.603			
120/80	1/2	4/6	sk=1.00, kr=3.00		0.596			
120/80	1/4	4/6	sk=1.00, kr=3.00		0.614			
120/80	1/1	4/6	sk=2.00, kr=6.00		0.597			
120/80	1/2	4/6	sk=2.00, kr=6.00		0.597			
120/80	1/4	4/6	sk=2.00, kr=6.00		0.609			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 34

Power Estimates for Conditions Indicating Adequate Type I Error Control (τ^2 = .33, δ=.8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00		0.636			
5/5	1/2	12/18	sk=0.00, kr=0.00					
5/5	1/4	12/18	sk=0.00, kr=0.00		0.631			
5/5	1/1	12/18	sk=1.00, kr=3.00		0.654			
5/5	1/2	12/18	sk=1.00, kr=3.00		0.670			
5/5	1/4	12/18	sk=1.00, kr=3.00		0.684			
5/5	1/1	12/18	sk=2.00, kr=6.00		0.684			
5/5	1/2	12/18	sk=2.00, kr=6.00		0.697			
5/5	1/4	12/18	sk=2.00, kr=6.00		0.700			
4/6	1/1	12/18	sk=0.00, kr=0.00		0.620			
4/6	1/2	12/18	sk=0.00, kr=0.00		0.635			
4/6	1/4	12/18	sk=0.00, kr=0.00		0.655	0.686		0.691
4/6	1/1	12/18	sk=1.00, kr=3.00		0.640			
4/6	1/2	12/18	sk=1.00, kr=3.00		0.676			
4/6	1/4	12/18	sk=1.00, kr=3.00		0.689	0.729		0.733
4/6	1/1	12/18	sk=2.00, kr=6.00		0.660			
4/6	1/2	12/18	sk=2.00, kr=6.00		0.703			
4/6	1/4	12/18	sk=2.00, kr=6.00			0.756		0.760
6/4	1/1	12/18	sk=0.00, kr=0.00		0.621			
6/4	1/2	12/18	sk=0.00, kr=0.00		0.613			
6/4	1/4	12/18	sk=0.00, kr=0.00		0.594			
6/4	1/1	12/18	sk=1.00, kr=3.00		0.654			
6/4	1/2	12/18	sk=1.00, kr=3.00		0.653			
6/4	1/4	12/18	sk=1.00, kr=3.00		0.651			
6/4	1/1	12/18	sk=2.00, kr=6.00		0.662			
6/4	1/2	12/18	sk=2.00, kr=6.00		0.675			
6/4	1/4	12/18	sk=2.00, kr=6.00		0.686			
20/20	1/1	12/18	sk=0.00, kr=0.00		0.874			
20/20	1/2	12/18	sk=0.00, kr=0.00					
20/20	1/4	12/18	sk=0.00, kr=0.00					
20/20	1/1	12/18	sk=1.00, kr=3.00		0.878			
20/20	1/2	12/18	sk=1.00, kr=3.00		0.881			
20/20	1/4	12/18	sk=1.00, kr=3.00		0.885			
20/20	1/1	12/18	sk=2.00, kr=6.00					
20/20	1/2	12/18	sk=2.00, kr=6.00					
20/20	1/4	12/18	sk=2.00, kr=6.00		0.889			
16/24	1/1	12/18	sk=0.00, kr=0.00		0.871			
16/24	1/2	12/18	sk=0.00, kr=0.00		0.874			
16/24	1/4	12/18	sk=0.00, kr=0.00		0.889			
16/24	1/1	12/18	sk=1.00, kr=3.00		0.879			
16/24	1/2	12/18	sk=1.00, kr=3.00		0.878			
16/24	1/4	12/18	sk=1.00, kr=3.00		0.894			
16/24	1/1	12/18	sk=2.00, kr=6.00		0.868			
16/24	1/2	12/18	sk=2.00, kr=6.00		0.880			
16/24	1/4	12/18	sk=2.00, kr=6.00		0.891			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.
*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Low power at sample sizes below 40 prohibited effective use of the RE and CR tests in the few conditions evidencing robustness. Permuted Q also produced low power for these conditions, until sample sizes rose to 40 and above.

Table 34 (continued)

Power Estimates for Conditions Indicating Adequate Type I Error Control (τ^2 = .33, δ=.8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00		0.869			
24/16	1/2	12/18	sk=0.00, kr=0.00		0.867			
24/16	1/4	12/18	sk=0.00, kr=0.00					
24/16	1/1	12/18	sk=1.00, kr=3.00		0.878			
24/16	1/2	12/18	sk=1.00, kr=3.00		0.876			
24/16	1/4	12/18	sk=1.00, kr=3.00		0.864			
24/16	1/1	12/18	sk=2.00, kr=6.00		0.867			
24/16	1/2	12/18	sk=2.00, kr=6.00		0.879			
24/16	1/4	12/18	sk=2.00, kr=6.00		0.875			
100/100	1/1	12/18	sk=0.00, kr=0.00		0.935			
100/100	1/2	12/18	sk=0.00, kr=0.00		0.935			
100/100	1/4	12/18	sk=0.00, kr=0.00		0.935			
100/100	1/1	12/18	sk=1.00, kr=3.00		0.937			
100/100	1/2	12/18	sk=1.00, kr=3.00					
100/100	1/4	12/18	sk=1.00, kr=3.00		0.939			
100/100	1/1	12/18	sk=2.00, kr=6.00		0.934			
100/100	1/2	12/18	sk=2.00, kr=6.00		0.939			
100/100	1/4	12/18	sk=2.00, kr=6.00		0.934			
80/120	1/1	12/18	sk=0.00, kr=0.00		0.936			
80/120	1/2	12/18	sk=0.00, kr=0.00		0.936			
80/120	1/4	12/18	sk=0.00, kr=0.00		0.937			
80/120	1/1	12/18	sk=1.00, kr=3.00					
80/120	1/2	12/18	sk=1.00, kr=3.00		0.939			
80/120	1/4	12/18	sk=1.00, kr=3.00		0.935			
80/120	1/1	12/18	sk=2.00, kr=6.00		0.932			
80/120	1/2	12/18	sk=2.00, kr=6.00		0.938			
80/120	1/4	12/18	sk=2.00, kr=6.00		0.933			
120/80	1/1	12/18	sk=0.00, kr=0.00		0.931			
120/80	1/2	12/18	sk=0.00, kr=0.00		0.930			
120/80	1/4	12/18	sk=0.00, kr=0.00		0.923			
120/80	1/1	12/18	sk=1.00, kr=3.00		0.926			
120/80	1/2	12/18	sk=1.00, kr=3.00		0.937			
120/80	1/4	12/18	sk=1.00, kr=3.00		0.928			
120/80	1/1	12/18	sk=2.00, kr=6.00		0.933			
120/80	1/2	12/18	sk=2.00, kr=6.00		0.932			
120/80	1/4	12/18	sk=2.00, kr=6.00		0.931			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Increasing K to 30 had an ameliorative effect on power for permuted Q, regardless of heterogeneity of effects at .33 (compare Table 32 to the above Table 34). As long as sample sizes exceeded the smallest level, power increased for this one test.

Table 35

Power Estimates for Conditions Indicating Adequate Type I Error Control (τ^2 = .33, δ=.8) at α=.10 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00		0.761			
5/5	1/2	12/18	sk=0.00, kr=0.00		0.753			
5/5	1/4	12/18	sk=0.00, kr=0.00		0.747			
5/5	1/1	12/18	sk=1.00, kr=3.00		0.774			
5/5	1/2	12/18	sk=1.00, kr=3.00		0.779			
5/5	1/4	12/18	sk=1.00, kr=3.00		0.800			
5/5	1/1	12/18	sk=2.00, kr=6.00		0.792			
5/5	1/2	12/18	sk=2.00, kr=6.00		0.799			
5/5	1/4	12/18	sk=2.00, kr=6.00		0.810			
4/6	1/1	12/18	sk=0.00, kr=0.00		0.747			
4/6	1/2	12/18	sk=0.00, kr=0.00		0.755	0.779		0.784
4/6	1/4	12/18	sk=0.00, kr=0.00		0.773	0.792		0.795
4/6	1/1	12/18	sk=1.00, kr=3.00		0.761	0.789		
4/6	1/2	12/18	sk=1.00, kr=3.00		0.782	0.807		0.812
4/6	1/4	12/18	sk=1.00, kr=3.00		0.799	0.821		
4/6	1/1	12/18	sk=2.00, kr=6.00		0.770			
4/6	1/2	12/18	sk=2.00, kr=6.00		0.809			
4/6	1/4	12/18	sk=2.00, kr=6.00		0.824	0.844		0.846
6/4	1/1	12/18	sk=0.00, kr=0.00		0.743	0.773		
6/4	1/2	12/18	sk=0.00, kr=0.00		0.733	0.764		
6/4	1/4	12/18	sk=0.00, kr=0.00		0.718			
6/4	1/1	12/18	sk=1.00, kr=3.00		0.771			
6/4	1/2	12/18	sk=1.00, kr=3.00		0.768			
6/4	1/4	12/18	sk=1.00, kr=3.00		0.768			
6/4	1/1	12/18	sk=2.00, kr=6.00		0.785			
6/4	1/2	12/18	sk=2.00, kr=6.00		0.787			
6/4	1/4	12/18	sk=2.00, kr=6.00		0.796			
20/20	1/1	12/18	sk=0.00, kr=0.00		0.932			
20/20	1/2	12/18	sk=0.00, kr=0.00		0.934			
20/20	1/4	12/18	sk=0.00, kr=0.00		0.932			
20/20	1/1	12/18	sk=1.00, kr=3.00		0.935			
20/20	1/2	12/18	sk=1.00, kr=3.00		0.938			
20/20	1/4	12/18	sk=1.00, kr=3.00		0.944			
20/20	1/1	12/18	sk=2.00, kr=6.00		0.931			
20/20	1/2	12/18	sk=2.00, kr=6.00		0.938			
20/20	1/4	12/18	sk=2.00, kr=6.00		0.944			
16/24	1/1	12/18	sk=0.00, kr=0.00		0.931			
16/24	1/2	12/18	sk=0.00, kr=0.00		0.932			
16/24	1/4	12/18	sk=0.00, kr=0.00		0.942			
16/24	1/1	12/18	sk=1.00, kr=3.00		0.938			
16/24	1/2	12/18	sk=1.00, kr=3.00		0.939			
16/24	1/4	12/18	sk=1.00, kr=3.00		0.944			
16/24	1/1	12/18	sk=2.00, kr=6.00		0.927			
16/24	1/2	12/18	sk=2.00, kr=6.00		0.937			
16/24	1/4	12/18	sk=2.00, kr=6.00		0.947			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The increase in nominal alpha served to increase power to a minimal extent (when sample sizes were small) for the RE and CR tests. Permuted Q's power did not change from the increase and remained high across all conditions.

Table 35 (continued)

Power Estimates for Conditions Indicating Adequate Type I Error Control (τ^2 = .33, δ=.8) at α=.10 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00		0.930			
24/16	1/2	12/18	sk=0.00, kr=0.00		0.928			
24/16	1/4	12/18	sk=0.00, kr=0.00		0.931			
24/16	1/1	12/18	sk=1.00, kr=3.00		0.933			
24/16	1/2	12/18	sk=1.00, kr=3.00		0.935			
24/16	1/4	12/18	sk=1.00, kr=3.00		0.932			
24/16	1/1	12/18	sk=2.00, kr=6.00		0.926			
24/16	1/2	12/18	sk=2.00, kr=6.00		0.935			
24/16	1/4	12/18	sk=2.00, kr=6.00		0.932			
100/100	1/1	12/18	sk=0.00, kr=0.00		0.971			
100/100	1/2	12/18	sk=0.00, kr=0.00		0.970			
100/100	1/4	12/18	sk=0.00, kr=0.00		0.968			
100/100	1/1	12/18	sk=1.00, kr=3.00		0.969			
100/100	1/2	12/18	sk=1.00, kr=3.00		0.974			
100/100	1/4	12/18	sk=1.00, kr=3.00		0.971			
100/100	1/1	12/18	sk=2.00, kr=6.00		0.970			
100/100	1/2	12/18	sk=2.00, kr=6.00		0.972			
100/100	1/4	12/18	sk=2.00, kr=6.00		0.969			
80/120	1/1	12/18	sk=0.00, kr=0.00		0.970			
80/120	1/2	12/18	sk=0.00, kr=0.00		0.970			
80/120	1/4	12/18	sk=0.00, kr=0.00		0.971			
80/120	1/1	12/18	sk=1.00, kr=3.00		0.972			
80/120	1/2	12/18	sk=1.00, kr=3.00		0.975			
80/120	1/4	12/18	sk=1.00, kr=3.00		0.970			
80/120	1/1	12/18	sk=2.00, kr=6.00		0.966			
80/120	1/2	12/18	sk=2.00, kr=6.00		0.972			
80/120	1/4	12/18	sk=2.00, kr=6.00		0.970			
120/80	1/1	12/18	sk=0.00, kr=0.00		0.968			
120/80	1/2	12/18	sk=0.00, kr=0.00		0.969			
120/80	1/4	12/18	sk=0.00, kr=0.00		0.960			
120/80	1/1	12/18	sk=1.00, kr=3.00		0.962			
120/80	1/2	12/18	sk=1.00, kr=3.00		0.966			
120/80	1/4	12/18	sk=1.00, kr=3.00		0.963			
120/80	1/1	12/18	sk=2.00, kr=6.00		0.970			
120/80	1/2	12/18	sk=2.00, kr=6.00		0.972			
120/80	1/4	12/18	sk=2.00, kr=6.00		0.967			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 36

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 1$, δ=.8) at α=.05 for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00		0.140			
5/5	1/2	4/6	sk=0.00, kr=0.00					
5/5	1/4	4/6	sk=0.00, kr=0.00					
5/5	1/1	4/6	sk=1.00, kr=3.00		0.137			
5/5	1/2	4/6	sk=1.00, kr=3.00		0.135			
5/5	1/4	4/6	sk=1.00, kr=3.00		0.152			
5/5	1/1	4/6	sk=2.00, kr=6.00		0.142			
5/5	1/2	4/6	sk=2.00, kr=6.00		0.141			
5/5	1/4	4/6	sk=2.00, kr=6.00		0.146			
4/6	1/1	4/6	sk=0.00, kr=0.00		0.149			
4/6	1/2	4/6	sk=0.00, kr=0.00		0.150			
4/6	1/4	4/6	sk=0.00, kr=0.00		0.140			
4/6	1/1	4/6	sk=1.00, kr=3.00		0.135			
4/6	1/2	4/6	sk=1.00, kr=3.00		0.147			
4/6	1/4	4/6	sk=1.00, kr=3.00		0.146			
4/6	1/1	4/6	sk=2.00, kr=6.00		0.145			
4/6	1/2	4/6	sk=2.00, kr=6.00		0.148			
4/6	1/4	4/6	sk=2.00, kr=6.00		0.158			
6/4	1/1	4/6	sk=0.00, kr=0.00		0.149			
6/4	1/2	4/6	sk=0.00, kr=0.00		0.141			
6/4	1/4	4/6	sk=0.00, kr=0.00					
6/4	1/1	4/6	sk=1.00, kr=3.00		0.153			
6/4	1/2	4/6	sk=1.00, kr=3.00		0.136			
6/4	1/4	4/6	sk=1.00, kr=3.00		0.144			
6/4	1/1	4/6	sk=2.00, kr=6.00		0.149			
6/4	1/2	4/6	sk=2.00, kr=6.00		0.146			
6/4	1/4	4/6	sk=2.00, kr=6.00		0.148			
20/20	1/1	4/6	sk=0.00, kr=0.00		0.172			
20/20	1/2	4/6	sk=0.00, kr=0.00					
20/20	1/4	4/6	sk=0.00, kr=0.00					
20/20	1/1	4/6	sk=1.00, kr=3.00					
20/20	1/2	4/6	sk=1.00, kr=3.00					
20/20	1/4	4/6	sk=1.00, kr=3.00		0.175			
20/20	1/1	4/6	sk=2.00, kr=6.00		0.172			
20/20	1/2	4/6	sk=2.00, kr=6.00		0.184			
20/20	1/4	4/6	sk=2.00, kr=6.00					
16/24	1/1	4/6	sk=0.00, kr=0.00		0.173			
16/24	1/2	4/6	sk=0.00, kr=0.00		0.178			
16/24	1/4	4/6	sk=0.00, kr=0.00		0.174			
16/24	1/1	4/6	sk=1.00, kr=3.00		0.181			
16/24	1/2	4/6	sk=1.00, kr=3.00		0.175			
16/24	1/4	4/6	sk=1.00, kr=3.00					
16/24	1/1	4/6	sk=2.00, kr=6.00					
16/24	1/2	4/6	sk=2.00, kr=6.00		0.167			
16/24	1/4	4/6	sk=2.00, kr=6.00		0.178			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

With K=10 and increased heterogeneity of effects at 1, permuted Q evidenced low power, rendering it an ineffective statistic under these conditions. Again, none of the other tests had adequate Type Ierror control. Therefore, concerns with the adequacy of power for these tests were not relevant.

Table 36 (continued)

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 1$, δ=.8) at α=.05 for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	4/6	sk=0.00, kr=0.00		0.168			
24/16	1/2	4/6	sk=0.00, kr=0.00		0.171			
24/16	1/4	4/6	sk=0.00, kr=0.00		0.165			
24/16	1/1	4/6	sk=1.00, kr=3.00		0.174			
24/16	1/2	4/6	sk=1.00, kr=3.00		0.168			
24/16	1/4	4/6	sk=1.00, kr=3.00		0.168			
24/16	1/1	4/6	sk=2.00, kr=6.00		0.174			
24/16	1/2	4/6	sk=2.00, kr=6.00		0.174			
24/16	1/4	4/6	sk=2.00, kr=6.00		0.165			
100/100	1/1	4/6	sk=0.00, kr=0.00		0.194			
100/100	1/2	4/6	sk=0.00, kr=0.00		0.195			
100/100	1/4	4/6	sk=0.00, kr=0.00		0.195			
100/100	1/1	4/6	sk=1.00, kr=3.00		0.186			
100/100	1/2	4/6	sk=1.00, kr=3.00		0.185			
100/100	1/4	4/6	sk=1.00, kr=3.00		0.187			
100/100	1/1	4/6	sk=2.00, kr=6.00		0.186			
100/100	1/2	4/6	sk=2.00, kr=6.00		0.182			
100/100	1/4	4/6	sk=2.00, kr=6.00		0.184			
80/120	1/1	4/6	sk=0.00, kr=0.00		0.190			
80/120	1/2	4/6	sk=0.00, kr=0.00		0.185			
80/120	1/4	4/6	sk=0.00, kr=0.00		0.190			
80/120	1/1	4/6	sk=1.00, kr=3.00		0.183			
80/120	1/2	4/6	sk=1.00, kr=3.00		0.186			
80/120	1/4	4/6	sk=1.00, kr=3.00		0.194			
80/120	1/1	4/6	sk=2.00, kr=6.00		0.189			
80/120	1/2	4/6	sk=2.00, kr=6.00		0.188			
80/120	1/4	4/6	sk=2.00, kr=6.00		0.199			
120/80	1/1	4/6	sk=0.00, kr=0.00		0.186			
120/80	1/2	4/6	sk=0.00, kr=0.00		0.185			
120/80	1/4	4/6	sk=0.00, kr=0.00		0.190			
120/80	1/1	4/6	sk=1.00, kr=3.00		0.198			
120/80	1/2	4/6	sk=1.00, kr=3.00		0.177			
120/80	1/4	4/6	sk=1.00, kr=3.00		0.180			
120/80	1/1	4/6	sk=2.00, kr=6.00		0.185			
120/80	1/2	4/6	sk=2.00, kr=6.00		0.182			
120/80	1/4	4/6	sk=2.00, kr=6.00		0.186			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 37

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 1$, δ=.8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00		0.389			
5/5	1/2	12/18	sk=0.00, kr=0.00		0.382			
5/5	1/4	12/18	sk=0.00, kr=0.00		0.386			
5/5	1/1	12/18	sk=1.00, kr=3.00		0.392			
5/5	1/2	12/18	sk=1.00, kr=3.00		0.383			
5/5	1/4	12/18	sk=1.00, kr=3.00		0.387			
5/5	1/1	12/18	sk=2.00, kr=6.00		0.396			
5/5	1/2	12/18	sk=2.00, kr=6.00		0.390			
5/5	1/4	12/18	sk=2.00, kr=6.00		0.398			
4/6	1/1	12/18	sk=0.00, kr=0.00		0.375			
4/6	1/2	12/18	sk=0.00, kr=0.00		0.389			
4/6	1/4	12/18	sk=0.00, kr=0.00		0.394			
4/6	1/1	12/18	sk=1.00, kr=3.00		0.385			
4/6	1/2	12/18	sk=1.00, kr=3.00					
4/6	1/4	12/18	sk=1.00, kr=3.00		0.398			
4/6	1/1	12/18	sk=2.00, kr=6.00		0.392			
4/6	1/2	12/18	sk=2.00, kr=6.00		0.403			
4/6	1/4	12/18	sk=2.00, kr=6.00		0.393			
6/4	1/1	12/18	sk=0.00, kr=0.00		0.366			
6/4	1/2	12/18	sk=0.00, kr=0.00		0.367			
6/4	1/4	12/18	sk=0.00, kr=0.00		0.362			
6/4	1/1	12/18	sk=1.00, kr=3.00		0.388			
6/4	1/2	12/18	sk=1.00, kr=3.00		0.395			
6/4	1/4	12/18	sk=1.00, kr=3.00		0.383			
6/4	1/1	12/18	sk=2.00, kr=6.00		0.370			
6/4	1/2	12/18	sk=2.00, kr=6.00		0.387			
6/4	1/4	12/18	sk=2.00, kr=6.00		0.378			
20/20	1/1	12/18	sk=0.00, kr=0.00		0.491			
20/20	1/2	12/18	sk=0.00, kr=0.00		0.492			
20/20	1/4	12/18	sk=0.00, kr=0.00		0.488			
20/20	1/1	12/18	sk=1.00, kr=3.00		0.478			
20/20	1/2	12/18	sk=1.00, kr=3.00		0.502			
20/20	1/4	12/18	sk=1.00, kr=3.00		0.485			
20/20	1/1	12/18	sk=2.00, kr=6.00		0.480			
20/20	1/2	12/18	sk=2.00, kr=6.00		0.490			
20/20	1/4	12/18	sk=2.00, kr=6.00		0.479			
16/24	1/1	12/18	sk=0.00, kr=0.00		0.482			
16/24	1/2	12/18	sk=0.00, kr=0.00		0.492			
16/24	1/4	12/18	sk=0.00, kr=0.00					
16/24	1/1	12/18	sk=1.00, kr=3.00		0.506			
16/24	1/2	12/18	sk=1.00, kr=3.00					
16/24	1/4	12/18	sk=1.00, kr=3.00		0.499			
16/24	1/1	12/18	sk=2.00, kr=6.00		0.496			
16/24	1/2	12/18	sk=2.00, kr=6.00		0.483			
16/24	1/4	12/18	sk=2.00, kr=6.00		0.493			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Though permuted Q's power improved with the increase in K at τ^2=1 and as sample size increased to 40, power was still insufficient. Low power was evident across all variants within this combination of heterogeneity of effects and effect size.

Table 37 (continued)

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 1$, $\delta=.8$) at $\alpha=.05$ for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00		0.487			
24/16	1/2	12/18	sk=0.00, kr=0.00		0.484			
24/16	1/4	12/18	sk=0.00, kr=0.00		0.487			
24/16	1/1	12/18	sk=1.00, kr=3.00		0.482			
24/16	1/2	12/18	sk=1.00, kr=3.00		0.493			
24/16	1/4	12/18	sk=1.00, kr=3.00		0.481			
24/16	1/1	12/18	sk=2.00, kr=6.00		0.480			
24/16	1/2	12/18	sk=2.00, kr=6.00		0.495			
24/16	1/4	12/18	sk=2.00, kr=6.00		0.474			
100/100	1/1	12/18	sk=0.00, kr=0.00		0.523			
100/100	1/2	12/18	sk=0.00, kr=0.00		0.524			
100/100	1/4	12/18	sk=0.00, kr=0.00		0.523			
100/100	1/1	12/18	sk=1.00, kr=3.00		0.519			
100/100	1/2	12/18	sk=1.00, kr=3.00		0.509			
100/100	1/4	12/18	sk=1.00, kr=3.00		0.525			
100/100	1/1	12/18	sk=2.00, kr=6.00		0.522			
100/100	1/2	12/18	sk=2.00, kr=6.00		0.515			
100/100	1/4	12/18	sk=2.00, kr=6.00		0.530			
80/120	1/1	12/18	sk=0.00, kr=0.00		0.532			
80/120	1/2	12/18	sk=0.00, kr=0.00		0.524			
80/120	1/4	12/18	sk=0.00, kr=0.00		0.518			
80/120	1/1	12/18	sk=1.00, kr=3.00		0.527			
80/120	1/2	12/18	sk=1.00, kr=3.00		0.518			
80/120	1/4	12/18	sk=1.00, kr=3.00		0.533			
80/120	1/1	12/18	sk=2.00, kr=6.00		0.525			
80/120	1/2	12/18	sk=2.00, kr=6.00		0.532			
80/120	1/4	12/18	sk=2.00, kr=6.00		0.522			
120/80	1/1	12/18	sk=0.00, kr=0.00		0.523			
120/80	1/2	12/18	sk=0.00, kr=0.00		0.508			
120/80	1/4	12/18	sk=0.00, kr=0.00		0.535			
120/80	1/1	12/18	sk=1.00, kr=3.00		0.519			
120/80	1/2	12/18	sk=1.00, kr=3.00		0.525			
120/80	1/4	12/18	sk=1.00, kr=3.00		0.529			
120/80	1/1	12/18	sk=2.00, kr=6.00		0.517			
120/80	1/2	12/18	sk=2.00, kr=6.00		0.515			
120/80	1/4	12/18	sk=2.00, kr=6.00		0.523			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 38

Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 1$, δ=.8) at α=.10 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00					
5/5	1/2	12/18	sk=0.00, kr=0.00		0.516			
5/5	1/4	12/18	sk=0.00, kr=0.00		0.513			
5/5	1/1	12/18	sk=1.00, kr=3.00		0.518			
5/5	1/2	12/18	sk=1.00, kr=3.00		0.514			
5/5	1/4	12/18	sk=1.00, kr=3.00		0.519			
5/5	1/1	12/18	sk=2.00, kr=6.00		0.530			
5/5	1/2	12/18	sk=2.00, kr=6.00		0.534			
5/5	1/4	12/18	sk=2.00, kr=6.00		0.525			
4/6	1/1	12/18	sk=0.00, kr=0.00		0.509			
4/6	1/2	12/18	sk=0.00, kr=0.00		0.520			
4/6	1/4	12/18	sk=0.00, kr=0.00		0.521			
4/6	1/1	12/18	sk=1.00, kr=3.00		0.516			
4/6	1/2	12/18	sk=1.00, kr=3.00		0.530			
4/6	1/4	12/18	sk=1.00, kr=3.00		0.531			
4/6	1/1	12/18	sk=2.00, kr=6.00		0.524			
4/6	1/2	12/18	sk=2.00, kr=6.00		0.528			
4/6	1/4	12/18	sk=2.00, kr=6.00		0.528			
6/4	1/1	12/18	sk=0.00, kr=0.00		0.504			
6/4	1/2	12/18	sk=0.00, kr=0.00					
6/4	1/4	12/18	sk=0.00, kr=0.00		0.494			
6/4	1/1	12/18	sk=1.00, kr=3.00		0.525			
6/4	1/2	12/18	sk=1.00, kr=3.00		0.526			
6/4	1/4	12/18	sk=1.00, kr=3.00		0.515			
6/4	1/1	12/18	sk=2.00, kr=6.00		0.503			
6/4	1/2	12/18	sk=2.00, kr=6.00		0.517			
6/4	1/4	12/18	sk=2.00, kr=6.00		0.513			
20/20	1/1	12/18	sk=0.00, kr=0.00		0.628			
20/20	1/2	12/18	sk=0.00, kr=0.00		0.627			
20/20	1/4	12/18	sk=0.00, kr=0.00		0.618			
20/20	1/1	12/18	sk=1.00, kr=3.00		0.618			
20/20	1/2	12/18	sk=1.00, kr=3.00		0.631			
20/20	1/4	12/18	sk=1.00, kr=3.00		0.614			
20/20	1/1	12/18	sk=2.00, kr=6.00		0.619			
20/20	1/2	12/18	sk=2.00, kr=6.00		0.628			
20/20	1/4	12/18	sk=2.00, kr=6.00		0.611			
16/24	1/1	12/18	sk=0.00, kr=0.00		0.615			
16/24	1/2	12/18	sk=0.00, kr=0.00		0.631			
16/24	1/4	12/18	sk=0.00, kr=0.00		0.627			
16/24	1/1	12/18	sk=1.00, kr=3.00		0.627			
16/24	1/2	12/18	sk=1.00, kr=3.00		0.620			
16/24	1/4	12/18	sk=1.00, kr=3.00		0.629			
16/24	1/1	12/18	sk=2.00, kr=6.00		0.626			
16/24	1/2	12/18	sk=2.00, kr=6.00		0.614			
16/24	1/4	12/18	sk=2.00, kr=6.00		0.621			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The increase in nominal alpha to .10 promoted further slight improvements in permuted Q's power, but still these increases in power remained inadequate for appropriate use under all of the τ^2=1 conditions.

Table 38 (continued)
Power Estimates for Conditions Indicating Adequate Type I Error Control ($\tau^2 = 1$, δ=.8) at α=.10 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00		0.620			
24/16	1/2	12/18	sk=0.00, kr=0.00		0.620			
24/16	1/4	12/18	sk=0.00, kr=0.00		0.623			
24/16	1/1	12/18	sk=1.00, kr=3.00		0.615			
24/16	1/2	12/18	sk=1.00, kr=3.00		0.627			
24/16	1/4	12/18	sk=1.00, kr=3.00		0.606			
24/16	1/1	12/18	sk=2.00, kr=6.00		0.620			
24/16	1/2	12/18	sk=2.00, kr=6.00		0.628			
24/16	1/4	12/18	sk=2.00, kr=6.00		0.610			
100/100	1/1	12/18	sk=0.00, kr=0.00		0.650			
100/100	1/2	12/18	sk=0.00, kr=0.00		0.662			
100/100	1/4	12/18	sk=0.00, kr=0.00		0.654			
100/100	1/1	12/18	sk=1.00, kr=3.00		0.646			
100/100	1/2	12/18	sk=1.00, kr=3.00		0.641			
100/100	1/4	12/18	sk=1.00, kr=3.00		0.658			
100/100	1/1	12/18	sk=2.00, kr=6.00		0.660			
100/100	1/2	12/18	sk=2.00, kr=6.00		0.650			
100/100	1/4	12/18	sk=2.00, kr=6.00		0.660			
80/120	1/1	12/18	sk=0.00, kr=0.00		0.661			
80/120	1/2	12/18	sk=0.00, kr=0.00		0.659			
80/120	1/4	12/18	sk=0.00, kr=0.00		0.651			
80/120	1/1	12/18	sk=1.00, kr=3.00		0.647			
80/120	1/2	12/18	sk=1.00, kr=3.00		0.655			
80/120	1/4	12/18	sk=1.00, kr=3.00		0.675			
80/120	1/1	12/18	sk=2.00, kr=6.00		0.663			
80/120	1/2	12/18	sk=2.00, kr=6.00		0.658			
80/120	1/4	12/18	sk=2.00, kr=6.00		0.655			
120/80	1/1	12/18	sk=0.00, kr=0.00		0.661			
120/80	1/2	12/18	sk=0.00, kr=0.00		0.646			
120/80	1/4	12/18	sk=0.00, kr=0.00		0.660			
120/80	1/1	12/18	sk=1.00, kr=3.00		0.650			
120/80	1/2	12/18	sk=1.00, kr=3.00		0.649			
120/80	1/4	12/18	sk=1.00, kr=3.00		0.655			
120/80	1/1	12/18	sk=2.00, kr=6.00		0.651			
120/80	1/2	12/18	sk=2.00, kr=6.00		0.651			
120/80	1/4	12/18	sk=2.00, kr=6.00		0.656			

Note: Shaded areas indicate those conditions for which a test demonstrated good power to detect true differences.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Summary of Power Estimates Results

When τ^2=0 and K=10 at nominal alpha, .05, power was low for each of the 5 tests in all conditions evidencing adequate Type I error, when sample sizes were small. With the increase in K to 30, the FE test offered the same number of effective, powerful conditions as with K=10 (see Table 30). But the RE test manifested an increased number of effective conditions for use (24 as compared to 19 in the previous K=10 conditions), primarily occurring in the equal and first group with greater sample size conditions. The CR test presented no change in the number of effective conditions and these were concentrated in the equal

sample size conditions. Permuted Q had an increased number of effective conditions both with the conditions with smaller and larger sample size.

With the increase in τ^2 to .33 at K=10 for nominal alpha, .05, sufficient power was absent for all tests, including permuted Q. Power remained limited for permuted Q, regardless of increases in sample size. With the introduction of effect sizes greater than 0, even permuted Q's effectiveness sharply deteriorated due to losses in power. The increase in nominal alpha to .10 for the τ^2=.33 condition at K=10 did not improve the power for permuted Q, the RE and CR tests. Both the regular Q and the FE test possessed good power when the primary study sample sizes were 40 or greater, but lacked robustness. The combined increase in heterogeneity and small K suppressed power for all of the tests. This pattern was evident as heterogeneity increased to 1 for both nominal alpha .05 and .10 when K=10.

Increasing K to 30 (with alpha set to .05) had an ameliorative effect on power for permuted Q, regardless of heterogeneity of effects being equal to .33. As long as sample sizes exceeded the smallest level, this test's power increased. The RE and CR tests showed slight improvement to robustness with the increase in effect size to .8. However, low power at sample sizes below 40 prohibited effective use of these tests in the conditions evidencing robustness.

The increase in nominal alpha served to increase power to a minimal extent (when sample sizes were small) for the RE and CR tests. Instead, the increase in heterogeneity of effects continued to suppress power for these tests, regardless of any additional changes in any of the factors investigated. Table 23 illustrates that power was still insufficient, despite a minimal increase, to permit effective use of the RE and CR tests. Permuted Q's power was enhanced only slightly from the increase in K=30 for the same τ^2 and remained high across all conditions. Therefore, the increase in K to 30 appeared to have more of an ameliorative impact than any other factor for increasing permuted Q's power at τ^2 =.33.

Once heterogeneity of effects increased to 1, permuted Q's power reduced dramatically. Regardless of any and all variations of K, nominal alpha and sample size, permuted Q's power remained low. None of the other tests attained adequate robustness for power to become a consideration. In summary, none of the tests offered any degree of effectiveness as heterogeneity of effects increased to 1 or when sample sizes were less than 40 with nominal α set to .05 (and K=10). Though permuted Q yielded

adequate Type I error control for all conditions investigated, it proved to be susceptible to Type II error under these conditions.

Power Estimates for Conditions Indicating Both Robustness and Good Power

This second series of results illustrate those conditions for which each test provided both adequate Type I error control and good power. Low power values were excluded. These tables (Tables 39-42) originated as Type I error estimate tables for each of the true null conditions with a true effect of .8. Tables were constructed only for those conditions in which at least one test obtained both adequate type I error control and good power for at least one condition. Once those conditions reflecting adequate robustness were identified and highlighted, data reflecting only good power for those conditions was presented. Therefore, these tables best illustrate optimal effectiveness of the five tests being evaluated.

Generally, adequate power is defined as probability estimates (corresponding false null estimates for the same τ^2 and δ combination) of .80 or better. This means that in 80% of the simulations, the test detected true differences between groups. Power estimates are only relevant for those conditions where a test has demonstrated adequate control of Type I error. As mentioned previously, only those conditions evidencing both robustness and good power by a test were included in the next set of tables. Therefore, the following tables (Tables 39-42) provide a picture of the conditions for optimal effective use of each test being evaluated.

Table 39

Power Estimates for Conditions Indicating Both Robustness & Good Power (τ^2= 0, δ= .8) at α=.05 for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	4/6	sk=0.00, kr=0.00					
5/5	1/2	4/6	sk=0.00, kr=0.00					
5/5	1/4	4/6	sk=0.00, kr=0.00					
5/5	1/1	4/6	sk=1.00, kr=3.00					
5/5	1/2	4/6	sk=1.00, kr=3.00					
5/5	1/4	4/6	sk=1.00, kr=3.00					
5/5	1/1	4/6	sk=2.00, kr=6.00					
5/5	1/2	4/6	sk=2.00, kr=6.00					
5/5	1/4	4/6	sk=2.00, kr=6.00					
4/6	1/1	4/6	sk=0.00, kr=0.00					
4/6	1/2	4/6	sk=0.00, kr=0.00					
4/6	1/4	4/6	sk=0.00, kr=0.00					
4/6	1/1	4/6	sk=1.00, kr=3.00					
4/6	1/2	4/6	sk=1.00, kr=3.00					
4/6	1/4	4/6	sk=1.00, kr=3.00					
4/6	1/1	4/6	sk=2.00, kr=6.00					
4/6	1/2	4/6	sk=2.00, kr=6.00					
4/6	1/4	4/6	sk=2.00, kr=6.00					
6/4	1/1	4/6	sk=0.00, kr=0.00					
6/4	1/2	4/6	sk=0.00, kr=0.00					
6/4	1/4	4/6	sk=0.00, kr=0.00					
6/4	1/1	4/6	sk=1.00, kr=3.00					
6/4	1/2	4/6	sk=1.00, kr=3.00					
6/4	1/4	4/6	sk=1.00, kr=3.00					
6/4	1/1	4/6	sk=2.00, kr=6.00					
6/4	1/2	4/6	sk=2.00, kr=6.00					
6/4	1/4	4/6	sk=2.00, kr=6.00					
20/20	1/1	4/6	sk=0.00, kr=0.00		0.917		0.966	0.962
20/20	1/2	4/6	sk=0.00, kr=0.00		0.917	0.957	0.965	0.958
20/20	1/4	4/6	sk=0.00, kr=0.00		0.904		0.969	0.961
20/20	1/1	4/6	sk=1.00, kr=3.00		0.912	0.962	0.970	0.963
20/20	1/2	4/6	sk=1.00, kr=3.00		0.925			
20/20	1/4	4/6	sk=1.00, kr=3.00		0.945			
20/20	1/1	4/6	sk=2.00, kr=6.00		0.917	0.960		0.961
20/20	1/2	4/6	sk=2.00, kr=6.00					
20/20	1/4	4/6	sk=2.00, kr=6.00		0.970			

Note: Empty shaded cells indicate those conditions with adequate Type I error control, but low power. Unshaded cells indicate conditions with poor Type I error control. Shaded cells containing data have both good Type I error and good power.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Though each of the tests presented at least one occasion of robustness when sample sizes were small, none of the tests, including permuted Q possessed power enough for effective application. After sample sizes increased to 40 and above, the RE, FE and CR tests all showed enhanced power when sample sizes were equal (particularly the FE and CR tests) and when the first group had a larger sample size than the second (particularly for the RE test).

Table 39 (continued)
Power Estimates for Conditions Indicating Both Robustness & Good Power (τ^2= 0, δ= .8) at α=.05 for K=10

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
16/24	1/1	4/6	sk=0.00, kr=0.00		0.904			
16/24	1/2	4/6	sk=0.00, kr=0.00					
16/24	1/4	4/6	sk=0.00, kr=0.00		0.937			
16/24	1/1	4/6	sk=1.00, kr=3.00		0.908	0.956		0.957
16/24	1/2	4/6	sk=1.00, kr=3.00		0.937			
16/24	1/4	4/6	sk=1.00, kr=3.00		0.959			
16/24	1/1	4/6	sk=2.00, kr=6.00		0.911	0.952		
16/24	1/2	4/6	sk=2.00, kr=6.00		0.956			
16/24	1/4	4/6	sk=2.00, kr=6.00		0.976			
24/16	1/1	4/6	sk=0.00, kr=0.00		0.907		0.961	
24/16	1/2	4/6	sk=0.00, kr=0.00		0.885	0.948		
24/16	1/4	4/6	sk=0.00, kr=0.00		0.870			
24/16	1/1	4/6	sk=1.00, kr=3.00		0.913	0.956	0.966	0.957
24/16	1/2	4/6	sk=1.00, kr=3.00			0.965	0.973	0.966
24/16	1/4	4/6	sk=1.00, kr=3.00		0.905	0.960	0.974	0.962
24/16	1/1	4/6	sk=2.00, kr=6.00		0.913	0.959		
24/16	1/2	4/6	sk=2.00, kr=6.00	0.828	0.935		0.979	0.974
24/16	1/4	4/6	sk=2.00, kr=6.00		0.948			
100/100	1/1	4/6	sk=0.00, kr=0.00	1.000	1.000			
100/100	1/2	4/6	sk=0.00, kr=0.00	1.000	1.000	1.000	1.000	1.000
100/100	1/4	4/6	sk=0.00, kr=0.00		1.000	1.000		1.000
100/100	1/1	4/6	sk=1.00, kr=3.00			1.000		1.000
100/100	1/2	4/6	sk=1.00, kr=3.00		1.000		1.000	
100/100	1/4	4/6	sk=1.00, kr=3.00		1.000			
100/100	1/1	4/6	sk=2.00, kr=6.00		1.000			
100/100	1/2	4/6	sk=2.00, kr=6.00		1.000			
100/100	1/4	4/6	sk=2.00, kr=6.00					
80/120	1/1	4/6	sk=0.00, kr=0.00	1.000		1.000	1.000	1.000
80/120	1/2	4/6	sk=0.00, kr=0.00		1.000			
80/120	1/4	4/6	sk=0.00, kr=0.00		1.000			
80/120	1/1	4/6	sk=1.00, kr=3.00		1.000	1.000		1.000
80/120	1/2	4/6	sk=1.00, kr=3.00		1.000			
80/120	1/4	4/6	sk=1.00, kr=3.00		1.000			
80/120	1/1	4/6	sk=2.00, kr=6.00		1.000			
80/120	1/2	4/6	sk=2.00, kr=6.00		1.000			
80/120	1/4	4/6	sk=2.00, kr=6.00		1.000			
120/80	1/1	4/6	sk=0.00, kr=0.00	1.000	1.000		1.000	1.000
120/80	1/2	4/6	sk=0.00, kr=0.00		1.000	1.000		
120/80	1/4	4/6	sk=0.00, kr=0.00		1.000			
120/80	1/1	4/6	sk=1.00, kr=3.00		1.000	1.000		1.000
120/80	1/2	4/6	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
120/80	1/4	4/6	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
120/80	1/1	4/6	sk=2.00, kr=6.00		1.000			
120/80	1/2	4/6	sk=2.00, kr=6.00	1.000	1.000		1.000	
120/80	1/4	4/6	sk=2.00, kr=6.00		1.000			

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Permuted Q consistently offered the majority of effective conditions across all factors of variance, sample size and population shape, once sample size increased to 40 and above.

Table 40

Power Estimates for Conditions Indicating Both Robustness & Good Power (τ^2= 0, δ= .8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00		0.869			
5/5	1/2	12/18	sk=0.00, kr=0.00		0.877			
5/5	1/4	12/18	sk=0.00, kr=0.00		0.872			
5/5	1/1	12/18	sk=1.00, kr=3.00					
5/5	1/2	12/18	sk=1.00, kr=3.00		0.916			
5/5	1/4	12/18	sk=1.00, kr=3.00		0.931			
5/5	1/1	12/18	sk=2.00, kr=6.00		0.930		0.928	
5/5	1/2	12/18	sk=2.00, kr=6.00		0.959			
5/5	1/4	12/18	sk=2.00, kr=6.00		0.967			
4/6	1/1	12/18	sk=0.00, kr=0.00		0.865			
4/6	1/2	12/18	sk=0.00, kr=0.00		0.880			
4/6	1/4	12/18	sk=0.00, kr=0.00		0.891			
4/6	1/1	12/18	sk=1.00, kr=3.00					
4/6	1/2	12/18	sk=1.00, kr=3.00		0.930			
4/6	1/4	12/18	sk=1.00, kr=3.00		0.942			
4/6	1/1	12/18	sk=2.00, kr=6.00		0.923			
4/6	1/2	12/18	sk=2.00, kr=6.00		0.962			
4/6	1/4	12/18	sk=2.00, kr=6.00		0.977			
6/4	1/1	12/18	sk=0.00, kr=0.00		0.870			
6/4	1/2	12/18	sk=0.00, kr=0.00		0.836			
6/4	1/4	12/18	sk=0.00, kr=0.00		0.814	0.836		0.839
6/4	1/1	12/18	sk=1.00, kr=3.00		0.902			
6/4	1/2	12/18	sk=1.00, kr=3.00		0.904			
6/4	1/4	12/18	sk=1.00, kr=3.00		0.907		0.926	
6/4	1/1	12/18	sk=2.00, kr=6.00		0.924			
6/4	1/2	12/18	sk=2.00, kr=6.00		0.949			
6/4	1/4	12/18	sk=2.00, kr=6.00		0.953			
20/20	1/1	12/18	sk=0.00, kr=0.00		1.000		1.000	1.000
20/20	1/2	12/18	sk=0.00, kr=0.00	0.046	1.000		1.000	1.000
20/20	1/4	12/18	sk=0.00, kr=0.00			1.000	1.000	1.000
20/20	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
20/20	1/2	12/18	sk=1.00, kr=3.00		1.000			
20/20	1/4	12/18	sk=1.00, kr=3.00		1.000			
20/20	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
20/20	1/2	12/18	sk=2.00, kr=6.00		1.000			
20/20	1/4	12/18	sk=2.00, kr=6.00		1.000			

Note: Empty shaded cells indicate those conditions with adequate Type I error control, but low power. Unshaded cells indicate conditions with poor Type I error control. Shaded cells containing data have both good Type I error and good power.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The greatest difference from Table 39 to 40 (with the increase in K to 30) was that permuted Q provided effective use conditions when sample sizes were small. The number of effective conditions yielded by the regular Q, RE, FE and CR tests increased when sample sizes reached 40 and above.

Table 40 (continued)

Power Estimates for Conditions Indicating Both Robustness & Good Power (τ^2= 0, δ= .8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
16/24	1/1	12/18	sk=0.00, kr=0.00		1.000		1.000	
16/24	1/2	12/18	sk=0.00, kr=0.00		1.000			
16/24	1/4	12/18	sk=0.00, kr=0.00		1.000			
16/24	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
16/24	1/2	12/18	sk=1.00, kr=3.00					
16/24	1/4	12/18	sk=1.00, kr=3.00		1.000			
16/24	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
16/24	1/2	12/18	sk=2.00, kr=6.00		1.000			
16/24	1/4	12/18	sk=2.00, kr=6.00		1.000			
24/16	1/1	12/18	sk=0.00, kr=0.00	0.978	1.000			
24/16	1/2	12/18	sk=0.00, kr=0.00		1.000	1.000		1.000
24/16	1/4	12/18	sk=0.00, kr=0.00		1.000	1.000		
24/16	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
24/16	1/2	12/18	sk=1.00, kr=3.00		1.000		1.000	1.000
24/16	1/4	12/18	sk=1.00, kr=3.00		1.000	1.000		1.000
24/16	1/1	12/18	sk=2.00, kr=6.00		1.000			
24/16	1/2	12/18	sk=2.00, kr=6.00		1.000	1.000	1.000	1.000
24/16	1/4	12/18	sk=2.00, kr=6.00		1.000			
100/100	1/1	12/18	sk=0.00, kr=0.00	1.000	1.000	1.000	1.000	1.000
100/100	1/2	12/18	sk=0.00, kr=0.00	1.000	1.000	1.000		1.000
100/100	1/4	12/18	sk=0.00, kr=0.00			1.000	1.000	1.000
100/100	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000		1.000
100/100	1/2	12/18	sk=1.00, kr=3.00		1.000		1.000	1.000
100/100	1/4	12/18	sk=1.00, kr=3.00		1.000			
100/100	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
100/100	1/2	12/18	sk=2.00, kr=6.00		1.000			
100/100	1/4	12/18	sk=2.00, kr=6.00		1.000			
80/120	1/1	12/18	sk=0.00, kr=0.00	1.000				
80/120	1/2	12/18	sk=0.00, kr=0.00		1.000			
80/120	1/4	12/18	sk=0.00, kr=0.00		1.000			
80/120	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000		
80/120	1/2	12/18	sk=1.00, kr=3.00					
80/120	1/4	12/18	sk=1.00, kr=3.00		1.000			
80/120	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
80/120	1/2	12/18	sk=2.00, kr=6.00		1.000			
80/120	1/4	12/18	sk=2.00, kr=6.00		1.000			
120/80	1/1	12/18	sk=0.00, kr=0.00	1.000	1.000		1.000	
120/80	1/2	12/18	sk=0.00, kr=0.00		1.000	1.000		
120/80	1/4	12/18	sk=0.00, kr=0.00		1.000			
120/80	1/1	12/18	sk=1.00, kr=3.00		1.000	1.000		1.000
120/80	1/2	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
120/80	1/4	12/18	sk=1.00, kr=3.00		1.000	1.000	1.000	1.000
120/80	1/1	12/18	sk=2.00, kr=6.00		1.000	1.000		
120/80	1/2	12/18	sk=2.00, kr=6.00	1.000				
120/80	1/4	12/18	sk=2.00, kr=6.00		1.000			

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The power of all tests was optimal in relationship to performance for all other combinations of heterogeneity of effects and effect size once K=30 and sample sizes were 40 or above.

Table 41

Power Estimates for Conditions Indicating Both Robustness & Good Power (τ^2= .33, δ= .8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00					
5/5	1/2	12/18	sk=0.00, kr=0.00					
5/5	1/4	12/18	sk=0.00, kr=0.00					
5/5	1/1	12/18	sk=1.00, kr=3.00					
5/5	1/2	12/18	sk=1.00, kr=3.00					
5/5	1/4	12/18	sk=1.00, kr=3.00					
5/5	1/1	12/18	sk=2.00, kr=6.00					
5/5	1/2	12/18	sk=2.00, kr=6.00					
5/5	1/4	12/18	sk=2.00, kr=6.00					
4/6	1/1	12/18	sk=0.00, kr=0.00					
4/6	1/2	12/18	sk=0.00, kr=0.00					
4/6	1/4	12/18	sk=0.00, kr=0.00					
4/6	1/1	12/18	sk=1.00, kr=3.00					
4/6	1/2	12/18	sk=1.00, kr=3.00					
4/6	1/4	12/18	sk=1.00, kr=3.00					
4/6	1/1	12/18	sk=2.00, kr=6.00					
4/6	1/2	12/18	sk=2.00, kr=6.00					
4/6	1/4	12/18	sk=2.00, kr=6.00					
6/4	1/1	12/18	sk=0.00, kr=0.00					
6/4	1/2	12/18	sk=0.00, kr=0.00					
6/4	1/4	12/18	sk=0.00, kr=0.00					
6/4	1/1	12/18	sk=1.00, kr=3.00					
6/4	1/2	12/18	sk=1.00, kr=3.00					
6/4	1/4	12/18	sk=1.00, kr=3.00					
6/4	1/1	12/18	sk=2.00, kr=6.00					
6/4	1/2	12/18	sk=2.00, kr=6.00					
6/4	1/4	12/18	sk=2.00, kr=6.00					
20/20	1/1	12/18	sk=0.00, kr=0.00		0.874			
20/20	1/2	12/18	sk=0.00, kr=0.00					
20/20	1/4	12/18	sk=0.00, kr=0.00					
20/20	1/1	12/18	sk=1.00, kr=3.00		0.878			
20/20	1/2	12/18	sk=1.00, kr=3.00		0.881			
20/20	1/4	12/18	sk=1.00, kr=3.00		0.885			
20/20	1/1	12/18	sk=2.00, kr=6.00					
20/20	1/2	12/18	sk=2.00, kr=6.00					
20/20	1/4	12/18	sk=2.00, kr=6.00		0.889			

Note: Empty shaded cells indicate those conditions with adequate Type I error control, but low power. Unshaded cells indicate conditions with poor Type I error control. Shaded cells containing data have both good Type I error and good power.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The increase in heterogeneity of effects to .33 (even with K=30) minimized the effectiveness of all tests, particularly with the small sample size of 10. Once sample sizes increased to 40 and above, only permuted Q demonstrated robustness and power. However, power was reduced, though sufficient for effectiveness for most conditions.

Table 41 (continued)
Power Estimates for Conditions Indicating Both Robustness & Good Power (τ^2= .33, δ= .8) at α=.05 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
16/24	1/1	12/18	sk=0.00, kr=0.00		0.871			
16/24	1/2	12/18	sk=0.00, kr=0.00		0.874			
16/24	1/4	12/18	sk=0.00, kr=0.00		0.889			
16/24	1/1	12/18	sk=1.00, kr=3.00		0.879			
16/24	1/2	12/18	sk=1.00, kr=3.00		0.878			
16/24	1/4	12/18	sk=1.00, kr=3.00		0.894			
16/24	1/1	12/18	sk=2.00, kr=6.00		0.868			
16/24	1/2	12/18	sk=2.00, kr=6.00		0.880			
16/24	1/4	12/18	sk=2.00, kr=6.00		0.891			
24/16	1/1	12/18	sk=0.00, kr=0.00		0.869			
24/16	1/2	12/18	sk=0.00, kr=0.00		0.867			
24/16	1/4	12/18	sk=0.00, kr=0.00					
24/16	1/1	12/18	sk=1.00, kr=3.00		0.878			
24/16	1/2	12/18	sk=1.00, kr=3.00		0.876			
24/16	1/4	12/18	sk=1.00, kr=3.00		0.864			
24/16	1/1	12/18	sk=2.00, kr=6.00		0.867			
24/16	1/2	12/18	sk=2.00, kr=6.00		0.879			
24/16	1/4	12/18	sk=2.00, kr=6.00		0.875			
100/100	1/1	12/18	sk=0.00, kr=0.00		0.935			
100/100	1/2	12/18	sk=0.00, kr=0.00		0.935			
100/100	1/4	12/18	sk=0.00, kr=0.00		0.935			
100/100	1/1	12/18	sk=1.00, kr=3.00		0.937			
100/100	1/2	12/18	sk=1.00, kr=3.00					
100/100	1/4	12/18	sk=1.00, kr=3.00		0.939			
100/100	1/1	12/18	sk=2.00, kr=6.00		0.934			
100/100	1/2	12/18	sk=2.00, kr=6.00		0.939			
100/100	1/4	12/18	sk=2.00, kr=6.00		0.934			
80/120	1/1	12/18	sk=0.00, kr=0.00		0.936			
80/120	1/2	12/18	sk=0.00, kr=0.00		0.936			
80/120	1/4	12/18	sk=0.00, kr=0.00		0.937			
80/120	1/1	12/18	sk=1.00, kr=3.00					
80/120	1/2	12/18	sk=1.00, kr=3.00		0.939			
80/120	1/4	12/18	sk=1.00, kr=3.00		0.935			
80/120	1/1	12/18	sk=2.00, kr=6.00		0.932			
80/120	1/2	12/18	sk=2.00, kr=6.00		0.938			
80/120	1/4	12/18	sk=2.00, kr=6.00		0.933			
120/80	1/1	12/18	sk=0.00, kr=0.00		0.931			
120/80	1/2	12/18	sk=0.00, kr=0.00		0.930			
120/80	1/4	12/18	sk=0.00, kr=0.00		0.923			
120/80	1/1	12/18	sk=1.00, kr=3.00		0.926			
120/80	1/2	12/18	sk=1.00, kr=3.00		0.937			
120/80	1/4	12/18	sk=1.00, kr=3.00		0.928			
120/80	1/1	12/18	sk=2.00, kr=6.00		0.933			
120/80	1/2	12/18	sk=2.00, kr=6.00		0.932			
120/80	1/4	12/18	sk=2.00, kr=6.00		0.931			

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 42

Power Estimates for Conditions Indicating Both Robustness & Good Power (τ^2= .33, δ= .8) at α=.10 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
5/5	1/1	12/18	sk=0.00, kr=0.00					
5/5	1/2	12/18	sk=0.00, kr=0.00					
5/5	1/4	12/18	sk=0.00, kr=0.00					
5/5	1/1	12/18	sk=1.00, kr=3.00					
5/5	1/2	12/18	sk=1.00, kr=3.00					
5/5	1/4	12/18	sk=1.00, kr=3.00		0.800			
5/5	1/1	12/18	sk=2.00, kr=6.00					
5/5	1/2	12/18	sk=2.00, kr=6.00		0.799			
5/5	1/4	12/18	sk=2.00, kr=6.00		0.810			
4/6	1/1	12/18	sk=0.00, kr=0.00					
4/6	1/2	12/18	sk=0.00, kr=0.00					
4/6	1/4	12/18	sk=0.00, kr=0.00					0.795
4/6	1/1	12/18	sk=1.00, kr=3.00					
4/6	1/2	12/18	sk=1.00, kr=3.00			0.807		0.812
4/6	1/4	12/18	sk=1.00, kr=3.00		0.799	0.821		
4/6	1/1	12/18	sk=2.00, kr=6.00					
4/6	1/2	12/18	sk=2.00, kr=6.00		0.809			
4/6	1/4	12/18	sk=2.00, kr=6.00		0.824	0.844		0.846
6/4	1/1	12/18	sk=0.00, kr=0.00					
6/4	1/2	12/18	sk=0.00, kr=0.00					
6/4	1/4	12/18	sk=0.00, kr=0.00					
6/4	1/1	12/18	sk=1.00, kr=3.00					
6/4	1/2	12/18	sk=1.00, kr=3.00					
6/4	1/4	12/18	sk=1.00, kr=3.00					
6/4	1/1	12/18	sk=2.00, kr=6.00					
6/4	1/2	12/18	sk=2.00, kr=6.00					
6/4	1/4	12/18	sk=2.00, kr=6.00		0.796			
20/20	1/1	12/18	sk=0.00, kr=0.00		0.932			
20/20	1/2	12/18	sk=0.00, kr=0.00		0.934			
20/20	1/4	12/18	sk=0.00, kr=0.00		0.932			
20/20	1/1	12/18	sk=1.00, kr=3.00		0.935			
20/20	1/2	12/18	sk=1.00, kr=3.00		0.938			
20/20	1/4	12/18	sk=1.00, kr=3.00		0.944			
20/20	1/1	12/18	sk=2.00, kr=6.00		0.931			
20/20	1/2	12/18	sk=2.00, kr=6.00		0.938			
20/20	1/4	12/18	sk=2.00, kr=6.00		0.944			
16/24	1/1	12/18	sk=0.00, kr=0.00		0.931			
16/24	1/2	12/18	sk=0.00, kr=0.00		0.932			
16/24	1/4	12/18	sk=0.00, kr=0.00		0.942			
16/24	1/1	12/18	sk=1.00, kr=3.00		0.938			
16/24	1/2	12/18	sk=1.00, kr=3.00		0.939			
16/24	1/4	12/18	sk=1.00, kr=3.00		0.944			
16/24	1/1	12/18	sk=2.00, kr=6.00		0.927			
16/24	1/2	12/18	sk=2.00, kr=6.00		0.937			
16/24	1/4	12/18	sk=2.00, kr=6.00		0.947			

Note: Empty shaded cells indicate those conditions with adequate Type I error control, but low power. Unshaded cells indicate conditions with poor Type I error control. Shaded cells containing data have both good Type I error and good power.

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

Table 42 (continued)

Power Estimates for Conditions Indicating Both Robustness & Good Power (τ^2= .33, δ= .8) at α=.10 for K=30

Primary Study Sample Sizes	Population Variances	N of studies	Population Shape	Reg Q	PQ_b	RE	FE	CR
24/16	1/1	12/18	sk=0.00, kr=0.00		0.930			
24/16	1/2	12/18	sk=0.00, kr=0.00		0.928			
24/16	1/4	12/18	sk=0.00, kr=0.00		0.931			
24/16	1/1	12/18	sk=1.00, kr=3.00		0.933			
24/16	1/2	12/18	sk=1.00, kr=3.00		0.935			
24/16	1/4	12/18	sk=1.00, kr=3.00		0.932			
24/16	1/1	12/18	sk=2.00, kr=6.00		0.926			
24/16	1/2	12/18	sk=2.00, kr=6.00		0.935			
24/16	1/4	12/18	sk=2.00, kr=6.00		0.932			
100/100	1/1	12/18	sk=0.00, kr=0.00		0.971			
100/100	1/2	12/18	sk=0.00, kr=0.00		0.970			
100/100	1/4	12/18	sk=0.00, kr=0.00		0.968			
100/100	1/1	12/18	sk=1.00, kr=3.00		0.969			
100/100	1/2	12/18	sk=1.00, kr=3.00		0.974			
100/100	1/4	12/18	sk=1.00, kr=3.00		0.971			
100/100	1/1	12/18	sk=2.00, kr=6.00		0.970			
100/100	1/2	12/18	sk=2.00, kr=6.00		0.972			
100/100	1/4	12/18	sk=2.00, kr=6.00		0.969			
80/120	1/1	12/18	sk=0.00, kr=0.00		0.970			
80/120	1/2	12/18	sk=0.00, kr=0.00		0.970			
80/120	1/4	12/18	sk=0.00, kr=0.00		0.971			
80/120	1/1	12/18	sk=1.00, kr=3.00		0.972			
80/120	1/2	12/18	sk=1.00, kr=3.00		0.975			
80/120	1/4	12/18	sk=1.00, kr=3.00		0.970			
80/120	1/1	12/18	sk=2.00, kr=6.00		0.966			
80/120	1/2	12/18	sk=2.00, kr=6.00		0.972			
80/120	1/4	12/18	sk=2.00, kr=6.00		0.970			
120/80	1/1	12/18	sk=0.00, kr=0.00		0.968			
120/80	1/2	12/18	sk=0.00, kr=0.00		0.969			
120/80	1/4	12/18	sk=0.00, kr=0.00		0.960			
120/80	1/1	12/18	sk=1.00, kr=3.00		0.962			
120/80	1/2	12/18	sk=1.00, kr=3.00		0.966			
120/80	1/4	12/18	sk=1.00, kr=3.00		0.963			
120/80	1/1	12/18	sk=2.00, kr=6.00		0.970			
120/80	1/2	12/18	sk=2.00, kr=6.00		0.972			
120/80	1/4	12/18	sk=2.00, kr=6.00		0.967			

*Reg Q=Regular Q; PQ_b=Permuted Q Between; RE=Random-effects Z; FE=Fixed-effects Z; CR=Conditionally Random Procedure

The increase in nominal alpha to .10 slightly facilitated the effectiveness of the permuted Q, RE and CR tests when sample sizes were small. But only permuted Q retained power and robustness when sample sizes were 40 and above. Permuted Q's effectiveness was high and slightly better with increased nominal alpha, .10.

Power Estimates Summary

In brief, increasing K tended to enhance the power of all of the tests, particularly permuted Q. Changing this variable in combination with changing τ^2 appeared to have some unique effect in that $\tau^2 = 0$ at K=10 resulted in several conditions for completely effective use of all tests. But only once K was elevated to 30 for $\tau^2 = .33$ did the permuted Q test work effectively (and then only as effectively as when $\tau^2 = 0$ at K=10, and not as effectively as when K=30). Raising nominal alpha from .05 to .10 mitigated this influence of increasing τ^2 on power, particularly for permuted Q. Furthermore, power and Type I error control improved as primary study sample sizes increased from 10 to 40. This influence was most salient when $\tau^2 = 0$. Surprisingly, the opposite pattern arose when heterogeneity of effects was .33 and nominal alpha was .10 at K=30. Effectiveness improved for the RE and CR tests at small sample sizes only.

When $\tau^2 = 0$ and alpha =.05, an increase in K from 10 to 30 (see Tables 39 and 40) most significantly affected permuted Q's power. When K=10, 48 of the 81 conditions had both adequate Type I error and good power. At K=30, the permuted Q not only demonstrated adequate Type I error, but also increased capacity for good power (78 of the 81 conditions). The other 4 tests (Regular Q, RE, FE and CR) produced far fewer conditions with both adequate Type I error and good power. The number of such conditions did not change appreciably from K=10 to 30, as with permuted Q (see Tables 39 and 40). Only the RE test produced an increase in the number of effective conditions (from 19 to 24) under these conditions.

With respect to the combination of $\tau^2 = .33$, $\delta = .8$, nominal alpha=.05 at K=10, a table of power estimates is not presented due to the fact that none of the conditions for any of the tests, including permuted Q, exhibited both good power and adequate Type I error control. Only once K was increased to 30 for the same set of conditions (see Table 41) did permuted Q yield both adequate Type I error and good power. Permuted Q's performance with this set of conditions was comparable to its performance at $\tau^2 = 0$, $\delta = .8$ at K=10 (Table 39) in that sufficient power was absent when sample sizes were small. Though good power was absent, the RE and CR tests had a few conditions for which each enabled adequate control of Type I error.

In order to determine the effect of nominal α, if any, alpha was elevated from .05 to .10 for the combination of τ^2 =.33, δ=.8 at K=10 and K=30. The combination at K=10 did not result in any conditions with adequate Type I error and good power. But at K=30, there was some positive effect on the power of permuted Q, and to a lesser extent for the RE and CR tests. When α = .05 under the same constraints, neither RE nor CR provided adequate Type I error control or good power for any of the conditions. Permuted Q produced 47 such conditions, as compared with 61 conditions under K=30 (see Tables 39 and 40).

Answers to Research Questions

1. To what extent is the Type I error rate of the fixed-effects Q (FE), permuted Q, random-effects (RE) and conditionally-random procedure (CR) maintained near the nominal alpha level across variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness and kurtosis?

a. Small K tended to inflate Type I error for all of the tests (except for Permuted Q), particularly as τ^2 increased.
b. Fluctuations in the number of primary studies influenced the Type I error rates as K varied from 10 to 30. There tended to be greater inflation of Type I error rates when the primary study sample sizes were small, especially when K was smaller.
c. Variance within primary studies did not inject significant changes in the median rejection rates of the RE, CR and Permuted Q tests. Variance had the greatest influence on the distribution of rejection rates for the regular Q test.
d. Heterogeneity of effects (τ^2) had the greatest impact on the performance of each of the tests. As τ^2 increased from 0 to .33, Type I error became increasingly inflated. Only Permuted Q was unaffected by the influence of varying the τ^2 values, as it constrained the margin error rates to nominal alpha, .05.
e. Skewness and kurtosis in isolation did not tend to contribute to inflated Type I error for any of the tests.

2. What is the relative statistical power of the fixed-effects Q (FE), permuted Q, random-effects (RE) and conditionally-random procedure (CR) given variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness and kurtosis?

a. Small K diminished power for all tests, including permuted Q.
b. As primary study sample sizes increased from 10 to 40, power tended to increase for all tests.
c. Differing the variance within primary studies did not exert notable influences on the power of any of the tests.
d. Increasing τ^2 did not enhance power for these tests (see Tables 30 and 32; Tables 31 and 34). This pattern continued whether K equaled 10 or 30.
e. Varying skewness and kurtosis had no remarkable effect on power.

Chapter Five

Interpretations and Conclusions

Following is a summary of the present research, including a review and discussion of the results, consideration of the limitations of the research design, and recommendations for future studies.

Summary

The purpose of this study was to investigate the power and Type I error control of the permuted Q, random-effects Z test, fixed-effects Z test, conditionally-random procedure and regular Q test under varying levels of heterogeneity of effects (τ^2=0, τ^2= .33, and τ^2=1) and at α level .05, as well as three variance ratios, two different K and N=10, 40 and 200. The relative effectiveness of these three tests (and one conditionally-random procedure) was compared under varying conditions of K, variance within groups, primary study sample sizes within meta-analytic studies, population shape, as well as τ^2. The comparison of the relative performance of these three tests of homogeneity of effects and the conditionally-random procedure within each set of controlled conditions should enhance the appropriateness of practitioners' test selection for meta-analysis.

The research questions propelling this investigation were the following:

1. To what extent is the Type I error rate of the regular fixed-effects Q (regQ), permuted Q_{bet}, fixed-effects Z (FE), random-effects Z (RE) and conditionally-random procedure (CR) maintained near the nominal alpha level across variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness/kurtosis?
2. What is the relative statistical power of the regular fixed-effects Q (regQ), permuted Q_{bet}, fixed-effects Z (FE), random-effects Z (RE) and conditionally-random procedure (CR) given variations in the number of primary studies included in the meta-analysis, sample sizes in primary studies, heterogeneity of variance, varying degrees of heterogeneity of effects (τ^2) and primary study skewness/kurtosis?

Results of each of the 1,458 experimental conditions arising from the factoring of seven independent variables across each of the three tests of homogeneity and conditionally-random procedure were presented. Effectiveness of each test was evaluated based on the proportion of the 5000 simulations of each meta-analytic condition reflecting adequate Type I error control at the nominal alpha level of .05. As defined by Bradley (1978), rejections above .055 are termed "inflated", whereas those empirical values below .045 are termed "conservative" for nominal α=.05. For nominal alpha level .10, rejections above .11 are "inflated", while those rejections below .09 are conservative.

This study is modeled after Harwell (1997) and Kromrey and Hogarty's (1998) experimental design. Specifically, it entails a 2 x 3 x 3 x 3 x 3 x 3 x 2 factorial design. The study also controls for between-studies variance, as suggested by Hedges and Vevea's (1998) study. The randomized factorial design includes seven independent variables: (1) number of studies within the meta-analysis (10 and 30); (2) primary study sample size (10, 40, 200); (3) score distribution skewness and kurtosis (0/0; 1/3; 2/6;); (4) equal or unequal (around typical sample sizes, 1:1; 4:6; and 6:4) within-group sample sizes; (5) equal or unequal group variances (1:1; 2:1; and 4:1;); (6) between-studies variance, τ^2 (0, .33, and 1); and (7) between-class effect size differences, δ_k (0 and .8). Data were obtained using two programs: one for null hypotheses (972 simulations) and the other for non-null hypotheses (486 simulations). Hence, the study incorporated 1,458 experimental conditions, illustrated in Figure 1. Simulated data from each sample were analyzed using each of five tests of homogeneity (includes one procedure).

The dependent variable is, in part, the proportion of conditions with adequate Type I error control at the nominal alpha level of .05. Although, not an original consideration of the study, the performance of each of these tests under nominal alpha level .10 was investigated under τ^2= .33 and 1.0. Additionally, estimates of statistical power were computed for those conditions where tests maintained adequate Type I error control. These power estimates indicate the degree to which a test reflects sensitivity to significant heterogeneity of effects, in the presence of violated assumptions.

There were nine hundred seventy-two simulated data conditions, consisting of six sets of null conditions. These sets (81 conditions per set) of null conditions entailed the following: τ^2 =0, δ=0; τ^2 =0, δ=.8; τ^2 =.33, δ=0; τ^2 =.33, δ=.8; τ^2 =1, δ=0; τ^2 =1, δ=.8. Each of these sets of conditions was submitted to a further condition, varying K. K equaled 10 for one (total number=486) and 30 (total

number=486) for the next. Additionally, 486 data conditions (three sets) were generated, utilizing the non-null program. These three sets consisted of the following: τ^2 =0, δ=.8; τ^2 =.33, δ=.8; and τ^2 =1, δ=.8. Again, each set of 81 conditions was simulated, assuming K=10 for one set and K=30 for the other. Five thousand iterations were simulated for each condition and an average rejection rate was calculated.

Results are presented as box and whisker plots of the Type I error rates, the proportion of simulations with adequate Type I error, average Type I error rates, and power value rates for each given test. The proportion of simulations was computed by adding up the number of conditions with adequate Type I error and then dividing that frequency by the number of conditions. Average Type I error rates were derived from the average of all error rates for a particular condition. Power value rates were based on the non-null conditions. When experimental conditions exhibited Type I error control, based on Bradley's criterion for each nominal alpha, power analyses were completed. First, power detection rates for each non-null condition were deemed either within good power limits (estimates greater than .795) or too low. For those conditions displaying both adequate Type I error control and good power, it was concluded that a particular test or tests demonstrated effectiveness under that set of conditions.

Results define the extent of Type I error control and the comparable degree of power of the five tests being investigated (regular Q serves as a baseline, as its purpose is somewhat different from the other tests). The effect of varying the within-study variance and population shape were of particular concern, as it has been demonstrated that increasing heterogeneity of variance within studies and increased skewness and kurtosis led to greater inflation of Type I error (Harwell, 1997; and Kromrey & Hogarty, 1998). Also of interest, Chang (1993) had noted that few had addressed the issue of Type II error with respect to regular Q. Increasing K with a small N in the primary studies had resulted in increased inflation of the Type I error when applying regular Q (Hedges & Vevea, 1998). No subsequent studies had examined the relative influence of increased heterogeneity of effects (τ^2) on the performance of regular Q, permuted Q, the RE test, FE test and the conditionally-random procedure. Lastly, the effect of altering the nominal alpha level from the commonly applied, .05, to a more liberal .10 had not been investigated in a comparative analysis of these five tests, with respect to control of Type I error or degree of power. Therefore, the overall purpose in conducting these analyses is to provide practitioners with guidance as to when best to apply each of the five tests under a given set of conditions.

As demonstrated previously, the permuted Q maintained the greatest degree of robustness under each set of conditions. However, it was surprising to find that it did not evidence good power under increasing heterogeneity of effects.

As illustrated by the box and whisker plots, changes in the population shape did not have an appreciable effect on Type I error, as previously reported (Harwell, 1997; Kromrey & Hogarty, 1998). As skewness/kurtosis increased from normal to more skewed and leptokurtotic, no significant changes were evident. Though inflated Type I error is evident for all tests but permuted Q, the degree of inflation did not appear to change appreciably from normal to more skewed/kurtotic extremes.

Consistent with the findings of the Hedges and Vevea (1998) study, the results of this investigation further demonstrate that the combination of increasing K with small N renders an increase in Type I error. When either N increases to 40 or above or K increases to 30, Type I error rates tended to decrease for all tests. As mentioned above, permuted Q was the only test relatively unaffected by these changes alone. Note that this finding is in isolation to the increase in heterogeneity of effects.

Investigating the issue of power, provisioned on the adequacy of Type I error control, presents a more complete picture of the effectiveness of each test. It was surprising to note how often adequate Type I error was not further augmented by corresponding good power, particularly with respect to the permuted Q. Conversely, many of the other four tests demonstrated insufficient control of Type I error.

As heterogeneity of effects increased from 0 to .33, there was a dramatic effect on the power of each of the tests. As τ^2 varied from 0 to .33, K played an integral role in the maintenance of Type I error control for all tests, except permuted Q. However, the same pattern proved relevant to the power of permuted Q. The effect from .33 to 1.0 was even more pronounced.

The effect of applying a more liberal nominal alpha did little to enhance the effectiveness of the tests. Under increasing heterogeneity of effects, applying a more liberal nominal alpha (from .05 to .10) permitted the permuted Q to achieve greater power (Type I error was well-maintained regardless of nominal alpha). But the actual increase in the frequency of effective conditions was minimal. Referring to Table 43, there were increases across nominal alpha levels at the K=30, τ^2=.33 condition. Permuted Q, RE and CR tests increased in effectiveness from 0 to low effectiveness (less than 25% of the conditions showed effective application of that specific test), when alpha went from .05 to .10.

Generally, the permuted Q maintained adequate Type I error control under all conditions of varying sample sizes within primary studies, population shape, variance within-studies, effect size (from 0 to .8) and K. Additionally, it did appear that increasing K had a positive effect on permuted Q's power. When the effectiveness of permuted Q did decline, it was due to low power.

Table 43
Effective Application of Five Meta-analytic Tests of Homogeneity for True Null Conditions

	α=.05		RegQ	PQB	RE	FE	CR		α=.10		RegQ	PQB	RE	FE	CR
K=10	τ^2=0, δ=.8	N<40	0	0	0	0	0	K=10	τ^2=0, δ=.8	N<40	0	0	0	0	0
		N>40	low	high	med	med	med			N>40	low	high	med	med	med
	τ^2=.33, δ=.8	N<40	0	0	0	0	0		τ^2=.33, δ=.8	N<40	0	0	0	0	0
		N>40	0	0	0	0	0			N>40	0	0	0	0	0
	τ^2=1, δ=.8	N<40	0	0	0	0	0		τ^2=1, δ=.8	N<40	0	0	0	0	0
		N>40	0	0	0	0	0			N>40	0	0	0	0	0
K=30	τ^2=0, δ=.8	N<40	0	high	low	low	low	K=30	τ^2=0, δ=.8	N<40	0	high	low	low	low
		N>40	low	high	med	med	med			N>40	low	high	med	med	med
	τ^2=.33, δ=.8	N<40	0	0	0	0	0		τ^2=.33, δ=.8	N<40	0	low	low	0	low
		N>40	0	high	0	0	0			N>40	0	high	0	0	0
	τ^2=1, δ=.8	N<40	0	0	0	0	0		τ^2=1, δ=.8	N<40	0	0	0	0	0
		N>40	0	0	0	0	0			N>40	0	0	0	0	0

low=low frequency for effectiveness-less than 25% of the conditions, but more than 0
high=high frequency for effectiveness-more than 75% of the conditions
med=medium frequency for effectiveness-between 25% and 75% of the conditions
0= Type I error rate is outside Bradley's criterion for robustness and/or power was low for all conditions

Note: effectiveness is determined when a test exhibits both adequate Type I error control and good power for a given condition.

The RE test and conditionally –random procedure generally exhibited the next greatest degree of robustness. In contrast to the RE and CR tests, both the FE and regular Q tests were more sensitive in terms of maintaining Type I error control under most changes in each of the treatment effects. When effectiveness diminished for any of these four tests, it was generally due to a lack of robustness.

Discussion

The results with respect to the overall robustness of permuted Q are consistent with previous research (Kromrey & Hogarty, 1998; and Hogarty & Kromrey, 1999). Over all of the varied conditions, permuted Q enabled sufficient Type I error control. But, there was an unexpected lack of power from one level of heterogeneity of effects to the next. Power diminished as heterogeneity of effects increased.

There was limited enhanced power for permuted Q as nominal alpha became more liberal from .05 to .10, evidenced at K=30, τ^2 =.33. The RE and CR tests also displayed similar small increases in frequency of effectiveness at this same level of heterogeneity of effects.

The introduction of heterogeneous variance tended to inflate Q's Type I error rate (Hogarty & Kromrey, 1999). In the present study, within-studies variance also impacted the frequency of conditions with Type I error control for the RE, FE and CR tests, particularly for τ^2 from 0 to .33. In general, as variance ratios increased, the frequency of conditions evidencing control of Type I error decreased. When τ^2 = 0, margin error rates for these 3 tests tended to be conservative. As τ^2 increased from 0 to .33, margin error rates for these 3 tests tended to inflate. As will be discussed, as heterogeneity of effects increased, Type I error control diminished substantially.

More specifically, when variance was equal, the greatest frequency of well-controlled Type I error conditions occurred for all 3 tests, regardless of whether K=10 or 30 for the τ^2 = 0, δ=.8 set of conditions (see table 4). Proportions of conditions with adequate Type I error for the RE, FE and CR tests were .41, .33 and .48, respectively for K=10. At K=30, proportions were .44, .22 and .26. Permuted Q presented better than 85% of its conditions with adequate Type I error at each of the variance ratios. This pattern occurred consistently for permuted Q across all variance ratios and for all combinations of heterogeneity of effects and delta. As for the other tests, few proportions greater than zero were evidenced for any other conditions beyond the τ^2 = 0, δ=.8 and τ^2 = 0, δ=0 (refer both to Table 43 and proportion tables 5-8).

As increases in variance ratios were introduced, the frequency of conditions with well-controlled Type I error for each of these tests fell consistently, especially for K=10. At K=30, the CR test had the same frequency of conditions with well-maintained Type I error control at variance ratios of 1/1 and 1 /2, with a drop at variance 1/4. The RE test manifested a notable drop in frequency of these well-controlled conditions from 1/1 to 1 /2, with a minimal increase at 1/4. The FE test exhibited a continued decrease in frequency of these conditions from the normal to more extreme variance conditions.

At τ^2 = .33 (when K=30) with a liberal nominal alpha of .10, the RE test had a slight increase in the frequency of conditions with well-maintained Type I error control as within-study variance increased (see table 33). The CR test had an equal number of well-maintained Type I error conditions at 1/2 and 1/4 and no conditions with well-maintained Type I error at the equal variance condition.

The combined effect of heterogeneous variances and unequal sample sizes was to reduce the power of any permutation test (Hogarty & Kromrey, 1999). But when the larger sample originated from the group with the larger variance, permutation tests had the highest power. Power was diminished for equal sample sizes or for negative pairings of sample size and population variance. Kromrey & Hogarty (1998) found these negative pairings of sample size and variance (where the first sample size is larger than that of the second group) result in increasingly inflated Type I error, as compared to the condition when there are equal sample sizes between groups and unequal variances. Alternatively, Harwell (1997) and Kromrey and Hogarty (1998) found that positive pairings (small unequal variance and small unequal sample sizes) resulted in conservative Type I error rates.

When N was 40 or greater, tests attained good power only with larger δ (Hogarty & Kromrey, 1999). Because only two effect sizes (0 and .8) were applied, this current study cannot shed additional light on changes in δ. However, in general, the number of conditions where any of the tests exhibited good power increased as the total sample size increased from 10 to 40, particularly for τ^2 =0 or .33.

The present study also demonstrated that the frequency of conditions with good power increased as either K increased or nominal alpha was expanded from .05 to .10, for all tests (for τ^2 =.33). At τ^2 =1, tests providing sufficient control of Type I error (the RE, CR, and permuted Q tests) also presented low power, regardless of whether K increased or nominal alpha was liberalized. Therefore, none of these tests would be effectively applied under these conditions.

The relationship between K and within-study sample size (N) has been discussed at length (Hedges & Vevea, 1998). Specifically for the regular Q, when N is small (less than 40) and constant, Type I error control deteriorates as K increases. The present study seems to bear out this result. But other tests, like Permuted Q, RE, and CR, Type I error control improves as K increases for the τ^2 =0 condition (when α=.05) and to a more limited extent for τ^2 =.33 (when α=.10).

Referring to Table 43, one can determine that if N was less than 40 with K less than 30, power and Type I error control were curtailed, even when τ^2 =0. But as K increased to 30, keeping all else equal, both power and Type I error control improved dramatically for all tests but regular Q. The ameliorative effect arose whether nominal alpha equaled .05 or .10 (for τ^2 =0).

Table 43 represents a compilation of the power estimate and power detection tables. It provides an overview of recommendations for the appropriate application of each of the tests evaluated by this study. The number of conditions was counted to derive a frequency for each distinct group of conditions (each τ^2 by δ by N and K group) for the two nominal alphas. Further clarification can be obtained from the power estimate and power detection tables.

Further evidence of the N to K relationship is evident, when τ^2 was raised to .33. In general, as τ^2 increased, Type I error increased. Only once K increased to 30 and alpha elevated to .10 (see Table 43) did robustness improve for all tests but permuted Q. When τ^2 =.33 (for α=.05) and sample size increased to 40 and above, permuted Q's effectiveness increased from 0 to high. At nominal alpha=.10, robustness showed a decrement from low effectiveness to 0 as sample size (N) increased to 40 and above. This pattern contradicted what has been reported in the literature as the impact on regular Q's performance. Also in contrast, permuted Q's effectiveness improved from low to medium, as N increased to 40 and above.

Moreover, Type I error control continued to erode for all tests but permuted Q as τ^2 increased to 1. Again, only an increase in nominal alpha to .10 ameliorated Type I error control for the RE and CR tests as τ^2 increased to .33, and this improvement was minimal (see Table 43). Permuted Q showed the greatest improvement as τ^2 increased to .33 and N increased to 40 (nominal alpha=.05).

In contrast to Hedges and Vevea's (1998) findings, but consistent with the results found by Kromrey and Hogarty (1998), permuted Q was the only statistic to provide rejection rates closely approximating nominal α, across all conditions. Even under conditions of extreme heterogeneity of effects (τ^2), permuted Q tended to maintain adequate Type I error control more often than either the RE or CR tests. However, as discovered by Hedges and Vevea, both of these latter tests did provide superior Type I error control (though still exceeding nominal α), as compared to either the regular Q or FE tests. Furthermore, Chang (1993) found that for Q, as K increased and sample size decreased, greater departures arose between the theoretical and simulated power. More specifically, Type I error results from Q's use under these conditions.

Increasing K has been thought to exacerbate the influence of heterogeneous effects (τ^2). In particular, Hedges and Vevea (1998) suggest that under such conditions RE and CR tests produce

conservative rejection rates. Chang (1993) contends that K had no such effect on the RE test, although total sample size within the primary studies did. In the present study, as the heterogeity of effects (τ^2) increased, RE and CR's rejection rates remained more conservative than Q's. However, the Type I error rate still exceeded nominal alpha when (τ^2) >0. This result occurred whether δ equaled 0 or .8. As (τ^2) increased from 0 to 1, while holding δ at 0, all of the tests, except permuted Q, produced inflated Type I error.

Practitioners responsible for evaluating treatments investigated through meta-analytic procedures seem faced with the choice of (a) applying a permuted test which though robust under most circumstances does not always exhibit good power, particularly under increasingly heterogeneous effects or (b) utilizing a test which does not maintain adequate control of Type I error across a wide variety of circumstances.

Researchers determined to evaluate treatments investigated by meta-analytic methods are advised to utilize statistical tests in the following manner:

When K =10, for τ^2=0 at nominal α=.05, permuted Q can produce both good power and a consistently well-controlled Type I error rate. The RE, FE and CR tests performed effectively in fewer than half of the cases. Similar performance can be expected at α=.10, holding all other factors constant.

Unfortunately, none of the 5 tests investigated proved effective for use when K=10 and sample size was less than 40. Specifically, all of the tests, but permuted Q, were not robust to Type I error and power was limited. Permuted Q, though robust, was ineffective due to low power. This problem was exacerbated as τ^2 increased to anything above 0. The same outcome occurred whether nominal alpha was .05 or .10.

As K increased to 30, all tests, except the regular Q, displayed a marked increase in effectiveness from K=10, at τ^2=0. Again, permuted Q evidenced the highest degree of effectiveness in terms of frequency of conditions where both Type I error was well-controlled and good power demonstrated. Again, the RE, FE and CR tests proved to be limited in the effectiveness when N was less than 40, whether nominal alpha equaled .05 or .10. As sample size increased to 40 and higher, effectiveness improved for all tests. Again, permuted Q manifested the greatest degree of effectiveness. Performance was comparable across alpha levels.

Once τ^2 increased to .33 (nominal alpha .05), none of the tests were effective until sample size increased to 40. And then, only permuted Q provided both effective detection of true differences and maintained robustness to residual variance. As τ^2 increased to .33 (nominal alpha .10) and sample size was less than 40, the permuted Q, RE and CR tests all demonstrated equal, but limited, effectiveness. This outcome is somewhat unusual in that typically more conditions evidence effective control of Type I error and good power as sample size increases. In this case, the RE and CR test actually proved to be more effective when sample sizes were small. As sample size increased to 40 and above, only the permuted Q (for both nominal alpha .05 and .10) demonstrated any effectiveness.

None of the tests were effective in providing both good power and robustness to Type I error as τ^2 increased to 1. This result was manifest whether nominal alpha was set to .05 or .10.

Hedges & Vevea (1998) recommend applying the random-effects model (RE or CR test) when K is large and generalizations are to be made to a broader universe of studies, as it provides more conservative estimates of the effects. Unfortunately, applying the random-effects test presents one with the dilemma of having a low frequency of effectiveness when N is less than 40 at τ^2=0, K=30 and medium frequency of effectiveness when N is 40 or greater under the same conditions. (It should be noted that this pattern of results was evident whether nominal alpha was set at .05 or .10.) However, this approach may be appropriate if one decides a priori that the generalizations they will be making go beyond the immediate sample or a sample with simple, well-defined characteristics. The random-effects approach provided no practical utility when τ^2 exceeded 0, unless nominal alpha was set to .10, N was less than 40, and τ^2=.33. The frequency of effectiveness was minimal, however, even under the latter set of conditions.

Permuted Q can best be applied to conditions similar to those in which the RE and CR tests can be effectively applied, but the investigator has decided *a priori* to draw generalizations more limited to the immediate sample of studies. The permuted Q has the added advantage of greater Type I error control under a wider variety of conditions than any of the other tests. Therefore, there are a greater number of instances in which the permuted Q would demonstrate greater robustness to the commission of Type I errors as well as a greater tendency to detect true differences. Lastly, the permuted Q can also be effectively applied when τ^2 increases to .33, as long as K is greater than 10. As with the other tests

investigated, it did not demonstrate practical utility when τ^2 increased to .33 or 1 at K=10 (whether nominal alpha=.05 or .10) and when $\tau^2 = 1$ (whether nominal alpha =.05 or .10) at K=30.

Limitations

Normality of effects is not being investigated. The random-effects test, for example, requires a normal distribution of effects (Raudenbush, 1994, p.317). Even a single outlying effect can introduce bias into the computation of Q. Only sensitivity analysis, wherein the outlier(s) are omitted, can accurately predict the consequences of skewed distributions of effects.

As regression is used after the computation of Q to compute the relationship of potential moderator variables and the effect size, there may be a problem with model overfit (Raudenbush, 1994). Model overfit occurs when there are multiple predictor or independent variables to a single or just a few criterion or dependent variables. This study incorporated 30 studies to a single dependent variable. Each of these studies introduces a multitude of variables not necessarily common to all of the studies in the collection. Having such a large K to dependent variable ratio could be problematic in the calculation of regression weights using Weighted Least Squares (typically completed after Q is computed and heterogeneity is determined).

This study investigated the effects of several varying conditions on the performance of five tests of homogeneity as applied to unconditional inferences only. However, most researchers still draw conclusions appropriate to the selection of unconditional inferences while inconsistently applying the fixed-effects model and corresponding tests (National Research Council, 1992). More specifically, unconditional inferences pertain to situations where one extends conclusions about a sample's performance, given some treatment, to a larger (usually less well-defined) population. Meta-analysts do not typically limit their conclusions about the performance of a particular sample to that sample. Such an approach severely curtails the generalization of their work, prohibiting the development of heuristic theory. Moreover, there is a common understanding that one can no longer assume the collection of a completely representative sample. With the prevalent use of the internet and other automated data sources, it is no longer possible to select a sample from a well-delimited population of studies. Therefore, conclusions about a sample must account for the added uncertainty inherent in the relationship of that sample to some larger, more abstract population to permit valid generalization.

Other potentially influential conditions were not examined. There was a handful of sample sizes and distribution shapes under investigation in the present study. Related to this issue, there are innumerable potential factors influencing the effect size in any given study (Fern & Monroe, 1996). For this reason, one can never be absolutely certain the variance is attributable to the independent variables under investigation.

Differing conditions change the robustness of the statistical properties. As Overton (1998) suggests, "...one of the most important considerations in selecting a meta-analysis model is the contextual conditions in which the effect of interest (e.g., a selection test's validity) is to be generalized in theory or in application" (p. 376). In other words, a statistic is only meaningful given the assumptions applicable to the situation in which it is being used. No single statistic can be expected to be robust to all conditions.

Furthermore, realistic data often entail widely differing data characteristics, not modeled by generated data sets. Computer programs simulated data to be analyzed within this Monte Carlo study in which distribution shapes, study sample sizes, extent of variance across studies and random variation in true effects have been controlled. Moreover, the models defining the statistical tools used were methodically alternated. For this reason, these results should be confirmed by direct application of these tests to data collected in actual educational settings.

Topics for Additional Research

Future research is required to consider the effect of heterogeneity of effects on the use of trimmed-d with non-normal populations (Hogarty & Kromrey, 1999). As many Federally-funded school programs are targeted to populations where negatively skewed performance is common, it would be important to understand how other statistics like the trimmed-d operate when incorporating various degrees of heterogeneity of effects. It is in such school programs where evaluation of treatment programs must take into account the unique characteristics of the performance-related data.

Additional study is warranted to determine the extent of the problem of model overfit in the recalculation of the regression once Q has been computed. If overfit is a problem, what measures will need to be taken in order to ensure proper computation of the regression weights? As the regression weights are required to compute the relationship of potential moderator variables and the effect size, the

problem of model overfit is an important issue in the final evaluation of treatment programs.

An investigation of the use of effect size indices other than the two-group sort incorporated in the same tests of homogeneity would be useful. School programs frequently organize students in one of several groups according to general ability level, thereby classifying students in more than two groups. Therefore, evaluating these tests of homogeneity using two-group effect size indices limits practitioners' understanding of how effectively these tests operate in realistic settings.

As schools do not have the luxury of consistently restricting class sizes to a given number, it would be worthwhile to evaluate the introduction of unequal sample sizes within any given condition. In this way, the behavior of each of the tests of homogeneity can be more realistically evaluated.

References

Abelson, R.P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik and J.H. Steiger (Eds.) *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 117-141.

Bangert-Drowns, R.L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99*, 3, 388-99.

Becker, B.J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics, 17, 4*, 341-362.

Becker, B.J. (1994). Combining significance levels. In *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Bollen, K.A. 1989. *Structural Equations with Latent Variables*. New York: John Wiley & Sons, Inc.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48, 3*, 378-399.

Chang, L. (1993). Power analysis *of the test of homogeneity in effect size meta-analysis*. Unpublished doctoral dissertation, Michigan State University, East Lansing.

Chatterjee, S. & Price, B. 1977. *Regression Analysis by Example*. New York: John Wiley & Sons.

Chelimsky, E. (1998). The role of experience in formulating theories of evaluation practice. *American Journal of Evaluation, 19, 1*, 35-55.

Chow, S.L. (1998). Precis of statistical significance: Rationale, validity, and utility. *Behavioral And Brain Sciences, 21*, 169-239.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45, 12*, 1304-1312.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences, Second Edition* . New Jersey: Lawrence Erlbaum Associates, Publishers.

Cooper, H. & Hedges, L.V. (Eds.) 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Cronbach, L.J., Gleser, G.C., Harinder, N., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 116-30.

Draper & Smith (1998). *Applied Regression Analysis, 3rd edition*.

Effron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37, 1*, 36-48.

Erez, A., Bloom, M.C. & Wells, M.T. (1996). Using random rather than fixed effects models in meta-analysis: implications for situational specificity and validity generalization. *Personnel Psychology, 49*, 275-306.

Fisher, R.A. (1959). *Statistical methods and scientific inference*. Edinburgh: Oliver &Boyd.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 4, 521-532.

Fleiss, J.L., & Gross, A.J. (1991). Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique. *Journal of Clinical Epidemiology*, *44*, 127-139.

Fowler, R.L. (1988). Estimating the standardized mean difference in intervention studies. *Journal of Educational Statistics*, *13*, *4*, 337-50.

Glass, G.V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*, 351-79.

Glass, G.V. & Hopkins, K. D. (1984). *Statistical Methods in Education and Psychology*, 2nd edition. Boston: Allyn & Bacon.

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-Analysis in Social Research*. Beverly Hills: Sage Publications.

Glass, G.V., Peckham, P.D. & Sanders, J.R.(1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-88.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer-Verlag.

Greenhouse, J.B. & Iyengar, S. (1994).

Harwell, M.R. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods*, *2*, 2, 219-31.

Harwell, M. (1997). An investigation of the Raudenbush (1988) test for studying variance heterogeneity. *The Journal of Experimental Education*, *65*, *2*, 181-190.

Harwell, M.R. (1992). Summarizing monte carlo results in methodological research. *Journal of Educational Statistics*, *17*, 4, 297-313.

Harwell, M.R., Rubinstein, E.N., Hayes, W.S., & Olds, C.C. (1992). Summarizing monte carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*, p. 315-39.

Hauck, W.W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician*,*38*, 3, 214-16.

Hayes, W.S., & Olds, C.C. (1992). Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*, 315-39.

Hedges, L.V. (1982). Fitting continuous models to effect size data. *Journal of Educational Statistics*, 14, 245-70.

Hedges, L.V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*, 2, 388-95.

Hedges, L.V. (1992). Meta-analysis. *Journal of Educational Statistics, 17, 4*, 279-96.

Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando: Academic Press, Inc.

Hedges, L.V. & Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3, 4*, 486-504.

Hoaglin, D.C. & Andrews, D.F. (1975). The reporting of computation-based results in statistics. *The American Statistician, 29, 3*, 122-26.

Hopkins, K.D. & Weeks, D.L. (1990). Tests for normality and measures of skewness and kurtosis: their place in research reporting. *Educational and Psychological Measurement, 50*, 717-29.

Inman, H.F. (1994). Karl Pearson and R.A. Fisher on statistical tests: A 1935 exchange from *Nature*. *The American Statistician, 48 , 1*, 2-10.

Kennedy, J.J. & Bush, A.J. (1985). *An Introduction to the Design and Analysis of Experiments in Behavioral Research*. Lanham:University Press of America.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C. & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analyses. *Review of Educational Research, 68, 3*, 350-86.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 5, 746-59.

Kraemer, H.C. & Andrews, G.A. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin, 91*, 404-412.

Kromrey, J.D., Lee, R.S., & Ferron, J.M. (1998). Permutation tests of equality of variances: An emprirical comparison of robustness and statistical power. Paper presented at the annual meeting of the American Educational Research Association (Chicago, IL, April, 1997).

Kromrey, J.D. & Hogarty, K.Y. (1998). Effect size estimates: An empirical study of their robustness in meta-analysis. Paper presented at the annual meeting of the Florida Educational Research Association (Orlando, FL, November, 1998).

Kromrey, J.D. & Larrimore, C.D. (1998). The robustness of meta-analytic tests for homogeneity of effect sizes: An empirical investigation. Paper presented at the annual meeting of the Florida Educational Research Association (Orlando, FL, November, 1998).

Kuhn, T. S.(1962). *The structure of scientific revolutions*.Chicago: The University of Chicago Press.

Langenfeld, T. E. & Coombs, W.T. (1998). The influence of total sample size, type of distribution, and ratio of population standard deviations on magnitude-of-effect statistics. Paper presented at the annual meeting of the American Educational Research Association (San Diego, CA, April, 1998).

Legendre, D.T. (1805). *Nouvelles method a la determination des orbites des cometes*. Paris: Courcier.

Lin, C. & Davenport, E.C. (1997). A weighted least squares approach to robustify least squares estimates. Paper presented at the annual meeting of the American Educational Research Association (Chicago, IL, March, 1997).

Lix, L.M. & Keselman, H.J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement, 58, 3,* 409-429.

Mathes, P.G. & Fuchs, L.S. (1994). The efficacy of peer tutoring in reading for students with mild disabilities: A best-evidence synthesis. *School Psychology Review, 23, 1,* 59-80.

Matt, G.E. & Cook, T.D. (1994). Threats to the validity of research syntheses. In *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

McCullagh, P. and Nelder, J.A. 1989. *Generalized Linear Models, Second Edition*. Boca Raton: Chapman & Hall/CRC.

Metsala, J.L., Stanovich, K.E. & Brown, G.D.A. (1998). Regularity effects and the phonological Deficit model of reading disabilities: A meta-analytic review. *Journal of Educational Psychology, 90, 2,* 279-93.

Mulaik, S.A., Raju, N.S. & Harshman, R.A. (1997). There is a time and a place for significance Testing. In *What If There Were No Significance Tests?* Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.

Noreen, E.W. (1989). *Computer Intensive Methods for Testing Hypotheses*. New York: Wiley.

O'Shaughnessy, T.E. & Swanson, H.L. (1998). Do immediate memory deficits in students with learning disabilities in reading reflect a developmental lag or deficit?: A selective meta-analysis of the literature. *Learning Disability Quarterly, 21,* 123-48.

Overton, R.C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3, 3,* 354-79.

Paunonen, S.V. & Gardner, R.C. (1991). Biases resulting from the use of aggregated variables in psychology. *Psychological Bulletin, 109, 3,* 520-523.

Pedhazer, E.J. 1982. *Multiple Regression in Behavioral Research*. Fort Worth: Holt, Rinehart and Winston, Inc.

Pinnell, G.S., DeFord, D.E., Bryk, A.S. & Seltzer, M. (1994). Comparing instructional models for the literacy education of high-risk first graders. *Reading Research Quarterly, 29,* 9-39.

Popper, K.R. (1968). *The logic of scientific discovery*. NewYork: Harper & Row, Publishers.

Rasmussen, J.L. & Dunlap, W.P. (1991). Dealing with nonnormal data: Parametric analysis of transformed data vs nonparametric analysis. *Educational and Psychological Measurement, 51,* 809-820.

Raudenbush, S.W. (1994). Random effects models. In *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, pg. 301-320.

Raudenbush, S.W. & Bryk, A.S. (1987). Examining correlates of diversity. *Journal of Educational Statistics, 12, 3,* 241-69.

Robey, R.R. (1990). The analysis of one-way between effects in fluency research. *Journal of Fluency Disorder*s, *15,* 275-89.

Robey, R.R. & Barcikowski, R.S. (1992). Type I error and the number of iterations in monte carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283-88.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in Psychological science. *American Psychologist, 44, 10*, 1276-1284.

Seltzer, M. (1991). The use of data augmentation in fitting hierarchical models to educational data. Unpublished doctoral dissertation, The University of Chicago, Chicago.

Serlin, R.C. (1987). Hypothesis testing, theory building, and the philosophy of science. *Journal of Counseling Psychology, 34, 4*, 365-371.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19, 1*, 1-19.

Shadish, W.R. & Haddock, C.K. (1994). Combining estimates of effect size. In *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, pg. 261-80.

Shanahan, T. & Barr, R. (1995). Reading recovery: An independent evaluation of the effects of an early instructional intervention for at-risk learners. *Reading Research Quarterly, 30,* 4, 958-96.

Snedecor, G.W. & Cochran, W.G. 1989. *Statistical Methods, eighth edition*. Ames: Iowa State University Press.

Thomas, H. (1986). Effect size standard errors for the non-normal non-identically distributed case. *Journal of Educational Statistics, 11, 4*, 293-303.

Thompson, B. (1998). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas. Paper presented at the annual meeting of the American Educational Research Association (San Diego, CA, April, 1998).

Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83-91.

Wachter, K.W. & Straf, M.L. (Eds.) 1990. *The Future of Meta-Analysis*. New York: Russell Sage Foundation.

Wang, L., Fan, X., & Willson, V.L. (1996). Effects of nonnormal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling, 3, 3*, 228-47.

Wang, M.C. & Bushman, B.J. (1999). *Integrating Results through Meta-Analytic Review Using SAS Software*.Cary, NC: SAS Institute Inc.

Weisberg, S. 1985. *Applied Linear Regression (2nd edition)*. New York: John Wiley & Sons.

Wilcox, R.R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research, 65*, 51-77.

Wilcox, R.R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology, 51*, 1-39.

Winer, B.J. 1971. *Statistical Principles in Experimental Design, 2nd edition*. New York: McGraw-Hill Book Company.

Wolf, F.M. (1990). Methodological observations on bias. In K.W. Wachter & M.L. Straf (Eds.), *The Future of Meta-Analysis*. New York: Russell Sage Foundation, pg. 139-151.

Yuen, K.K. & Dixon, W.J. (1973). The approximate behaviour and performance of the two-sample
trimmed *t*. *Biometrika*, *60*, *2*, 369-74.

Zeng, L. & Cope, R. T. (1995). Standard error of linear equating for the counterbalanced design. *Journal of Educational and Behavioral Statistics*, *20*, *4*, 337-348.

Zucker, D.M. (1990). An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement*, *50*, 731-738.

Appendices

```
* option ps=59 ls=132 pageno=1;
proc printto print='c:\my documents\dissertation
results\testnull100203.out';
proc iml;
* +-----------------------------------------------------+
   ROBUSTQ.SAS
   Changes required to execute the program:

    *NN0 - *NNB        Sample sizes
    *SPEC0 - *SPEC1    True False Moderating Null Hypothesis
    *S1 - *S4          Variances
    *SHAPE1 - *SHAPE5  Skewness and Kurtosis
    *K1 - *K3          N of studies in each meta-analysis
    *delta0 - *delta5  Population mean differences
  +-----------------------------------------------------+;

* +-----------------------------------------------------+
    Define parameters for execution of the simulation
  +-----------------------------------------------------+;
   replicat=5000;  * N of meta-analyses to simulate;

  dlta=0;
 *delta1 dlta=0.8;

 *K1 KK={2,3};       * N of studies in each meta-analysis;
  KK={4,6};
 *K3 KK={12,18};


  *NN2 njs={ 5, 5};
  *NN3 njs={20,20};
  *NN4 njs={4,6};
   njs={16,24};
  *NN7 njs={80,120};
  *NN8 njs={6,4}; * Note: NN8 - NN10 reverse pairing with unequal
variances;
  *       Use these only for non-null conditions;

  *NN0 njs={120,80};
  *NNA njs={100,100};
  *NNB njs={24,16};

 *S1 sds={1.0,1.0};
  sds={1.0,2.0};
 *S3 sds={1.0,4.0};
  *pooled=SQRT ((njs` * sds)/sum (njs));

   POOLED_VAR=(0.5)#SUM(sds);

POOLED_SD=SQRT(POOLED_VAR);
  *Tau2=0;
  *Tau2=.33;
  Tau2=1.0;
  * Tau2=5.0;
```

```
MEANDELTA=DLTA#POOLED_SD;
VARDELTA=TAU2#POOLED_VAR;
  mu1={0.0,0.0};    * Pop means for experimental and control
boys;
  mu2={0.0,0.0};    * Pop means for experimental and control
girls;


   specific=0;      * Null condition for moderation effect;
  *SPEC1 specific=1;* Non-null condition for moderation effect;

* +--------------------------------+
    Fleishman Transformations
    to nonnormality
  +--------------------------------+;
          * The following give sk= 0, kr= 0;
 *SHAPE 1b=1;
 *SHAPE 1c=0;
 *SHAPE 1d=0;

          * The following give sk= 1.00, kr= 3.00;
 *SHAPE 2b= .83221632289426;
 *SHAPE 2c= .12839670935047;
 *SHAPE 2d= .04803205907079;

          * The following give sk= 2.00, kr= 6.00;
b= 0.82632385761082;
c= 0.31374908500462;
d= 0.02270660525731;
* +--------------------+
    Initialize counters
  +--------------------+;
  rejq101 = 0;
  rejq105 = 0;
  rejq110 = 0;

  rejreq101 = 0;
  rejreq105 = 0;
  rejreq110 = 0;

  rejreq201 = 0;
  rejreq205 = 0;
  rejreq210 = 0;

  rejqb01=0;
  rejqb05=0;
  rejqb10=0;

  rejqb201=0;
  rejqb205=0;
  rejqb210=0;

  rejprob_REZ101=0;
  rejprob_REZ105=0;
  rejprob_REZ110=0;
```

```
  rejprob_FEZ101=0;
  rejprob_FEZ105=0;
  rejprob_FEZ110=0;

  rejprob_conrand01=0;
  rejprob_conrand05=0;
  rejprob_conrand10=0;

  nsamples=0;
* +-----------------------------------------+
   Subroutine to generate a random sample.
   User specifies the population mean and
   standard deviation. For population shapes,
   Fleishman constants are used.

   Inputs to the subroutine are
     NN - desired sample size
     mu - population mean
     variance - population variance
     bb,cc,dd - Fleishman constants

   Outputs are
     Rawdata - column vector of NN observations
               from the specified population
  +-----------------------------------------+;
start gendata(NN,variance,bb,cc,dd,mu,rawdata);
  seed1=round(1000000*ranuni(0));
  rawdata=rannor(repeat(seed1,nn,1));
  rawdata = (-1*cc) + (bb*rawdata) + (cc*rawdata##2) +
(dd*rawdata##3);
  rawdata = (rawdata * SQRT(variance)) + mu;
finish;
* +---------------------------------+
   Direct resampling for randomization
  +---------------------------------+;
start resamp(x);
n=Nrow(x);
allnbut=n-1;
  do i = 1 to allnbut;
     * +-----------------------------------------+
        Randomly select rows from the matrix X to
        create the matrix NEWM. Sampling is without
        replacement so that the matrix NEWM has the
        same data as X, but in random order
       +-----------------------------------------+;
```

```
 ranrow = round(uniform(0)*(n - i + 0.999)+0.5);
    if i = 1 then do;

      newm = x[ranrow,];
    end;
    if i > 1 then do;
       newm = newm//x[ranrow,];
    end;
    if ranrow > 1 then do;
        if ranrow < (n-(i-1)) then
                 x = x[1:ranrow-1,]//x[ranrow+1:n-(i-1),];
        if ranrow = n-(i-1) then x=x[1:(n-i),];
    end;
    if ranrow = 1 then x = x[2:n-(i-1),];
  end;
  newm = newm//x;
  x = newm;
* print x;
finish;
* +-------------------------------------------------------+
   Subroutine to calculate the Q test
   of homogeneity.
   Inputs to the subroutine are
     di_vec - column vector of effect sizes (d)
     n_vec  - matrix (k X 2) of sample sizes
              corresponding to each effect size

   Outputs are
     QQ = the obtained value of Q
     d_plus = weighted mean d value
     d_star = iteratively weighted mean d value
     prob_qq1 = chi-square probability associated with QQ

  +-------------------------------------------------------+;
*start calcq(di_vec,n_vec,qq,d_plus,prob_qq1);

* calculate variance for each effect size;

*  k = nrow(di_vec);
*  var_di=J(k,1,0);
*  do i = 1 to k;
*    var_di[i,1] =
((n_vec[i,1]+n_vec[i,2])/(n_vec[i,1]#n_vec[i,2])) +
((di_vec[i,1]##2)/(2#(n_vec[i,1]+n_vec[i,2])));
* end;

* calculate weighted mean effect size;

*  d_plus = 0;
*  sum_wt = 0;
*  do i = 1 to k;
*    d_plus = d_plus + di_vec[i,1]/var_di[i,1];
*    sum_wt = sum_wt + var_di[i,1]##-1;
```

```
*  end;
*  d_plus = d_plus/sum_wt;

* calculate Q;
*  QQ = 0;
*  do i = 1 to k;
*    QQ = QQ + ((di_vec[i,1] - d_plus)##2/var_di[i,1]);
*  end;
*  prob_qq1 = 1 - PROBCHI(QQ,k-1);
      * print di_vec var_di;
      * print d_plus qq prob_qq1;
*  finish;
*+------------------------------------------------------+
Subroutine to calculate the REQ test
   of homogeneity.
   Inputs to the subroutine are
       KK     - column vector of N of studies in each class
     di_vec - column vector of effect sizes (d)
     n_vec  - matrix (k X 2) of sample sizes
              corresponding to each effect size

   Outputs are
     reqq = the obtained value of Q
     d_plusrq = weighted mean d value
     d_starrq = iteratively weighted mean d value
     prob_req = chi-square probability associated with req
  +------------------------------------------------------+;
start
calcreq(KK,di_vec,n_vec,RSS_wls1,B_wls2,B_wlsi,vartheta,cov_B,cov
_B2,SE_B, SE_B2);

* calculate variance for each effect size;

  k = nrow(di_vec);
  var_di=J(k,1,0);
  Vi=J(k,1,0);

  X=J(k,1,1);
  do i = 1 to k;
    var_di[i,1] =
((n_vec[i,1]+n_vec[i,2])/(n_vec[i,1]#n_vec[i,2])) +
((di_vec[i,1]##2)/(2#(n_vec[i,1]+n_vec[i,2])));
Vi[i,1]=var_di[i,1]##-1;
  end;
 * print X;
 * print var_di di_vec Vi;

  B_ols =INV(X`*X)*X`*di_vec;
  M=X*INV(X`*X)*X`;
  NOBS =NROW(di_vec);
  IOBS = I (NOBS);
  RSS = di_vec`*(IOBS - M)* di_vec;
```

```
*const1=(J(1, NOBS, 1)*Vi) - TRACE(X`*DIAG(Vi)*X*INV(X`*X));
  const1=(J(1, NOBS, 1)*Vi##-1)-TRACE(X`*DIAG(Vi##-
1)*X*INV(X`*X));
  *A vector of variances will need to be created where we take
the reciprocal= var_di and call in the Vi;
  const2=NOBS - NCOL(X);
  vartheta=(RSS - const1)/const2;
  if vartheta<0 then vartheta=0;
  wi=Vi##-1 + J(NOBS, 1, 1)*vartheta;
  wi = wi##-1;
  wi2=Vi;
  *prob_req = 1 - PROBCHI(req,k-1);
 * print 'Initial Run Using OLS';
 * print B_ols RSS vartheta const1 const2 vi wi wi2;

  B_wls=INV(X`*DIAG(wi2)*X)*X`*DIAG(wi2)*di_vec;
  RSS_wls1=(di_vec -X*B_wls)`*DIAG(wi2)*(di_vec-X*B_wls);
 * print 'This equals Q test:'RSS_wls1;
  X=J(k,2,1);
    do i=1 to k;
         if i<=KK[1,1] then X[i,2] =0;
      end;
*  print X;
  *+-----------------------------------+
  Weighted least squares estimation using wi
  as variance estimates
  +-----------------------------------+;
  B_wls = INV(X`*DIAG(wi)*X)*X`*DIAG(wi)*di_vec;
  RSS_wls = (di_vec - X*B_wls)`*DIAG(wi)*(di_vec-X*B_wls);
  cov_b = INV(X`*DIAG(wi)*X);
  SE_B = SQRT(vecdiag(cov_b));

  B_wls2=INV(X`*DIAG(wi2)*X)*X`*DIAG(wi2)*di_vec;
  RSS_wls2=(di_vec - X*B_wls2)`*DIAG(wi2)*(di_vec-X*B_wls2);
  cov_b2= INV(X`*DIAG(wi2)*X);
  SE_B2= SQRT(vecdiag(cov_b2));

 * print 'Running Thru WLS with wi as the weights';
 * print B_wls RSS_wls cov_b SE_B;

 * print 'Running Thru WLS with wi2 as the weights';
 * print B_wls2 RSS_wls2 cov_b2 SE_B2;
  *+-----------------------------------+
  Maximum Likelihood Estimation
  +-----------------------------------+;
  change=1;
  iterate=1;

  do until(change<.0000000001);
  wi=(Vi##-1) + J(NOBS,1,1)*vartheta;
  wi=wi##-1;
  *wi2=Vi##-1;
  wi2=Vi;
```

```
B_wlsi = INV(X`*DIAG(wi)*X)*X`*DIAG(wi)*di_vec;
*  print wi B_wlsi;
  RSS_i= (di_vec - X*B_wlsi)`*DIAG(wi)*(di_vec - X*B_wlsi);
  r_vec=(di_vec-X*B_wlsi)##2;
  *varnew=SUM(wi##2#(r_vec - vi))/(wi`*wi);
  varnew=SUM(wi##2#(r_vec - var_di))/(wi`*wi);
  if varnew<0 then varnew=0;
  change=abs(vartheta - varnew);
  B_prt = B_wlsi`;
*  print 'Maximum Likelihood Algorithm';
 * print iterate vartheta varnew change B_prt RSS_i;
vartheta = varnew;
  iterate = iterate +1;
  end;
  wi=(Vi##-1) + J(NOBS,1,1)*vartheta;
  wi=wi##-1;

  cov_b = INV(X`*DIAG(wi)*X);
  cov_b2= INV(X`*DIAG(wi2)*X);
  SE_B = SQRT(vecdiag(cov_b));
  SE_B2= SQRT(vecdiag(cov_b2));
 * print 'Last Commands in Routine';
 * print cov_b cov_b2;
 * print se_b se_b2;
*prob_req = 1 - PROBCHI(req,k-1);
finish;

* +--------------------------------------+
    Subroutine to calculate Qb test of
    homogeneity of effect sizes across
    classes.
    Inputs to the subroutine are
      dp_vec1-column vector of study effect sizes for boys
      dp_vec2-column vector of study effect sizes for girls
      n_vec-matrix (K X 2) of sample sizes
            corresponding to each indivi-
            dual study
      KK - column vector with N of studies on boys and on girls
     Outputs are
       Qb=the obtained value of Qb
       d_plspls=grand mean effect size
       prob_qb=chi-square probability
               associated with Qb
+--------------------------------------+;
start calcqb (dp_vec1,dp_vec2,n_vec,KK,K,qb,d_plspls,prob_qb);

* print 'This is within the CALCB Subroutine';
* print dp_vec1 dp_vec2 n_vec KK K;

*calculate variance for each effect size for the studies on boys;

n_vec1=n_vec[1:kk[1,1],];
var_di1=J(kk[1,1], 1, 0);
```

```
do i=1 to kk[1,1];
var_di1[i,1]=((n_vec1[i,1] +
n_vec1[i,2])/(n_vec1[i,1]#n_vec1[i,2]))+
((dp_vec1[i,1]##2)/(2#(n_vec1[i,1]+n_vec1[i,2])));
end;

*calculate weighted mean effect size per class;

dp1=0;
sum_wt=0;
do i=1 to kk[1,1];
dp1=dp1 + dp_vec1[i,1]/var_di1[i,1];
sum_wt=sum_wt + var_di1[i,1]##-1;
end;
dp1=dp1/sum_wt;
* print 'These are calculations for boys';
* print n_vec1 var_di1 dp1;
*calculate variance for each effect size for the studies on
girls;

n_vec2=n_vec[(kk[1,1]+1):K,];
var_di2=J(kk[2,1], 1, 0);
do i=1 to kk[2,1];
var_di2[i,1]=((n_vec2[i,1] +
n_vec2[i,2])/(n_vec2[i,1]#n_vec2[i,2]))+
((dp_vec2[i,1]##2)/(2#(n_vec2[i,1]+n_vec2[i,2])));
end;

*calculate weighted mean effect size for girls;

dp2=0;
sum_wt=0;
do i=1 to kk[2,1];
dp2=dp2 + dp_vec2[i,1]/var_di2[i,1];
sum_wt=sum_wt + var_di2[i,1]##-1;
end;
dp2=dp2/sum_wt;
* print 'These are calculations for girls';
* print n_vec2 var_di2 dp2;

*calculate weighted grand mean (d++);
dpall=dp_vec1//dp_vec2;
varall=var_di1//var_di2;
d_plspls=0;
sum_wt=0;
do i=1 to k;
d_plspls=d_plspls + dpall[i,1]/varall[i,1];
sum_wt=sum_wt + varall[i,1]##-1;
end;
d_plspls=d_plspls/sum_wt;

*calculate Qb;
```

```
Qb=0;
do i=1 to kk[1,1];
Qb=Qb + ((dp1 - d_plspls)##2/var_di1[i,1]);
end;
do i=1 to kk[2,1];
qb=qb+ (dp2-d_plspls)##2/var_di2[i,1];
end;
prob_qb=1-PROBCHI (Qb, 1);
* print d_plspls qb prob_qb;
finish;

* +-------------------------------------------------------------
------------+
    Subroutine to calculate exact (and approximate) permutation
test of
    homogeneity of effect sizes across classes (Qb). For K = 5
and K = 10, the
    test is exact. For K = 30, the test is approximate, based on
a sample of
    1000 permutations of the obtained effect sizes.
Inputs to the subroutine are
      dp_vec1-column vector of study effect sizes for boys
      dp_vec2-column vector of study effect sizes for girls
      n_vec-matrix (K X 2) of sample sizes
            corresponding to each individual study
      KK - column vector with N of studies on boys and on girls
      K - total number of studies in the meta-analysis
      Q_real - obtained value of Qb on the actual study data

     Outputs are
       prob_qb2 - permutation probability associated with Qb

+-------------------------------------------------------------
------------+;
start Qb_exact (dp_vec1,dp_vec2,n_vec,KK,K,Q_real,prob_qb2);
  dpall = dp_vec1//dp_vec2;
  prob_qb2 = 0;
    perm = 0;
  if K = 5 then do;
    do i = 1 to K - 1;
      do j = 2 to K;
         if i < j then do;
           dvect1 = dpall[i,];
           dvect1 = dvect1//dpall[j,];
           nvect = n_vec[i,];
           nvect = nvect//n_vec[j,];
             cdt2 = 0;
           do z = 1 to K;
             if (z ^= i & z ^= j) then do;
              dvect2=dvect2//dpall[z,];
              nvect = nvect//n_vec[z,];
              end;
           end;
           run calcqb
```

```
(dvect1,dvect2,nvect,KK,K,qbtemp,d_plspl,probqbt);
           if Qbtemp < Q_real then prob_qb2 = prob_qb2 + 1;
           perm = perm + 1;
           free dvect1 dvect2 nvect;
         end;
      end;
    end;
    prob_qb2 = 1 - (prob_qb2 / perm);
    end;
  if K = 10 then do;
    do i = 1 to K - 3;
      do j = 2 to K - 2;
        do l = 3 to K - 1;
          do m = 4 to K;
              if (i<j & j<l & l<m) then do;
                dvect1 = dpall[i,];
                dvect1 = dvect1//dpall[j,];
                dvect1 = dvect1//dpall[l,];
                dvect1 = dvect1//dpall[m,];
                nvect = n_vec[i,];
                nvect = nvect//n_vec[j,];

      nvect = nvect//n_vec[l,];
                nvect = nvect//n_vec[m,];

                do z = 1 to K;
                 if (z ^= i & z ^= j & z ^= l & z ^= m) then do;
                   dvect2=dvect2//dpall[z,];
                   nvect = nvect//n_vec[z,];
                   end;
                end;
               run calcqb
(dvect1,dvect2,nvect,KK,K,qbtemp,d_plspl,probqbt);
                if Qbtemp < Q_real then prob_qb2 = prob_qb2 + 1;
                perm = perm + 1;
                free dvect1 dvect2 nvect;
              end;
          end;
        end;
      end;
    end;
    prob_qb2 = 1 - (prob_qb2 / perm);
    end;
if K=30 then do;
  dpN=dpall||n_vec;
  do i=1 to 1000;
   run resamp (dpN);
   dvect1=dpN[1:12,1];
   dvect2=dpN[13:30,1];
   nvect=dpN[, 2:3];
```

```
 run calcqb (dvect1,dvect2,nvect, KK, K, qbtemp, d_plspls,
probqbt);
 if qbtemp < Q_real then prob_qb2=prob_qb2+1;

 perm=perm+1;
 free dvect1 dvect2 nvect;
end;
prob_qb2=1 - (prob_qb2/perm);
end;
finish;
* +---------------------------------+
     Bubble sort
      X = matrix to be sorted
      N = number of rows in matrix
      C = column number to sort by
  +---------------------------------+;
start bubble(x,n,c);
do i = 1 to n;
  do j = 1 to n-1;
    if x[J,C] > x[J+1,C] then do;
       temp = x[J+1,];
       x[J+1,] = x[J,];
       x[J,] = temp;
    end;
  end;
end;
finish;


* +-------------------------------------+
   Main program
   Generates samples, calls subroutines,
   computes rejection rates.
  +-------------------------------------+;

do rep=1 to replicat;              * This starts the big do loop;

k=sum(kk);

  n_vec=J(K,2,0);
  do study = 1 to k;            * Inner loop for primary studies;
  ALPHA_K=RANNOR(0)#SQRT(VARDELTA)+MEANDELTA;
  Mean_exp=mu1[1,1]+alpha_k;
      n1=rannor(0)+ njs[1,1];
      n1=round(n1);
      if n1<3 then n1=3;

      n2=rannor (0)+njs[2,1];
      n2=round(n2);
      if n2<3 then n2=3;
```

```
 n_vec[study,1]=n1;
  n_vec[study,2]=n2;
      *print 'Group Sizes:' Rep Study n1 n2;
     * generate a vector of scores for each group;

if (study <= kk[1,1]) then do;
     run gendata(n1,sds[1,1],b,c,d,mu1[1,1],z1);
     run gendata(n2,sds[2,1],b,c,d,mean_exp,z2);
end;
if (study > kk[1,1]) then do;
     run gendata (n1,sds[1,1],b,c,d, mu2[1,1], z1);
     run gendata (n2,sds[2,1],b,c,d, mean_exp,z2);
end;

 * calculate sample means, SS, di and cd for primary studies;

        xbar1 = (J(1,n1,1)*z1)/n1;
        xbar2 = (J(1,n2,1)*z2)/n2;
        ss1 = (J(1,n1,1)*(z1##2)) - ((J(1,n1,1)*z1)##2/n1);
        ss2 = (J(1,n2,1)*(z2##2)) - ((J(1,n2,1)*z2)##2/n2);
            njstemp = n1//n2;
        di = ((xbar1 - xbar2)/sqrt((ss1 + ss2)/(J(1,2,1)*njstemp
- 2 )))#(1 - (3/(4#(J(1,2,1)*njstemp)-9)));
        if study = 1 then di_vec = di;
        if study > 1 then di_vec = di_vec//di;

  end;                                    * End inner loop for
primary studies;

* Calculate the regular parametric Q tests;

  *n_vec = njs*J(1,k,1);
  *n_vec = T(n_vec);
* run calcq(di_vec,n_vec,q,d_plus,prob_q1);
 *print 'From CALCQ';
* print di_vec n_vec q d_plus prob_q1;

run
calcreq(KK,di_vec,n_vec,RSS_wls1,B_wls2,B_wlsi,vartheta,cov_B,cov
_B2,SE_B,SE_B2);
prob_q1=1-probchi(RSS_wls1,k-1);
*print 'FROM CALCREQ';
*print di_vec n_vec RSS_wls1 B_wls2 B_wlsi vartheta cov_B cov_B2
SE_B SE_B2;

FE_T=B_wls2[2,1]/SE_B2[2,1];
PROB_FE=2#(1-probt(abs(FE_T),K-2));
PROB_FEZ=2#(1-probnorm(abs(FE_T)));
*print 'Fixed Effects Test:' fe_t prob_fe prob_fez;

RE_T = B_wlsi[2,1]/se_b[2,1];
PROB_RE=2#(1-probt(abs(RE_T),K-2));
PROB_REZ=2#(1-probnorm(abs(RE_T)));
```

```
*--------------------------------;
*record reject and fail-to-reject probabilities for the
random- and fixed-effects magnitude of effects tests.
*--------------------------------;
if prob_REZ < .01 then rejprob_REZ101 = rejprob_REZ101+1;
if prob_REZ < .05 then rejprob_REZ105 = rejprob_REZ105+1;
if prob_REZ < .10 then rejprob_REZ110 = rejprob_REZ110+1;
if prob_FEZ < .01 then rejprob_FEZ101 = rejprob_FEZ101+1;
if prob_FEZ < .05 then rejprob_FEZ105 = rejprob_FEZ105+1;
if prob_FEZ < .10 then rejprob_FEZ110 = rejprob_FEZ110+1;


*print 'Random Effects Test:'RE_t prob_RE prob_REZ;

*calculate Qb tests;

  run calcqb (di_vec[1:kk[1,1],],di_vec[kk[1,1]+1:k,], n_vec, kk,
k, qb, d_plspls, prob_qb);
*   print 'From CALCQB';
*   print qb d_plspls prob_qb;

  *-------------------------------------;
  *Subroutine to run conditionally-random procedure
  *conrand=conditionally-random procedure
  *-------------------------------------;
  conrand=prob_q1;
  if prob_q1<.05 then conrand=prob_REZ;
  if prob_q1>.05 then conrand=prob_FEZ;

if conrand < .01 then rejprob_conrand01 = rejprob_conrand01+1;
  if conrand < .05 then rejprob_conrand05 = rejprob_conrand05+1;
  if conrand < .10 then rejprob_conrand10 = rejprob_conrand10+1;

  * print 'Conditionally Random Test:'conrand;

 * Permutation test of Qb;

  run Qb_exact
(di_vec[1:kk[1,1],],di_vec[kk[1,1]+1:k,],n_vec,KK,K,FE_T##2,prob_
qb2);

     * Record reject/fail to reject for each test;

       if prob_q1 < .01 then rejq101 = rejq101+1;
         if prob_q1 < .05 then rejq105 = rejq105+1;
         if prob_q1 < .10 then rejq110 = rejq110+1;

       if prob_qb2 < .01 then rejqb201=rejqb201+1;
       if prob_qb2 < .05 then rejqb205=rejqb205+1;
       if prob_qb2 < .10 then rejqb210=rejqb210+1;

       nsamples=nsamples+1;
end;
 *end the big loop;
```

```
rejprob_REZ101 = rejprob_REZ101/nsamples;
rejprob_REZ105 = rejprob_REZ105/nsamples;
rejprob_REZ110 = rejprob_REZ110/nsamples;

rejprob_FEZ101 = rejprob_FEZ101/nsamples;
rejprob_FEZ105 = rejprob_FEZ105/nsamples;
rejprob_FEZ110 = rejprob_FEZ110/nsamples;

rejprob_conrand01 = rejprob_conrand01/nsamples;
rejprob_conrand05 = rejprob_conrand05/nsamples;
rejprob_conrand10 = rejprob_conrand10/nsamples;
* +-------------------------------------------------------+
   Convert counts of rejected hypotheses into proportions
  +-------------------------------------------------------+;
     rejq101 = rejq101/nsamples;
       rejq105 = rejq105/nsamples;
       rejq110 = rejq110/nsamples;

       *rejreq101 = rejreq101/nsamples;
       *rejreq105 = rejreq105/nsamples;
       *rejreq110 = rejreq110/nsamples;

       *rejreq201 = rejreq201/nsamples;
       *rejreq205 = rejreq205/nsamples;
       *rejreq210 = rejreq210/nsamples;

     *rejqb01=rejqb01/nsamples;
     *rejqb05=rejqb05/nsamples;
     *rejqb10=rejqb10/nsamples;


       rejqb201=rejqb201/nsamples;
     rejqb205=rejqb205/nsamples;
     rejqb210=rejqb210/nsamples;

*print 'Tests of Homogeneity of Effect Sizes';

PRINT njs sds dlta kk specific b Tau2 nsamples;

*print 'sk= 0.00, kr=  0.00';
*SHAPE2  print 'sk= 1.00, kr=  3.00';
*SHAPE3  print 'sk= 1.50, kr=  5.00';
*SHAPE4  print 'sk= 2.00, kr=  6.00';
*SHAPE5  print 'sk= 0.00, kr= 25.00';

PRINT  rejq101, rejq105, rejq110;
*PRINT  rejreq101, rejreq105, rejreq110, rejreq201, rejreq205,
rejreq210;
PRINT  rejqb201, rejqb205, rejqb210;
PRINT rejprob_REZ101, rejprob_REZ105, rejprob_REZ110;
PRINT rejprob_FEZ101,rejprob_FEZ105,rejprob_FEZ110;
PRINT rejprob_conrand01, rejprob_conrand05, rejprob_conrand10;
quit;
```

```
* option ps=59 ls=132 pageno=1;
proc printto print='c:\my documents\dissertation \test48.out';
proc iml;
* +-----------------------------------------------------+
   ROBUSTQ.SAS
   Changes required to execute the program:
    *NN0 - *NNB        Sample sizes
    *SPEC0 - *SPEC1    True False Moderating Null Hypothesis
    *S1 - *S4          Variances
    *SHAPE1 - *SHAPE5  Skewness and Kurtosis
    *K1 - *K3          N of studies in each meta-analysis
    *delta0 - *delta5  Population mean differences
  +-----------------------------------------------------+;


* +-----------------------------------------------------+
    Define parameters for execution of the simulation
  +-----------------------------------------------------+;
   replicat=5000;  * N of meta-analyses to simulate;



  *delta0 dlta=0;
  dlta=0.8;

 *K1 KK={2,3};       * N of studies in each meta-analysis;
  KK={4,6};
 *K3 KK={12,18};

  *NN2 njs={ 5, 5};
  *NN3 njs={20,20};

  *NN5 njs={4,6};
  *NN6 njs={16,24};
  *NN7 njs={80,120};
  *NN8 njs={6,4}; * Note: NN8 - NN10 reverse pairing with unequal
variances;
  *       Use these only for non-null conditions;


   njs={120,80};
  *NNA njs={100,100};
  *NNB njs={24,16};


 *S1 sds={1.0,1.0};
 *S2 sds={1.0,2.0};
  sds={1.0,4.0};
  *pooled=SQRT ((njs` * sds)/sum (njs));
   POOLED_VAR=(0.5)#SUM(sds);
   POOLED_SD=SQRT(POOLED_VAR);
  Tau2=0;
```

```
*Tau2=.33;
  *Tau2=1.0;

  * Tau2=5.0;
MEANDELTA=DLTA#POOLED_SD;
VARDELTA=TAU2#POOLED_VAR;
  mu1={0.0,0.0};    * Pop means for experimental and control
boys;
  mu2={0.0,0.0};    * Pop means for experimental
  and control girls;

  *SPEC0 specific=0;      * Null condition for moderation effect;
   specific=1;* Non-null condition for moderation effect;



* +--------------------------------+
    Fleishman Transformations
    to nonnormality
  +--------------------------------+;
          * The following give sk= 0, kr= 0;
 *SHAPE 1b=1;
 *SHAPE 1c=0;
 *SHAPE 1d=0;

          * The following give sk= 1.00, kr= 3.00;
 *SHAPE 2b= .83221632289426;
 *SHAPE 2c= .12839670935047;
 *SHAPE 2d= .04803205907079;


          * The following give sk= 2.00, kr= 6.00;
 b= 0.82632385761082;
 c= 0.31374908500462;
 d= 0.02270660525731;


* +--------------------+
    Initialize counters
  +--------------------+;
  rejq101 = 0;
  rejq105 = 0;
  rejq110 = 0;

  rejreq101 = 0;
  rejreq105 = 0;
  rejreq110 = 0;

  rejreq201 = 0;
  rejreq205 = 0;
  rejreq210 = 0;

  rejqb01=0;
  rejqb05=0;
  rejqb10=0;
```

```
 rejqb201=0;
  rejqb205=0;
  rejqb210=0;

  rejprob_REZ101=0;
  rejprob_REZ105=0;
  rejprob_REZ110=0;

  rejprob_FEZ101=0;
  rejprob_FEZ105=0;
  rejprob_FEZ110=0;

  rejprob_conrand01=0;
  rejprob_conrand05=0;
  rejprob_conrand10=0;

  nsamples=0;

* +-----------------------------------------+
   Subroutine to generate a random sample.
   User specifies the population mean and
   standard deviation. For population shapes,
   Fleishman constants are used.

   Inputs to the subroutine are
     NN - desired sample size
     mu - population mean
     variance - population variance
     bb,cc,dd - Fleishman constants

   Outputs are
     Rawdata - column vector of NN observations
               from the specified population
  +-----------------------------------------+;
start gendata(NN,variance,bb,cc,dd,mu,rawdata);
  seed1=round(1000000*ranuni(0));
  rawdata=rannor(repeat(seed1,nn,1));
  rawdata = (-1*cc) + (bb*rawdata) + (cc*rawdata##2) +
(dd*rawdata##3);
  rawdata = (rawdata * SQRT(variance)) + mu;
finish;

* +---------------------------------+
   Direct resampling for randomization
  +---------------------------------+;
start resamp(x);
n=Nrow(x);
allnbut=n-1;
  do i = 1 to allnbut;
```

```
   * +-----------------------------------------+
       Randomly select rows from the matrix X to
       create the matrix NEWM. Sampling is without
       replacement so that the matrix NEWM has the
       same data as X, but in random order
      +-----------------------------------------+;
    ranrow = round(uniform(0)*(n - i + 0.999)+0.5);
    if i = 1 then do;
       newm = x[ranrow,];
    end;
    if i > 1 then do;
       newm = newm//x[ranrow,];
    end;
    if ranrow > 1 then do;
        if ranrow < (n-(i-1)) then
                 x = x[1:ranrow-1,]//x[ranrow+1:n-(i-1),];
        if ranrow = n-(i-1) then x=x[1:(n-i),];
    end;
    if ranrow = 1 then x = x[2:n-(i-1),];
  end;
  newm = newm//x;
  x = newm;
* print x;
finish;

* +-------------------------------------------------------+
   Subroutine to calculate the Q test
   of homogeneity.
   Inputs to the subroutine are
     di_vec - column vector of effect sizes (d)
     n_vec  - matrix (k X 2) of sample sizes
              corresponding to each effect size

   Outputs are
     QQ = the obtained value of Q
     d_plus = weighted mean d value
     d_star = iteratively weighted mean d value
     prob_qq1 = chi-square probability associated with QQ

  +-------------------------------------------------------+;
*start calcq(di_vec,n_vec,qq,d_plus,prob_qq1);

* calculate variance for each effect size;

*  k = nrow(di_vec);
*  var_di=J(k,1,0);
*  do i = 1 to k;
*    var_di[i,1] =
((n_vec[i,1]+n_vec[i,2])/(n_vec[i,1]#n_vec[i,2])) +
((di_vec[i,1]##2)/(2#(n_vec[i,1]+n_vec[i,2])));
* end;

* calculate weighted mean effect size;
```

```
*  d_plus = 0;
*  sum_wt = 0;
*  do i = 1 to k;
*    d_plus = d_plus + di_vec[i,1]/var_di[i,1];
*    sum_wt = sum_wt + var_di[i,1]##-1;
*  end;
*  d_plus = d_plus/sum_wt;

* calculate Q;

*  QQ = 0;
*  do i = 1 to k;
*    QQ = QQ + ((di_vec[i,1] - d_plus)##2/var_di[i,1]);
*  end;
*  prob_qq1 = 1 - PROBCHI(QQ,k-1);
      * print di_vec var_di;
      * print d_plus qq prob_qq1;
*  finish;
*+-----------------------------------------------------------+
Subroutine to calculate the REQ test
   of homogeneity.
   Inputs to the subroutine are
       KK     - column vector of N of studies in each class
     di_vec - column vector of effect sizes (d)
     n_vec  - matrix (k X 2) of sample sizes
              corresponding to each effect size

   Outputs are
     reqq = the obtained value of Q
     d_plusrq = weighted mean d value
     d_starrq = iteratively weighted mean d value
     prob_req = chi-square probability associated with req

  +-----------------------------------------------------------+;
start
calcreq(KK,di_vec,n_vec,RSS_wls1,B_wls2,B_wlsi,vartheta,cov_B,cov
_B2,SE_B, SE_B2);

* calculate variance for each effect size;

  k = nrow(di_vec);
  var_di=J(k,1,0);
  Vi=J(k,1,0);

  X=J(k,1,1);
  do i = 1 to k;
    var_di[i,1] =
((n_vec[i,1]+n_vec[i,2])/(n_vec[i,1]#n_vec[i,2])) +
((di_vec[i,1]##2)/(2#(n_vec[i,1]+n_vec[i,2])));
Vi[i,1]=var_di[i,1]##-1;
  end;
 * print X;
 * print var_di di_vec Vi;
```

```
B_ols =INV(X`*X)*X`*di_vec;
  M=X*INV(X`*X)*X`;
  NOBS =NROW(di_vec);
  IOBS = I (NOBS);
  RSS = di_vec`*(IOBS - M)* di_vec;
  *const1=(J(1, NOBS, 1)*Vi) - TRACE(X`*DIAG(Vi)*X*INV(X`*X));
  const1=(J(1, NOBS, 1)*Vi##-1)-TRACE(X`*DIAG(Vi##-
1)*X*INV(X`*X));
  *A vector of variances will need to be created where we take
the reciprocal= var_di and call in the Vi;
  const2=NOBS - NCOL(X);
  vartheta=(RSS - const1)/const2;
  if vartheta<0 then vartheta=0;
  wi=Vi##-1 + J(NOBS, 1, 1)*vartheta;
  wi = wi##-1;
  wi2=Vi;
  *prob_req = 1 - PROBCHI(req,k-1);
 * print 'Initial Run Using OLS';
 * print B_ols RSS vartheta const1 const2 vi wi wi2;

  B_wls=INV(X`*DIAG(wi2)*X)*X`*DIAG(wi2)*di_vec;
  RSS_wls1=(di_vec -X*B_wls)`*DIAG(wi2)*(di_vec-X*B_wls);
 * print 'This equals Q test:'RSS_wls1;
  X=J(k,2,1);
    do i=1 to k;
         if i<=KK[1,1] then X[i,2] =0;
      end;
*  print X;
  *+-----------------------------------+
  Weighted least squares estimation using wi
  as variance estimates
  +-----------------------------------+;
  B_wls = INV(X`*DIAG(wi)*X)*X`*DIAG(wi)*di_vec;
  RSS_wls = (di_vec - X*B_wls)`*DIAG(wi)*(di_vec-X*B_wls);
  cov_b = INV(X`*DIAG(wi)*X);
  SE_B = SQRT(vecdiag(cov_b));

  B_wls2=INV(X`*DIAG(wi2)*X)*X`*DIAG(wi2)*di_vec;
  RSS_wls2=(di_vec - X*B_wls2)`*DIAG(wi2)*(di_vec-X*B_wls2);
  cov_b2= INV(X`*DIAG(wi2)*X);
  SE_B2= SQRT(vecdiag(cov_b2));

 * print 'Running Thru WLS with wi as the weights';
 * print B_wls RSS_wls cov_b SE_B;

 * print 'Running Thru WLS with wi2 as the weights';
 * print B_wls2 RSS_wls2 cov_b2 SE_B2;
  *+-----------------------------------+
  Maximum Likelihood Estimation
  +-----------------------------------+;
  change=1;
  iterate=1;
```

```
do until(change<.0000000001);
  wi=(Vi##-1) + J(NOBS,1,1)*vartheta;
  wi=wi##-1;
  *wi2=Vi##-1;
  wi2=Vi;

  B_wlsi = INV(X`*DIAG(wi)*X)*X`*DIAG(wi)*di_vec;
*  print wi B_wlsi;
  RSS_i= (di_vec - X*B_wlsi)`*DIAG(wi)*(di_vec - X*B_wlsi);
  r_vec=(di_vec-X*B_wlsi)##2;
  *varnew=SUM(wi##2#(r_vec - vi))/(wi`*wi);
  varnew=SUM(wi##2#(r_vec - var_di))/(wi`*wi);
  if varnew<0 then varnew=0;
  change=abs(vartheta - varnew);
  B_prt = B_wlsi`;
*  print 'Maximum Likelihood Algorithm';
 * print iterate vartheta varnew change B_prt RSS_i;
  vartheta = varnew;
  iterate = iterate +1;
  end;
  wi=(Vi##-1) + J(NOBS,1,1)*vartheta;
  wi=wi##-1;

  cov_b = INV(X`*DIAG(wi)*X);
  cov_b2= INV(X`*DIAG(wi2)*X);
  SE_B = SQRT(vecdiag(cov_b));
  SE_B2= SQRT(vecdiag(cov_b2));
 * print 'Last Commands in Routine';
 * print cov_b cov_b2;
 * print se_b se_b2;
*prob_req = 1 - PROBCHI(req,k-1);
finish;

* +--------------------------------------+
    Subroutine to calculate Qb test of
    homogeneity of effect sizes across
    classes.
    Inputs to the subroutine are
      dp_vec1-column vector of study effect sizes for boys
      dp_vec2-column vector of study effect sizes for girls
      n_vec-matrix (K X 2) of sample sizes
            corresponding to each indivi-
            dual study
      KK - column vector with N of studies on boys and on girls
     Outputs are
       Qb=the obtained value of Qb
       d_plspls=grand mean effect size
       prob_qb=chi-square probability
               associated with Qb
+--------------------------------------+;
start calcqb (dp_vec1,dp_vec2,n_vec,KK,K,qb,d_plspls,prob_qb);
* print 'This is within the CALCB Subroutine';
```

```
* print dp_vec1 dp_vec2 n_vec KK K;

*calculate variance for each effect size for the studies on boys;

n_vec1=n_vec[1:kk[1,1],];
var_di1=J(kk[1,1], 1, 0);
do i=1 to kk[1,1];
var_di1[i,1]=((n_vec1[i,1] +
n_vec1[i,2])/(n_vec1[i,1]#n_vec1[i,2]))+
((dp_vec1[i,1]##2)/(2#(n_vec1[i,1]+n_vec1[i,2])));
end;

*calculate weighted mean effect size per class;

dp1=0;
sum_wt=0;
do i=1 to kk[1,1];
dp1=dp1 + dp_vec1[i,1]/var_di1[i,1];
sum_wt=sum_wt + var_di1[i,1]##-1;
end;
dp1=dp1/sum_wt;
* print 'These are calculations for boys';
* print n_vec1 var_di1 dp1;
*calculate variance for each effect size for the studies on
girls;

n_vec2=n_vec[(kk[1,1]+1):K,];
var_di2=J(kk[2,1], 1, 0);
do i=1 to kk[2,1];
var_di2[i,1]=((n_vec2[i,1] +
n_vec2[i,2])/(n_vec2[i,1]#n_vec2[i,2]))+
((dp_vec2[i,1]##2)/(2#(n_vec2[i,1]+n_vec2[i,2])));
end;

*calculate weighted mean effect size for girls;

dp2=0;
sum_wt=0;
do i=1 to kk[2,1];
dp2=dp2 + dp_vec2[i,1]/var_di2[i,1];
sum_wt=sum_wt + var_di2[i,1]##-1;
end;
dp2=dp2/sum_wt;
* print 'These are calculations for girls';
* print n_vec2 var_di2 dp2;

*calculate weighted grand mean (d++);
dpall=dp_vec1//dp_vec2;
varall=var_di1//var_di2;
d_plspls=0;
sum_wt=0;
do i=1 to k;
```

```
d_plspls=d_plspls + dpall[i,1]/varall[i,1];
sum_wt=sum_wt + varall[i,1]##-1;
end;
d_plspls=d_plspls/sum_wt;

*calculate Qb;

Qb=0;
do i=1 to kk[1,1];
Qb=Qb + ((dp1 - d_plspls)##2/var_di1[i,1]);
end;
do i=1 to kk[2,1];
qb=qb+ (dp2-d_plspls)##2/var_di2[i,1];
end;
prob_qb=1-PROBCHI (Qb, 1);
* print d_plspls qb prob_qb;
finish;

* +-------------------------------------------------------------
------------+
    Subroutine to calculate exact (and approximate) permutation
test of
    homogeneity of effect sizes across classes (Qb). For K = 5
and K = 10, the
    test is exact. For K = 30, the test is approximate, based on
a sample of
    1000 permutations of the obtained effect sizes.

    Inputs to the subroutine are
      dp_vec1-column vector of study effect sizes for boys
      dp_vec2-column vector of study effect sizes for girls
      n_vec-matrix (K X 2) of sample sizes
            corresponding to each individual study
      KK - column vector with N of studies on boys and on girls
      K - total number of studies in the meta-analysis
      Q_real - obtained value of Qb on the actual study data

     Outputs are
       prob_qb2 - permutation probability associated with Qb

+-------------------------------------------------------------
------------+;
start Qb_exact (dp_vec1,dp_vec2,n_vec,KK,K,Q_real,prob_qb2);
  dpall = dp_vec1//dp_vec2;
  prob_qb2 = 0;
    perm = 0;
  if K = 5 then do;
    do i = 1 to K - 1;
      do j = 2 to K;
         if i < j then do;
           dvect1 = dpall[i,];
           dvect1 = dvect1//dpall[j,];
           nvect = n_vec[i,];
```

```
            nvect = nvect//n_vec[j,];
              cdt2 = 0;
            do z = 1 to K;
              if (z ^= i & z ^= j) then do;
               dvect2=dvect2//dpall[z,];
               nvect = nvect//n_vec[z,];
               end;
            end;
            run calcqb
(dvect1,dvect2,nvect,KK,K,qbtemp,d_plspl,probqbt);
            if Qbtemp < Q_real then prob_qb2 = prob_qb2 + 1;
            perm = perm + 1;
            free dvect1 dvect2 nvect;
          end;
        end;
      end;
      prob_qb2 = 1 - (prob_qb2 / perm);
      end;
    if K = 10 then do;
      do i = 1 to K - 3;
        do j = 2 to K - 2;
          do l = 3 to K - 1;
            do m = 4 to K;
                if (i<j & j<l & l<m) then do;
                  dvect1 = dpall[i,];
                  dvect1 = dvect1//dpall[j,];
                  dvect1 = dvect1//dpall[l,];
                  dvect1 = dvect1//dpall[m,];
                  nvect = n_vec[i,];
                  nvect = nvect//n_vec[j,];
                  nvect = nvect//n_vec[l,];
                  nvect = nvect//n_vec[m,];

                  do z = 1 to K;
                   if (z ^= i & z ^= j & z ^= l & z ^= m) then do;
                     dvect2=dvect2//dpall[z,];
                     nvect = nvect//n_vec[z,];
                     end;
                  end;
                 run calcqb
(dvect1,dvect2,nvect,KK,K,qbtemp,d_plspl,probqbt);
                  if Qbtemp < Q_real then prob_qb2 = prob_qb2 + 1;
                  perm = perm + 1;
                  free dvect1 dvect2 nvect;
                end;
            end;
          end;
        end;
      end;
      prob_qb2 = 1 - (prob_qb2 / perm);
      end;
if K=30 then do;
  dpN=dpall||n_vec;
  do i=1 to 1000;
```

```
  run resamp (dpN);
   dvect1=dpN[1:12,1];
   dvect2=dpN[13:30,1];
   nvect=dpN[, 2:3];

 run calcqb (dvect1,dvect2,nvect, KK, K, qbtemp, d_plspls,
probqbt);
 if qbtemp < Q_real then prob_qb2=prob_qb2+1;

 perm=perm+1;
 free dvect1 dvect2 nvect;
end;
prob_qb2=1 - (prob_qb2/perm);
end;
finish;
* +---------------------------------+
     Bubble sort
      X = matrix to be sorted
      N = number of rows in matrix
      C = column number to sort by
  +---------------------------------+;
start bubble(x,n,c);
do i = 1 to n;
  do j = 1 to n-1;
    if x[J,C] > x[J+1,C] then do;
       temp = x[J+1,];
       x[J+1,] = x[J,];
       x[J,] = temp;
    end;
  end;
end;
finish;

* +-------------------------------------+
   Main program
   Generates samples, calls subroutines,
   computes rejection rates.
  +-------------------------------------+;

do rep=1 to replicat;                       * This starts the big
do loop;

k=sum(kk);

  n_vec=J(K,2,0);
  do study = 1 to k;                        * Inner loop for
primary studies;
  ALPHA_K=RANNOR(0)#SQRT(VARDELTA)+MEANDELTA;
ALPHA_K_GIRLS = RANNOR(0)#SQRT(VARDELTA);
Mean_exp=mu1[1,1]+alpha_k;
MEAN_GIRLS=mu2[1,1]+ ALPHA_K_GIRLS;
      n1=rannor(0)+ njs[1,1];
      n1=round(n1);
      if n1<3 then n1=3;
```

```
      n2=rannor (0)+njs[2,1];
      n2=round(n2);
      if n2<3 then n2=3;

  n_vec[study,1]=n1;
  n_vec[study,2]=n2;
      *print 'Group Sizes:' Rep Study n1 n2;
     * generate a vector of scores for each group;

if (study <= kk[1,1]) then do;
     run gendata(n1,sds[1,1],b,c,d,mu1[1,1],z1);
     run gendata(n2,sds[2,1],b,c,d,mean_exp,z2);
end;
if (study > kk[1,1]) then do;
     run gendata (n1,sds[1,1],b,c,d, mu2[1,1], z1);
     run gendata (n2,sds[2,1],b,c,d, mean_girls,z2);
end;

 * calculate sample means, SS, di and cd for primary studies;

        xbar1 = (J(1,n1,1)*z1)/n1;
        xbar2 = (J(1,n2,1)*z2)/n2;
        ss1 = (J(1,n1,1)*(z1##2)) - ((J(1,n1,1)*z1)##2/n1);
        ss2 = (J(1,n2,1)*(z2##2)) - ((J(1,n2,1)*z2)##2/n2);
            njstemp = n1//n2;
        di = ((xbar1 - xbar2)/sqrt((ss1 + ss2)/(J(1,2,1)*njstemp
- 2 )))#(1 - (3/(4#(J(1,2,1)*njstemp)-9)));
        if study = 1 then di_vec = di;
        if study > 1 then di_vec = di_vec//di;



  end;                      * End inner loop for primary studies;

     * Calculate the regular parametric Q tests;

  *n_vec = njs*J(1,k,1);
  *n_vec = T(n_vec);
* run calcq(di_vec,n_vec,q,d_plus,prob_q1);
 *print 'From CALCQ';
* print di_vec n_vec q d_plus prob_q1;

run
calcreq(KK,di_vec,n_vec,RSS_wls1,B_wls2,B_wlsi,vartheta,cov_B,cov
_B2,SE_B,SE_B2);
prob_q1=1-probchi(RSS_wls1,k-1);
*print 'FROM CALCREQ';
*print di_vec n_vec RSS_wls1 B_wls2 B_wlsi vartheta cov_B cov_B2
SE_B SE_B2;

FE_T=B_wls2[2,1]/SE_B2[2,1];
PROB_FE=2#(1-probt(abs(FE_T),K-2));
```

```
PROB_FEZ=2#(1-probnorm(abs(FE_T)));
*print 'Fixed Effects Test:' fe_t prob_fe prob_fez;

RE_T = B_wlsi[2,1]/se_b[2,1];
PROB_RE=2#(1-probt(abs(RE_T),K-2));
PROB_REZ=2#(1-probnorm(abs(RE_T)));

*--------------------------------;
*record reject and fail-to-reject probabilities for the
random- and fixed-effects magnitude of effects tests.
*--------------------------------;
if prob_REZ < .01 then rejprob_REZ101 = rejprob_REZ101+1;
if prob_REZ < .05 then rejprob_REZ105 = rejprob_REZ105+1;
if prob_REZ < .10 then rejprob_REZ110 = rejprob_REZ110+1;
if prob_FEZ < .01 then rejprob_FEZ101 = rejprob_FEZ101+1;
if prob_FEZ < .05 then rejprob_FEZ105 = rejprob_FEZ105+1;
if prob_FEZ < .10 then rejprob_FEZ110 = rejprob_FEZ110+1;


*print 'Random Effects Test:'RE_t prob_RE prob_REZ;

*calculate Qb tests;

  run calcqb (di_vec[1:kk[1,1],],di_vec[kk[1,1]+1:k,], n_vec, kk,
k, qb, d_plspls, prob_qb);
*   print 'From CALCQB';
*   print qb d_plspls prob_qb;

  *-------------------------------------;
  *Subroutine to run conditionally-random procedure
  *conrand=conditionally-random procedure
  *-------------------------------------;
  conrand=prob_q1;
  if prob_q1<.05 then conrand=prob_REZ;
  if prob_q1>.05 then conrand=prob_FEZ;

  if conrand < .01 then rejprob_conrand01 = rejprob_conrand01+1;
  if conrand < .05 then rejprob_conrand05 = rejprob_conrand05+1;
  if conrand < .10 then rejprob_conrand10 = rejprob_conrand10+1;

  * print 'Conditionally Random Test:'conrand;

 * Permutation test of Qb;

  run Qb_exact
(di_vec[1:kk[1,1],],di_vec[kk[1,1]+1:k,],n_vec,KK,K,FE_T##2,prob_
qb2);

     * Record reject/fail to reject for each test;

       if prob_q1 < .01 then rejq101 = rejq101+1;
         if prob_q1 < .05 then rejq105 = rejq105+1;
```

```
 if prob_q1 < .10 then rejq110 = rejq110+1;

       if prob_qb2 < .01 then rejqb201=rejqb201+1;
       if prob_qb2 < .05 then rejqb205=rejqb205+1;
       if prob_qb2 < .10 then rejqb210=rejqb210+1;

       nsamples=nsamples+1;


end;                                          *end the big loop;

rejprob_REZ101 = rejprob_REZ101/nsamples;
rejprob_REZ105 = rejprob_REZ105/nsamples;
rejprob_REZ110 = rejprob_REZ110/nsamples;

rejprob_FEZ101 = rejprob_FEZ101/nsamples;
rejprob_FEZ105 = rejprob_FEZ105/nsamples;
rejprob_FEZ110 = rejprob_FEZ110/nsamples;

rejprob_conrand01 = rejprob_conrand01/nsamples;
rejprob_conrand05 = rejprob_conrand05/nsamples;
rejprob_conrand10 = rejprob_conrand10/nsamples;
* +-----------------------------------------------------+
   Convert counts of rejected hypotheses into proportions
  +-----------------------------------------------------+;
     rejq101 = rejq101/nsamples;
       rejq105 = rejq105/nsamples;
       rejq110 = rejq110/nsamples;

       *rejreq101 = rejreq101/nsamples;
       *rejreq105 = rejreq105/nsamples;
       *rejreq110 = rejreq110/nsamples;

       *rejreq201 = rejreq201/nsamples;
       *rejreq205 = rejreq205/nsamples;
       *rejreq210 = rejreq210/nsamples;

     *rejqb01=rejqb01/nsamples;
     *rejqb05=rejqb05/nsamples;
     *rejqb10=rejqb10/nsamples;

     rejqb201=rejqb201/nsamples;
     rejqb205=rejqb205/nsamples;
     rejqb210=rejqb210/nsamples;

*print 'Tests of Homogeneity of Effect Sizes';

PRINT njs sds dlta kk specific b Tau2 nsamples;
```

```
*print 'sk= 0.00, kr=  0.00';
*SHAPE2  print 'sk= 1.00, kr=  3.00';
*SHAPE3  print 'sk= 1.50, kr=  5.00';
*SHAPE4  print 'sk= 2.00, kr=  6.00';
*SHAPE5  print 'sk= 0.00, kr= 25.00';

PRINT  rejq101, rejq105, rejq110;
*PRINT  rejreq101, rejreq105, rejreq110, rejreq201, rejreq205,
rejreq210;
PRINT  rejqb201, rejqb205, rejqb210;
PRINT rejprob_REZ101, rejprob_REZ105, rejprob_REZ110;
PRINT rejprob_FEZ101,rejprob_FEZ105,rejprob_FEZ110;
PRINT rejprob_conrand01, rejprob_conrand05, rejprob_conrand10;
quit;
```

About the Author

Lisa Aaron received a Bachelor's Degree in Psychology from Emory University in 1987 and a M.S. in Rehabilitation Counseling from Georgia State University in 1990. She practiced vocational rehabilitation counseling after graduation from the Master's program, until she entered the Ph.D. program at the University of South Florida in 1992.

While in the Ph.D. program at the University of South Florida, Ms. Aaron worked as a member of the design team for the Teaching of Higher Order Thinking. She also worked as the Supervisor of Title I Evaluation for the Pasco County School District, serving migrant and low socio-economic students. She has presented papers at regional and national meetings of the American Educational Research Association. She provides consultation to the Working for Success organization affiliated with Hesed House, dedicated to job placement of the homeless.