A POWER ANALYSIS OF THE TEST OF HOMOGENEITY IN EFFECT-SIZE META-ANALYSIS

By

Lin Chang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology,
and Special Education

1992

ABSTRACT

POWER ANALYSIS OF

THE TEST OF HOMOGENEITY IN META-ANALYSIS

By

Lin Chang

The power of homogeneity tests in both fixed- and random-effects models in effect-size meta-analyses is studied. Power functions are approximated and simulated. The impact of the power of the homogeneity test on statistical errors of subsequent tests of effect magnitude is also examined. The homogeneity test or $\underline{H}$ statistic had an asymptotic central chi-squared distribution when effect sizes were homogeneous. When the effect sizes were not homogeneous, under the fixed-effects model, the distribution of the $\underline{H}$ statistic was well approximated by a noncentral chi-squared distribution. The probability of a type I error (a false rejection) was higher than the preset α level when study effects were from many small samples. In order to maintain the desired significance level, meta-analysts were advised to lower the nominal type I error rate for reviews with many small samples. The non-null distribution of the homogeneity test $\underline{H}_+$ under the random-effects model is

approximated well by a combination of many noncentral chi-squared distributions. Power values were compared for subsequent tests of effect magnitude ($\underline{z}$ tests) calculated with the fixed-effects variance ($\underline{z}_F$) versus tests with the random-effects variance ($\underline{z}_R$) in the presence of a statistical error at stage one of testing. When the stage-one test of homogeneity was falsely accepted, the subsequent fixed-effects test ($\underline{z}_F$) was slightly more powerful than the appropriate random-effects test ($\underline{z}_R$). When the stage-one test of homogeneity was falsely rejected, the subsequent random-effects test ($\underline{z}_R$) was much less powerful than the correct fixed-effects test ($\underline{z}_F$). To prevent the random-effects test ($\underline{z}_R$) from being falsely applied, reviewers could either apply other approaches to prevent the use of the test until more is learned about the estimator of parameter variance used in the random-effects test, or reviewers could lower the Type I error rate (the possibility of false rejection) for the homogeneity test at stage one.

To my parents Yu-Tai and Jen-Pin Han Chang

ACKNOWLEDGEMENT

Many thanks are due to those who have supported me in the completion of this dissertation. I am grateful for God's sufficient mercy and provision.

First I thank my advisor Dr. Betsy Becker, who was also a great friend. She understood me well despite cultural differences and helped me learn to write statistical problems. I was often inspired by her persistent encouragement. I deeply appreciate her patience and availability to me, especially for the time and energy she spent with me outside her office hours.

I sincerely thank all my committee members, Dr. Steve Raudenbush for his consistent support and helpful suggestions; Dr. James Stapleton for his constructive advice and insightful assistance in the mathematical part of my dissertation; and, last but not least, Dr. Susan Phillips for her friendly encouragement and sincere concerns during the process of the completion of this paper. I also thank computer consultants Ryan Simmons and Randy Foutiu for their useful assistance in computer operation.

I thank my parents for the way they raised me and for their unconditional love for and trust in me. Finally, I thank my loving and supporting husband, Jacob Chi, for his tacit understanding and belief in me.

TABLE OF CONTENTS

Page

LIST OF TABLES x

LIST OF FIGURES xvii

CHAPTER

I. INTRODUCTION 1

Meta-analysis in Educational Research . . 1
Purpose of the Study 2
Need for a Power Study of the Homogeneity Test in Effect-size Meta-analysis 4
Definition of Statistical Power 4
Importance to the Test of Fit 6
Power of the Homogeneity Test 7
Need for a Power Study 7
Comparison to the Unbalanced Analysis of Variance Case 8

II. STATEMENT OF THE PROBLEM 11

Power of the Statistical Test in Empirical Research 11
Power of the Homogeneity Test in Meta-analysis 12

III. POWER OF HOMOGENEITY TESTS IN EFFECT-SIZE ANALYSES 17

Definitions and Notation 17
Population Effect Size 17
Glass's Estimate of Effect Size 17
Unbiased Estimate of Effect Size . . . 18
Analytical Approximation of Power 20
Effect-size Analyses for Fixed-Effects Models 20
Hypotheses 20
Homogeneity Test Statistic 21
Distribution of the Homogeneity Test for Fixed-effects Models 22
Theorem 23
Proof 23

Effect-size Analyses for Random-Effects Models 25
Hypotheses 27
Homogeneity Test Statistic 27
Distribution of the Homogeneity Test for Random-effects Models 28
Theorem 28
Proof 28

IV. SIMULATION OF THE DISTRIBUTIONS OF THE STATISTICS FOR POWER UNDER FIXED- OR RANDOM-EFFECTS MODELS 31

Parameters of the Simulation Study 31
Number of Effect Sizes 32
Sample Sizes 33
Population Effect Sizes 36
Variance of Population Effects 37
Design of the Simulation Study 37
Computation for Simulated Distributions . 40
Fixed-effects Models 40
Random-effects Models 41
Test for Goodness of Fit 42
Results 43
Power Discrepancies for Fixed-effects Models 43
Number of Effect Sizes ($\underline{k}$) 46
Sample Sizes ($\underline{N}$) 47
Sampling Fractions (π_i) 49
Sampling Ratios (ϕ_i) 51
Patterns of Effect-size Parameters. 52
Summary 58
Power Discrepancies on Random-effects Models 60
Power Analysis 64
Fixed-effects Model 64
Random-effects Model 77

V. THE INFLUENCE OF THE SIGNIFICANCE LEVEL AND POWER OF THE FIRST STAGE TEST ON THE SECOND STAGE TEST: A SEQUENTIALLY RELATED TESTING PROCEDURE 80

Two-stage Testing 82
Influence of Sequentially Related Hypothesis Testing on Statistical Errors 83
Acceptance of the Overall Homogeneity Test 83
Rejection of the Overall Homogeneity Test 85
Summary 86

Simulation of Power for Sequential Tests . 89
Factors for Simulation of Subsequent z Tests 90
Results 92
Simulated vs. Theoretical Power Values. 92
Fixed-effects Tests 93
Random-effects Tests 103
Summary 112
Power of z Based on Decisions about Homogeneity 116
Homogeneous Population Effects . . 116
Heterogeneous Population Effects . 119
Summary 120
Adjustment to Maintain the Desired Testing Error Rates 120

VI. CONCLUSIONS AND IMPLICATIONS 125

Example 125
Summary 127
The Power of the Homogeneity Test 128
The Power of the z Test 129
Practical Implications 131
Suggestions for Further Research 131

APPENDIX A: CHOOSE NUMBER OF REPLICATIONS 133

APPENDIX B: SUPPLEMENTARY TABLES 134

APPENDIX C: POWER TABLES 144

APPENDIX D: FIGURES 165

APPENDIX E: LIST OF SYNTHESIZED STUDIES 192

BIBLIOGRAPHY 195

LIST OF TABLES

Table		Page
1.	Sampling Fractions for Power Study	35
2.	Paired $\underline{t}$ Test between Theoretical and Simulated Power (α = 0.05) for Fixed-effects Model	44
3.	Crosstabulation of Discrepancies and $\underline{k}$ . .	45
4.	Crosstabulation of Significant Discrepancies by $\underline{k}$	46
5.	Crosstabulation of Significant Discrepancies by Sample Size	48
6.	Crosstabulation of Significant Discrepancies by $\underline{N}$ and $\underline{k}$	49
7.	Crosstabulation of Significant Discrepancies by $\pi_{\underline{i}}$	50
8.	Crosstabulation of Significant Discrepancies by $\pi_{\underline{i}}$ and $\underline{k}$	51
9.	Crosstabulation of Significant Discrepancies by $\phi_{\underline{i}}$	52
10.	Crosstabulation of Significant Discrepancies by $\phi_{\underline{i}}$ and $\underline{k}$	52
11.	Crosstabulation of Significant Discrepancies by Pattern of $\delta_{\underline{i}}$s	53
12.	Crosstabulation of Significant Discrepancies by Pattern of $\delta_{\underline{i}}$s and $\underline{k}$	54
13.	Crosstabulation of Significant Discrepancies by Pattern of $\delta_{\underline{i}}$s, $\underline{N}$, and $\underline{k}$	55
14.	Means of Significant Discrepancies by Pattern of $\delta_{\underline{i}}$s, $\underline{N}$, and $\underline{k}$	58

15. Paired t Test between Theoretical and Simulated Power for Random-effects Model 60

16. Frequency Table for Significant Discrepancies for Random-effects Model 62

17. Analysis of Variance for Power of H . . . 67

18. Means of Theoretical Power of H by Pattern of δ_is, N, and k 68

19. Means of Simulated Power of H by Pattern of δ_is by N and k 69

20. Means of Simulated Power of H for Homogeneous δ_is by N by k ($\delta = 0$) 70

21. ANOVA on Power of H for δ_is with One Extreme Value 71

22. Mean of Power of H for δ_is with One Extreme Value by N and k 71

22.a Mean of Simulated Power of H for δ_is with One Extreme Value by N and k 72

23. ANOVA on Power of H for δ_is with Two Extreme Values 72

24. Mean of Power of H for δ_is with Two Extreme Values by N and k 73

24.a Mean of Simulated Power of H for δ_is with Two Extreme Values by N by k 73

25. ANOVA on Power of H for Three Equal Subsets of δ_is 74

26. ANOVA on Power of H for Three Equal Subsets of δ_is by N and k 74

26.a Mean of Simulated Power of H for Three Equal Subsets of δ_is by N and k 75

27. ANOVA on Power of H for Five Equal Subsets of δ_is 75

28. Mean of Power of H for Five Equal Subsets of δ_is by N and k 76

28.a Mean of Simulated Power of H for Five Equal Subsets of δ_is by N and k 76

29. Mean of Power of $\underline{H}_+$ at $\alpha = 0.05$ for $\mu_\delta = 0$ for the Random-effects Model 79

30. Two-Stage Testing Errors 87

31. Paired $\underline{t}$ Tests on Mean Theoretical and Simulated $\underline{z}_F$ Power for Homogeneous Effects with $\delta = 0$ ($\alpha = 0.05$) 94

32. Paired $\underline{t}$ Tests on Mean Theoretical and Simulated $\underline{z}_F$ Power for Homogeneous Effects with $\delta > 0$ ($\alpha = 0.05$) 95

33. Frequencies of Significant Discrepancies for Power of $\underline{z}_F$ by $\underline{k}$ for Homogeneous Effects with $\delta > 0$ 96

34. Frequencies of Significant Discrepancies for Power of $\underline{z}_F$ by $\underline{N}$ for Homogeneous Effects with $\delta > 0$ 96

35. Frequencies of Significant Discrepancies for Power of $\underline{z}_F$ by $\pi_{\underline{i}}$ for Homogeneous Effects with $\delta > 0$ 97

36. Frequencies of Significant Discrepancies for Power of $\underline{z}_F$ by $\phi_{\underline{i}}$ for Homogeneous Effects with $\delta > 0$ 97

37. Frequencies of Significant Discrepancies for Power of $\underline{z}_F$ by δ for Homogeneous Effects with $\delta > 0$ 98

38. Paired $\underline{t}$ Tests on Theoretical and Simulated Power of $\underline{z}_F$ for Heterogeneous Effects . . 99

39. Frequencies of Significant Discrepancies of Power of $\underline{z}_F$ for Heterogeneous Effects . . 100

40. Frequencies of Significant Discrepancies for Power of $\underline{z}_F$ by $\underline{N}$ for Heterogeneous Effects 101

41. Significant Discrepancies for Power of $\underline{z}_F$ by Pattern of $\delta_{\underline{i}}$ for Heterogeneous Effects. . 101

42. Frequencies of Significant Discrepancies by Power of $\underline{z}_F$ by $\pi_{\underline{i}}$ for Heterogeneous Effects 102

43. Frequencies of Significant Discrepancies for Power of z_F by ϕ_i for Heterogeneous Effects 102

44. Paired t Tests on Mean Theoretical and Simulated Power of z_R for Homogeneous Effects with $\delta = 0$ ($\alpha = 0.05$) 104

45. Paired t Tests on Mean Theoretical and Simulated Power of z_R for Homogeneous Effects with $\delta > 0$ ($\alpha = 0.05$) 105

46. Frequencies of Significant Discrepancies for Power of z_R by k for Homogeneous Effects with $\delta > 0$ 106

47. Frequencies of Significant Discrepancies for Power of z_R by N for Homogeneous Effects with $\delta > 0$ 107

48. Frequencies of Significant Discrepancies for Power of z_R by π_i for Homogeneous Effects with $\delta > 0$ 107

49. Frequencies of Significant Discrepancies for Power of z_R by ϕ_i for Homogeneous Effects with $\delta > 0$ 108

50. Frequencies of Significant Discrepancies for Power of z_R by δ for Homogeneous Effects with $\delta > 0$ 108

51. Paired t Tests on Theoretical and Simulated Power of z_R for Heterogeneous Effects . . 109

52. Frequencies of Significant Discrepancies for Power of z_R by k for Heterogeneous Effects 110

53. Frequencies of Significant Discrepancies for Power of z_R by N for Heterogeneous Effects 111

54. Frequencies of Significant Discrepancies for Power of z_R by π_i for Heterogeneous Effects 111

55. Frequencies of Significant Discrepancies for Power of z_R by ϕ_i for Heterogeneous Effects 112

56. Significant Discrepancies for Power of z_R by Pattern of δ_i for Heterogeneous Effects 112

57. Paired t Tests on Power (sizea0 of z_F versus z_R for Heterogeneous Effects with $\delta = 0$ ($\alpha = 0.05$) and Homogeneity Was Rejected. . 117

58. Mean z Power values of z_F versus z_R for Homogeneous Effects with $\delta > 0$ ($\alpha = 0.05$) and Homogeneity Was Rejected 118

59. Mean z Power values of z_F versus z_R for Heterogeneous Effects ($\alpha = 0.05$) and Homogeneity Was Accepted 119

60. Computation of Noncentrality Parameter for the One-Extreme-Value Example 126

61. Computation of Noncentrality Parameter for the Three-Equal-Values Example 126

62. Values of Sample Sizes Used in the Simulation 134

63. Values of δ_is used in the Simulation for $k = 2$ 140

64. Values of δ_is used in the Simulation for $k = 5$ 140

65. Values of δ_is used in the Simulation for $k = 10$ 141

66. Values of δ_is used in the Simulation for $k = 30$ 142

67. Mean of Power for δ_is with One Extreme Value by N and k ($\alpha = 0.10$) 144

67.a Mean of Simulated Power for δ_is with One Extreme Value by N and k ($\alpha = 0.10$) . . . 145

68. Mean of Power for δ_is with Two Extreme Values by N and k ($\alpha = 0.10$) 146

68.a Mean of Simulated Power for δ_is with Two Extreme Values by N and k ($\alpha = 0.10$) . . . 146

69. Mean of Power for Three Equal Subsets of δ_is by N and k ($\alpha = 0.10$) 147

69.a Mean of Simulated Power for Three Equal Subsets of δ_is by N and k ($\alpha = 0.10$) . . . 148

70. Mean of Power for Five Equal Subsets of δ_is by N and k ($\alpha = 0.10$) 149

70.a Mean of Simulated Power for Five Equal Subsets of δ_is by N and k ($\alpha = 0.10$) . . . 149

71. Mean of Power for δ_is with One Extreme Value by N and k ($\alpha = 0.025$) 150

71.a Mean of Simulated Power for δ_is with One Extreme Value by N and k ($\alpha = 0.025$) . . . 151

72. Mean of Power for δ_is with Two Extreme Values by N and k ($\alpha = 0.10$) 152

72.a Mean of Simulated Power for δ_is with Two Extreme Values by N and k ($\alpha = 0.025$) . . 152

73. Mean of Power for Three Equal Subsets of δ_is by N and k ($\alpha = 0.025$) 153

73.a Mean of Simulated Power for Three Equal Subsets of δ_is by N and k ($\alpha = 0.025$) . . 154

74. Mean of Power for Five Equal Subsets of δ_is by N and k ($\alpha = 0.025$) 155

74.a Mean of Simulated Power for Five Equal Subsets of δ_is by N and k ($\alpha = 0.025$) . . 155

75. Mean of Power for δ_is with One Extreme Value by N and k ($\alpha = 0.01$) 156

75.a Mean of Simulated Power for δ_is with One Extreme Value by N and k ($\alpha = 0.01$) . . . 157

76. Mean of Power for δ_is with Two Extreme Values by N and k ($\alpha = 0.01$) 158

76.a Mean of Simulated Power for δ_is with Two Extreme Values by N and k ($\alpha = 0.01$) . . . 158

77. Mean of Power for Three Equal Subsets of δ_is by N by k ($\alpha = 0.01$) 159

77.a Mean of Simulated Power for Three Equal Subsets by N and k ($\alpha = 0.01$) 160

78. Mean of Power for Five Equal Subsets of δ_is by N and k ($\alpha = 0.01$) 161

78.a Mean of Simulated Power for Five Equal Subsets of δ_is by N and k ($\alpha = 0.01$) . . . 161

79. Mean of Power of H_+ at $\alpha = 0.10$ for $\mu_\delta = 0$ for the Random-effects Model 162

80. Mean of Power of H_+ at $\alpha = 0.025$ for $\mu_\delta = 0$ for the Random-effects Model 163

81. Mean of Power of H_+ at $\alpha = 0.01$ for $\mu_\delta = 0$ for the Random-effects Model 164

LIST OF FIGURES

Figure		Page
4.1.0	Frequencies of Absolute Significant Discrepancies	57
4.1.1	Power Curve with $\underline{k}$ = 2 (α = 0.05) for Fixed-effects Models (One Extreme Value)	165
4.1.2	Power Curve with $\underline{k}$ = 5 (α = 0.05) for Fixed-effects Models (One Extreme Value)	166
4.1.3	Power Curve with $\underline{k}$ = 10 (α = 0.05) for Fixed-effects Models (One Extreme Value)	167
4.1.4	Power Curve with $\underline{k}$ = 30 (α = 0.05) for Fixed-effects Models (One Extreme Value)	168
4.2.1	Power Curve with $\underline{k}$ = 10 (α = 0.05) for Fixed-effects Models (Two Extreme Values) . . .	169
4.2.2	Power Curve with $\underline{k}$ = 30 (α = 0.05) for Fixed-effects Models (Two Extreme Values) . . .	170
4.3.1	Power Curve with $\underline{k}$ = 5 (α = 0.05) for Fixed-effects Models (Three Equal Values) . . .	171
4.3.2	Power Curve with $\underline{k}$ = 10 (α = 0.05) for Fixed-effects Models (Three Equal Values) . . .	172
4.3.3	Power Curve with $\underline{k}$ = 30 (α = 0.05) for Fixed-effects Models (Three Equal Values) . . .	173
4.4.1	Power Curve with $\underline{k}$ = 10 (α = 0.05) for Fixed-effects Models (Five Equal Values)	174
4.4.2	Power Curve with $\underline{k}$ = 30 (α = 0.05) for Fixed-effects Models (Five Equal Values)	175
4.5.1	Power Curve with $\underline{k}$ = 2 (α = 0.05) for Random-effects Models with μ_δ = 0	176
4.5.2	Power Curve with $\underline{k}$ = 5 (α = 0.05) for Random-effects Models with μ_δ = 0	177

4.5.3 Power Curve with $\underline{k}$ = 10 (α = 0.05) for Random-effects Models with μ_δ = 0 178

4.5.4 Power Curve with $\underline{k}$ = 30 (α = 0.05) for Random-effects Models with μ_δ = 0 179

4.6.1 Power Curve with $\underline{k}$ = 2 (α = 0.05) for Random-effects Models with μ_δ = 0.10 180

4.6.2 Power Curve with $\underline{k}$ = 5 (α = 0.05) for Random-effects Models with μ_δ = 0.10 181

4.6.3 Power Curve with $\underline{k}$ = 10 (α = 0.05) for Random-effects Models with μ_δ = 0.10 182

4.6.4 Power Curve with $\underline{k}$ = 30 (α = 0.05) for Random-effects Models with μ_δ = 0.10 183

4.7.1 Power Curve with $\underline{k}$ = 2 (α = 0.05) for Random-effects Models with μ_δ = 0.25 184

4.7.2 Power Curve with $\underline{k}$ = 5 (α = 0.05) for Random-effects Models with μ_δ = 0.25 185

4.7.3 Power Curve with $\underline{k}$ = 10 (α = 0.05) for Random-effects Models with μ_δ = 0.25 186

4.7.4 Power Curve with $\underline{k}$ = 30 (α = 0.05) for Random-effects Models with μ_δ = 0.25 187

4.8.1 Power Curve with $\underline{k}$ = 2 (α = 0.05) for Random-effects Models with μ_δ = 0.50 188

4.8.2 Power Curve with $\underline{k}$ = 5 (α = 0.05) for Random-effects Models with μ_δ = 0.50 189

4.8.3 Power Curve with $\underline{k}$ = 10 (α = 0.05) for Random-effects Models with μ_δ = 0.50 190

4.8.4 Power Curve with $\underline{k}$ = 30 (α = 0.05) for Random-effects Models with μ_δ = 0.50 191

CHAPTER I

INTRODUCTION

Meta-analysis in Educational Research

The application of quantitative methods in synthesizing and analyzing the results of related studies has been of growing interest to researchers in the social sciences. As the number of related studies increases, drawing conclusions about research questions becomes less straightforward than it has been. Study results may be consistent with or contradictory to each other. Features of the related studies including sample sizes, experimental treatment conditions, and sampled populations differ from study to study. Drawing reasonable conclusions from those related yet varied studies is the challenge for researchers.

Research reviewers utilize the results of many related studies rather than results of single studies to draw inferences. Such synthetic research is known as "meta-analysis", a term coined by Glass (1976) to mean the "analysis of analyses."

Various methods of research synthesis have been used for many decades (e.g., since Tippett, 1931). The procedure of meta-analysis in the social sciences was popularized by Glass (1976), and has been developed by Rosenthal (1978), Rosenthal and Rubin (1979), Pillemer and Light (1980),

Cooper (1982), Hedges and Olkin (1985) and others in the last decade. This work has enabled research syntheses to become quantitatively more precise through the analysis of standardized effect sizes from primary studies.

Chang and Becker (1987) examined an empirical application of three main approaches in meta-analysis: vote counts and vote-counting estimation procedures (e.g., Hedges, 1986; Hedges & Olkin, 1980, 1985), tests of combined significance (e.g., Fisher, 1932; Rosenthal, 1978; Tippett, 1931), and analyses of effect sizes (e.g., Hedges & Olkin, 1985). Chang and Becker compared the hypotheses, statistical properties, and possible conclusions drawn from the three approaches. In contrasting these methods, they identified several areas for further research, noting in particular a lack of information on the power of tests of homogeneity of effect-size analyses.

Purpose of the Study

The purpose of this research is to study the power of tests of homogeneity in effect-size analyses. The power of homogeneity tests in both fixed- and random-effects models in meta-analyses is studied. Power functions are approximated and simulated. In addition, since typical effect-size analyses involve tests for at least two stages, the influence of the power of the homogeneity test on the statistical errors of the subsequent tests is examined.

Power analysis of statistical tests is essential and often ignored by empirical researchers (Brewer, 1972; Cohen, 1962, 1973, 1977; Daly & Hexamer, 1983; and Sedlmeier & Gigerenzer, 1989). Without information on power, interpretation of the results of statistical tests can be very difficult. A null hypothesis may be accepted either because the null hypothesis is true, or because the statistical test had insufficient power to detect a true alternative hypothesis, or because by chance the result was small by sampling error even when the test had sufficient power. Brewer (1972) and Cohen (1962, 1965) found that the neglect of power analysis has resulted in generally low power in research. Brewer argued that lower power affects the validity of what otherwise would be a proper rejection of H_0 based on the research data. Cohen (1973) emphasized power analysis as "the only rational guide to planning the relevant details of the research" (p. 227).

This study approximates power functions and serves empirical meta-analysts by enabling them to estimate the power of their statistical tests against an array of possible outcomes. I will do a numerical simulation of power values for homogeneity tests in effect-size meta-analyses. Comparisons will be made between power values calculated through theoretical approximations and simulated values. Power tables will be constructed. The influence of the power of the homogeneity test on subsequent effect-

magnitude tests will also be examined. Below, I start by briefly reviewing the concept of power and discussing the importance of power analysis, especially for homogeneity tests.

Need for a Power Study of Homogeneity Tests in Meta-analysis

Definition of Statistical Power

Two types of error are involved in statistical hypothesis testing. The type I error occurs if the researcher rejects a null hypothesis when the null hypothesis is actually true. A researcher commits a type II error when accepting (failing to reject) a false null hypothesis. The probability of the type I error is usually denoted as α, whereas the probability of the type II error is denoted as β. Statistical power is defined as the probability of rejecting a false null hypothesis, and is denoted $1 - \beta$.

Educational researchers have tended to be more concerned about type I errors than about type II errors. In setting α, the researcher imagines the null hypothesis to be true and then considers the risk of falsely rejecting H_0. On the other hand, in considering power, the researcher imagines the treatment to have "the minimum effect size" worth detecting and then considers the risk of falsely accepting H_0. Researchers limit the probability of a type I error by setting low α levels, such as .05, .01, etc. Given

certain preset or fixed α levels, they then try to increase power. For instance, they may increase sample sizes to increase the statistical power $(1 - \beta)$.

By setting low α levels rather than controlling the β level, educational researchers are conservative about accepting a new alternative hypothesis over an existing null hypothesis. The existing null hypothesis will be retained unless there is enough evidence against it. This conservative attitude in considering new alternative hypotheses in educational settings is often practical. It reflects concern over possible extra time or extra cost if changes are involved. Nevertheless, the tradeoff for a conservative attitude is the increased possibility of making a type II error.

This conservative attitude is reasonable in the context of rejecting the null hypothesis, because rejecting a null hypothesis does not cause a type II error. However, when the null hypothesis is accepted (which sometimes results from a "conservative attitude"), one needs to have reasonably high power in order to be comfortable that the acceptance of the null hypothesis implies a small or non-existent effect. Thus, apart from limiting the type I error, a power analysis is always valuable in research planning.

Empirical researchers often may not report the power of their statistical tests for two reasons. First, the power

functions of some tests are not available, and second, some researchers do not emphasize the importance of power.

Importance to the Test of Fit

The type I error is of primary concern and is often used as the criterion for decisions in statistical tests. However, one needs to be as concerned or more concerned about limiting the type II error when testing for fit.

The purpose of tests of "fit" is to test the hypothesis that certain expectations about a distribution (under H_0) are correct and that the obtained data are actually from the population specified by the hypothetical model (Hays, 1981). The difference between tests of fit and other tests is an implied "attitude." In the ordinary test, researchers usually accept H_0 unless the treatment effect is significantly large. Therefore, researchers limit α values in ordinary tests. In the test of fit, one tends to accept H_a unless the obtained data fit H_0. That is, the researcher assumes the data do not fit and seeks evidence that they do (i.e., seeks to accept H_0). Logically, one should limit β in the test of fit. If applying a "conservative attitude" to the tests of fit, researchers should limit β rather than α, because in the tests of fit, the conservative researcher would rather "accept" H_a. Hence, to be consistent with a "conservative attitude," one would emphasize statistical power $(1 - \beta)$ more in testing for fit than in ordinary tests. Also, since the test of fit is usually a preliminary

test to other tests, for one to proceed comfortably with the assumption of data being "fit" the power of the test of fit need to be high.

Power of the Homogeneity Test in Effect-size Meta-analysis

The simplest homogeneity test in meta-analysis (Hedges & Olkin, 1985) examines whether all the studies share a common effect size. Unless the effect sizes are shown to be homogeneous, they are treated as heterogeneous. Thus, the homogeneity test can be viewed as a test of fit. A power study for the homogeneity test is important because the homogeneity test is a test of fit. An analysis of the power of homogeneity tests in meta-analyses not only will aid our understanding of how homogeneity tests relate to other meta-analysis summaries, as suggested by Chang and Becker (1987) but also is essential per se.

Need for a Power Study

A power study can provide more understanding about the homogeneity test. Practically, a power analysis can examine how sensitive the test of homogeneity in meta-analysis is to such important factors as the number of studies to be integrated, sample sizes in each study, magnitudes of effect sizes, and other factors. Thus, the examination of the power of the homogeneity test is significant for both theoretical and practical reasons. Based on the results of this study, meta-analysts will be able to estimate the statistical power of the homogeneity test prior to their

analysis, recognize factors influencing the power of the test, and when possible choose appropriate values for those influencing factors which can be manipulated to maintain reasonable levels of power in their applications. Even if they are unable (or choose not) to manipulate factors, researchers will at least be able to evaluate how much power they can obtain, based on this power analysis.

Comparison to the Unbalanced Analysis of Variance Case

Parallels can be drawn between research synthesis and the analysis of variance (ANOVA). Hypothesis testing in ANOVA involves certain assumptions: observations are random samples drawn from normally distributed populations; the numerator and denominator of the F ratio are independent and (under H_0) estimate the same population variance , σ^2_ϵ. In ANOVA models, the total variation in scores is partitioned. For example, the simplest ANOVA model partitions the total variation into two parts, the between-groups variation and the within-group variation. The ratio of the between-groups variation to the within-group variation has an F-distribution (under H_0) and is used to test, for example, the hypothesis of equal group means in the one-way case.

As with the analysis of variance, there are two models for the population parameters in meta-analysis: the fixed-effects case, and the random-effects case. In the fixed-effects case, the population effect sizes are assumed to be

constants (or the variance of population effect size is zero). By contrast in the random-effects case, the population effect sizes are random variables. Therefore, in the random-effects case, population effect-sizes have a variance greater than zero.

In combining results, studies have been treated as a blocking variable (Snedecor & Cochran, 1967; and Rosenthal, 1978) in ANOVA. When the studies are regarded as a random factor and when the Treatment × Studies effect is large, this interaction effect is used as the appropriate error term. In the fixed-effects case for effect sizes, Hedges and Olkin (1985) and others (e.g., Pigott, 1986) also have drawn analogies between the effect-size meta-analysis and the analysis of variance.

However, for combining studies, the homogeneity test proposed by Hedges and Olkin is often more accurate than the F based on the Treatment × Studies effect as an index of the extent to which effect sizes vary across the groups. This statement is true primarily because in combining studies, the scales of measurement of the variables usually are not the same across studies, whereas in ordinary ANOVA, treatment groups within an experiment or study usually are measured on the same scale.

Also, the assumption of the homogeneity of variance for ANOVA is often violated when standard (unweighted) ANOVA is applied to meta-analysis data. Studies in meta-analysis

thus often cannot be treated as blocks in an ANOVA where the assumption is that comparable measurements are used. However, weighted ANOVA where scores are weighted by their precision would be appropriate, or if all of the reviewed studies measure the outcome variable on a single metric and if sample sizes (*N*s) are same (i.e., if homogeneity of variance exists) then one could use the "treatment × blocks (studies)" ANOVA to examine whether different studies have different treatment effects.

Caution needs to be taken in making homogeneity of variance assumptions in meta-analysis. In combining studies, the sample sizes of the studies are almost always different across studies. When studies do have equal sample sizes, one might treat the study effects as having equal variances (which depend mainly on the sample sizes). However, more realistically, most studies will not be based on the same sample sizes, thus the homogeneity of variance in combining studies cannot typically be assumed. Therefore, Hedges and Olkin's homogeneity tests for effect-size meta-analysis are often necessary, and usually more accurate than *F* tests in ANOVA. The homogeneity test proposed by Hedges and Olkin does not require the assumption of homogeneity of variance across the effect sizes. And the homogeneity test can be applied to studies with unequal sample sizes.

CHAPTER II

STATEMENT OF THE PROBLEM

Power of the Statistical Test in Empirical Research

As Cohen (1962) indicated nearly three decades ago, the power of statistical tests in empirical research is rarely reported. This is still true today. Though many researchers have recognized the importance of statistical power, few estimate and report the power of statistical tests in their studies. For example, a review of studies for the last ten years in the Journal of Research in Science Teaching (1980-1990) shows that few researchers (less than 5%) report power based on their proposed treatment effects or sample sizes.

Theoretically, the power of a statistical test to detect some alternative hypotheses (versus a given null hypothesis) should be computed before the initiation of a study. Without information on power, the test's conclusion may be questionable. When the power of a test is reasonably high, the decision about the hypothesis is likely to be a valid one. However, when the power of a test is low, the decision about the hypothesis may be confounded and confusing. Specifically, when the probability of rejecting the null hypothesis is low, the null hypothesis may be

accepted because it is true or because of low power. Tversky and Kahneman (1971) even suggested that research studies can be wasteful, as the interpretation of results is quite difficult with tests having low power.

Overall (1969) argued that when a test has low power, the probability of rejecting the true null hypothesis (α) may be only slightly smaller than the probability of rejecting a false null hypothesis ($1 - \beta$). "As a consequence, false rejections of valid null hypotheses may constitute a large proportion of all significant results" (Overall, 1969, p. 286).

As defined in Bayes' theorem, the ratio of the probability of invalid rejection of H_0 to the total probability of rejecting H_0 depends upon (1) the simple α specified by the investigator, (2) the power of the test, and (3) the a priori probability that the null hypothesis is valid (Overall, 1969). With low power, and if "the a priori probability of validity for the null hypothesis is substantial, an even larger proportion of significant results may be due to chance" (Overall, 1969, p. 286). Overall's message supports the emphasis on the power analysis of the homogeneity test in combining studies.

Power of the Homogeneity Test in Meta-analysis

The test for homogeneity of effect sizes has been suggested of having "excessively high statistical power

(Hunter et al., 1982)". In detecting a true difference, the concept of a test being "too powerful" is often not a concern. A powerful test can have a problem when the false rejection rate (or the type I error rate) exceeds the nominal level. Alexander et al. (1989) examined the chi-square test of homogeneity of effect sizes when the test is applied to correlation coefficients. Their results showed that the test on untransformed *r*s has excessively high Type I error rates but the test performs nominally for Fisher's *r*-to-*z* transformation. However, the power of test for homogeneity of effect sizes are yet to be studied.

As mentioned above, in meta-analysis the effect-size analysis can involve two levels of statistical tests. Before testing the magnitude of the average of the effect sizes drawn from related studies, one typically examines whether the studies share a common effect size. The reviewer first tests the homogeneity of the effect sizes drawn from various studies; and then tests if the common or average effect shared by those studies is greater than zero.

Low statistical power from the first-stage homogeneity test can also affect the second-stage test of the magnitude of the common effect size. When power is low, the null hypothesis for the homogeneity test tends to be accepted; that is, the effect sizes from studies are assumed to be homogeneous. The subsequent test for the magnitude of the common effect size may be wrong (or misleading) if the

effect sizes were actually heterogeneous and this has not been detected.

In the extreme case, if the power of the homogeneity test is approximately zero, one would always falsely accept the hypothesis that the effects are from the same population (i.e., effects are homogeneous). Subsequent tests of effect magnitude would be based on the average effect size, which would be wrongly assumed to be the common effect. The test for the magnitude of the effect then will generally be too lenient, and the concept of the common effect is misleading. By assuming that the test of fit has adequate power, the researcher also assumes that subsequent tests will behave as they should. Thus a power analysis for the test of fit in meta-analysis has indirect benefits as well.

Another situation using two-stage testing involves the homogeneity-of-variance test in analysis of variance. Suppose the within-group variances σ^2_i are very different from group to group. In this case, the standard ANOVA would be unjustified. Here the researcher also goes through two stages: (1) testing homogeneity of variance across the groups; and (2) if homogeneity is retained, proceeding with the ANOVA. Testing at stage 2 will only be valid if the H_0 at stage 1 is true. In other words, the test at stage 2 will lack validity if the result in stage 1 is a type II error, wherein the H_0 of homogeneity is falsely retained.

A similar analogy is (1) testing the blocks by

treatment interaction in a two-way ANOVA design; and (2) if the interaction effect is judged to be zero, one can either (a) pool the interaction sum of squares into the error sum of squares, or (b) form a one-way model with treatment effect as the only factor by pooling sums of squares for blocks and the interactions into the error sum of squares. Fabian (1991) pointed out that to proceed as if the interactions were zero after rejecting the zero-interaction hypothesis may give incorrect decisions with a large probability. Fabian further studied whether considering the power of the test and obtaining information on the neglected interactions can provide improved methods for obtaining "(1) an interval estimate of one of the cell expectations, (2) a simultaneous interval estimate of the cell expectations, and (3) an estimate of the cell with the largest expectation" (p.362). Fabian concluded that replacing the two-way model by the one-way model is a better method.

In the effect-size meta-analyses, the goal is to estimate the overall average treatment effects. The procedures also differ from the ANOVA analogy. When effect sizes are determined to be consistent, the variation between the population effect sizes will too be ignored. However, instead of pooling error sum of squares as in the ANOVA, the fixed-effects model excluding the variation of population effect-sizes will be applied. Power of the homogeneity test is again important because one can examine whether similar

recommendation to the two-way ANOVA with blocks design will be made to the effect-size meta-analyses.

CHAPTER III

POWER OF HOMOGENEITY TESTS IN EFFECT-SIZE ANALYSES

In this section, notation and definitions are given for the statistics used in this paper. Second, procedures are outlined for effect-size meta-analyses for both fixed- and random-effects models. And third, the power of the tests of homogeneity in effect-size meta-analyses for both fixed- and random-effects models is studied.

Definitions and Notation

Population Effect Size

Consider the ith of a series of k studies each comparing two groups. The population effect size for the two groups within study i is defined as

$$\delta_i = (\mu_i^E - \mu_i^C) / \sigma_i, \qquad i = 1, \ldots, k, \tag{1}$$

where μ_i^E and μ_i^C are the population means in the ith study on some outcome variable Y_i, in the experimental and control groups, respectively, and σ_i is the common population standard deviation for study i.

Glass's Estimator of Effect Size

Glass's estimator of effect size is often used in integrative reviews. (Examples can be found in some reviews

in the Appendix.) Glass (1976) estimated the population effect size by the sample standardized mean difference. The formula for Glass's effect size for the ith study of a set of k studies is

$$g_i = (\bar{Y}_i^E - \bar{Y}_i^C)/S_i, \quad (2)$$

where $\bar{Y}_i^E$ and $\bar{Y}_i^C$ are the sample means in the ith study for the experimental and control groups, and S_i is the pooled sample standard deviation from the usual two-sample t test for experimental and control groups. We assume that Y_i^E, $i = 1,\ldots, n_i^E$, and Y_i^C, $i = 1,\ldots, n_i^C$, are independent and normal with means μ_i^E and μ_i^C, respectively, and common population variance σ_i^2. This is the usual t test assumption.

Unbiased Estimator of Effect Size

Glass's estimator of the population effect size is biased. Hedges (1981) obtained a corrected effect size d_i, which is the minimum variance unbiased estimator of δ_i. The unbiased estimator is approximately

$$d_i = c(m_i)g_i \quad (3)$$

where,

$$c(m_i) \approx 1 - 3/(4m_i - 1), \text{ and}$$
$$m_i = n_i^E + n_i^C - 2.$$

The large-sample distribution of d_i tends towards normality. Hedges and Olkin noted (1985, p. 86) that if n_i^E

and n_i^C increase at the same rate (that is, if n_i^E/N_i and n_i^C/N_i are fixed, where N_i is $n_i^E + n_i^C$) then the asymptotic distribution of d_i is normal with mean δ_i and asymptotic variance $\sigma^2(d_i)$. We may write

$$d_i \sim N(\delta_i, \sigma^2(d_i)), \tag{4}$$

where the variance of d_i is approximated by,

$$\sigma^2(d_i) = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{\delta_i^2}{2(n_i^E + n_i^C)}. \tag{5}$$

The variance of d_i, $\sigma^2(d_i)$, is estimated by $\hat{\sigma}^2(d_i)$, a sample estimate of $\sigma^2(d_i)$, where d_i is substituted for δ_i in formula (5). I do not use the notation $\sigma^2(\delta_i)$ to denote the variance of d_i, to avoid confusion with σ^2_δ introduced below. According to Hedges and Olkin (1985, p. 193; also Hedges, 1983), the exact conditional variance $\sigma^2(d_i|\delta_i)$ of d_i is

$$\sigma^2(d_i|\delta_i) = a_i/\tilde{n}_i + (a_i - 1)\,\delta^2{}_i, \tag{6}$$

where $\tilde{n}_i = n_i^E n_i^C / (n_i^E + n_i^C)$,

$$a_i = m_i(c(m_i))^2 / (m_i - 2),$$

and $m_i = n_i^E + n_i^C - 2$,

and $c(m_i)$ is approximated as in (3).

Analytical Approximation of Power

Effect-size Analyses for Fixed-Effects Models

In this section, I review methods for effect-size meta-analyses in the fixed-effects case. The procedures for analysis and the statistical tests used are briefly described. Full details are given in Hedges and Olkin (1985).

Hypotheses. In effect-size analyses, an estimate of effect size is first calculated for each study using (2) and (3) above. Combining these estimates, one can obtain an overall estimate of effect size. Reviewers are usually interested in testing the magnitude of the overall effect size. Typically one tests the null hypothesis of no effect.

Hedges (1982) indicated that if the underlying population effect sizes from a series of studies are not identical, representing the results of a set of studies by a single estimate of effect size can be misleading. Hedges developed a two-stage testing procedure for effect-size meta-analysis in the fixed-effects case. At the first stage, one tests the homogeneity of the effect sizes from all the collected studies, and decides if the studies share a common population effect size. If the studies are not homogeneous, the studies probably do not share a common population effect size. The reviewer next may attempt to "model" or describe the studies with categorical or regression models using study features as factors or

predictors or may decide to adopt a random-effects approach. If the studies are homogeneous, one can test the hypothesis that the magnitude of the common effect size equals zero at the second stage of testing.

Hypotheses examined in the two-stage testing are:

$$H_{01}: \quad \delta_1 = \delta_2 = \ldots = \delta_k = \delta, \text{ and} \tag{7}$$

$$H_{02}: \quad \delta = 0. \tag{7a}$$

Homogeneity test statistic. The statistic for the homogeneity test of H_{01}, proposed by Hedges and Olkin (1985), is

$$H = \sum_{i=1}^{k} \frac{(d_i - d.)^2}{\hat{\sigma}^2(d_i)} \overset{A}{\sim} \chi^2_{(k-1)}, \tag{8}$$

under H_{01}, and where

$$d. = \frac{\sum_{i=1}^{k} \hat{\sigma}^{-2}(d_i)\, d_i}{\sum_{i=1}^{k} \hat{\sigma}^{-2}(d_i)}, \tag{9}$$

is the average of d_is, weighted by the precision of each d_i.

Hedges and Olkin (1985, p. 112) noted that if the sample sizes of the experimental and control groups in each of the k studies, n_1^E, ..., n_k^E, n_1^C, ..., n_k^C, increase at the same rates (as n_i^E/N_i, n_i^C/N_i remain fixed, where N_i is the total sample size for study i), then the null distribution of $d.$ tends to normality with a mean

$$\delta_{\cdot} = \frac{\sum_{i=1}^{k} \sigma^{-2}(d_i)\ \delta_i}{\sum_{i=1}^{k} \sigma^{-2}(d_i)}, \tag{10}$$

and a variance

$$\sigma^2(d_{\cdot}) = \frac{1}{\sum_{i=1}^{k} \sigma^{-2}(d_i)}, \tag{11}$$

where $\sigma^2(d_i)$ is defined in (5).

When the hypothesis of homogeneity is retained, one tests H_{02} by drawing a normal confidence interval around the weighted average $d_{\cdot}$, or by doing a z test since $d_{\cdot}$ is asymptotically normally distributed with a mean of the common effect δ (if all δ_is equal δ, then $\delta_{\cdot} = \delta$), and a variance of $\sigma^2(d_{\cdot})$.

Distribution of the homogeneity test for fixed-effects models. As stated before, when the d_is are asymptotically normal, under the null hypothesis where the k studies share a common effect size, then the homogeneity test statistic, H has an approximate central chi-square distribution with (k-1) degrees of freedom. When the δ_is are not the same across the k studies, H has an approximate noncentral chi-square distribution with (k-1) degrees of freedom and a noncentrality parameter

$$\lambda_{\cdot} = \sum_{i=1}^{k} \frac{(\delta_i - \delta_{\cdot})^2}{\sigma^2(d_i)}, \qquad (12)$$

where $\delta_{\cdot}$ is the weighted mean of δ_is shown in (10).

Theorem. Let $d_1, \ldots, d_k$ be defined as in (4) and the homogeneity test H be defined as in (8). Then when H_0: $\delta_1 = \ldots = \delta_k = \delta$ is true, $H \sim \chi^2_{k-1}$, and when H_0 is false $H \sim \chi^2_{k-1}(\lambda_{\cdot})$ where $\lambda_{\cdot}$ is defined in (12).

Proof: We observe $d_1, \ldots, d_k$ independently, each with a mean δ_i, and a variance $\sigma^2(d_i)$, that is,

$$\underset{\sim}{d} = \begin{bmatrix} d_1 \\ \vdots \\ d_k \end{bmatrix} \overset{A}{\sim} N_k(\underset{\sim}{\delta}, \operatorname{diag}(\sigma^2(d_1), \ldots, \sigma^2(d_k))), \qquad (13)$$

where $\delta = (\delta_1, \ldots, \delta_k)'$. We wish to test the hypotheses

H_0: $\delta_1 = \delta_2 = \ldots = \delta_k = \delta$ versus (14)

H_a: At least one δ_i is different, for $i = 1, \ldots, k$.

The null hypothesis can be rewritten in matrix form as

$$H_0: \underset{\sim}{\delta} = \delta \underset{\sim}{J} \qquad (15)$$

for some constant δ, where $\underset{\sim}{\delta} = (\delta_1, \ldots, \delta_k)'$, and $\underset{\sim}{J} = (1, \ldots, 1)'$. Let $w_i = d_i/\sigma(d_i)$ denote d_i weighted by its precision (or, the inverse of its standard error), so that the vector of w_is is normally distributed with a mean vector of δ_is weighted by these precisions, denoted as vector $\underset{\sim}{\mu_w}$, and with a variance matrix equal to the $k \times k$ identity matrix, $\underset{\sim}{I_k}$. In matrix form,

$$\underset{\sim}{w} = \begin{bmatrix} \underline{d}_1/\sigma(\underline{d}_1) \\ \vdots \\ \vdots \\ \vdots \\ \underline{d}_k/\sigma(\underline{d}_k) \end{bmatrix} \tag{16}$$

$$= \begin{bmatrix} 1/\sigma(\underline{d}_1) & \cdots & 0 \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ 0 & \cdots & 1/\sigma(\underline{d}_k) \end{bmatrix} \underset{\sim}{d} \qquad \overset{A}{\underset{\sim}{}} \; N\,[\underset{\sim}{\mu}_{\underline{w}}, \underset{\sim}{I}_{\underline{k}}],$$

where $\underset{\sim}{\mu}_{\underline{w}} = (\delta_1/\sigma(\underline{d}_1), \ldots, \delta_{\underline{k}}/\sigma(\underline{d}_{\underline{k}}))'$. Let vector $\underset{\sim}{x}_0 = (1/\sigma(\underline{d}_1), \ldots, 1/\sigma(\underline{d}_{\underline{k}}))'$. Under the null hypothesis, $\underset{\sim}{\mu}_{\underline{w}} = \delta\underset{\sim}{x}_0$. The projection of vector $\underset{\sim}{w}$ on $\underset{\sim}{x}_0$ is

$$p(\underset{\sim}{w}|\underset{\sim}{x}_0) = \frac{\underset{\sim}{x}_0'\underset{\sim}{w}}{\|\underset{\sim}{x}_0\|^2}\,\underset{\sim}{x}_0 = \frac{\sum_{\underline{i}=1}^{\underline{k}} \underline{w}_{\underline{i}}/\sigma(\underline{d}_{\underline{i}})}{\sum_{\underline{i}=1}^{\underline{k}} 1/\sigma^2(\underline{d}_{\underline{i}})}\,\underset{\sim}{x}_0 \tag{17}$$

$$= \frac{\sum_{\underline{i}=1}^{\underline{k}} \underline{d}_{\underline{i}}/\sigma^2(\underline{d}_{\underline{i}})}{\sum_{\underline{i}=1}^{\underline{k}} 1/\sigma^2(\underline{d}_{\underline{i}})}\,\underset{\sim}{x}_0 = \underline{d}_{\cdot}\underset{\sim}{x}_0,$$

where $\underline{d}_{\cdot}$ is defined in (9). The projection of vector $\underset{\sim}{w}$ on the entire space other than the space spanned by vector $\underset{\sim}{x}_0$ is, by definition, the difference between the vector $\underset{\sim}{w}$ and its projection on vector $\underset{\sim}{x}_0$:

$$\underset{\sim}{w} - p(\underset{\sim}{w}|\underset{\sim}{x}_0) = \begin{bmatrix} \frac{d_i}{\sigma(d_1)} \\ \vdots \\ \frac{d_k}{\sigma(d_k)} \end{bmatrix} - \underline{d}. \begin{bmatrix} \frac{1}{\sigma(d_1)} \\ \vdots \\ \frac{1}{\sigma(d_k)} \end{bmatrix} = \begin{bmatrix} \frac{d_1 - \underline{d}.}{\sigma(d_1)} \\ \vdots \\ \frac{d_k - \underline{d}.}{\sigma(d_k)} \end{bmatrix}. \tag{18}$$

The vector $\underset{\sim}{w} - p(\underset{\sim}{w}|\underset{\sim}{x}_0)$ has multivariate normal distribution with a mean vector of $(\delta_i - \delta.)/\sigma(d_i)$. The squared length of the above vector is

$$\| \underset{\sim}{w} - p(\underset{\sim}{w}|\underset{\sim}{x}_0) \|^2 = \sum_{i=1}^{k} \frac{(d_i - \underline{d}.)^2}{\sigma^2(d_k)}, \tag{19}$$

which is asymptotically distributed as a noncentral chi-square with (k - 1) degrees of freedom and a noncentrality parameter, say $\lambda.$, where

$$\lambda. = \sum_{i=1}^{k} \frac{(\delta_i - \delta.)^2}{\sigma^2(d_i)}. \tag{20}$$

Under the null hypothesis where δ_is are equal and $\lambda.$ is zero, the H statistic is asymptotically distributed as a central chi-square with (k - 1) degrees of freedom.‖

Effect-size Analyses for Random-Effects Models

Unlike the fixed-effects case where the population effect sizes, the δ_is (i.e., δ_1, ..., δ_k), are fixed constants, in the random-effects case the δ_is are sampled from some population. Cronbach (1980) argued that in educational research each treatment site (or study) may be a

sample from some universe of related sites rather than from a single population. Under the random-effects model variations in treatments are viewed as more or less effective in producing an outcome. In other words, in the random-effects model there is no "single" true (population) effect. The true effects are from a distribution of effects with some variance.

Since random-effects models assume that true values of the effect sizes are sampled from a distribution, the sources of variation in observed effects are at least two. One is the variability in effect-size parameters in the population distribution of effects. Another is the variability in the estimator about the true parameter value for a particular study (due to sampling error).

The simplest case of a random-effects model specifies that $\underline{d}_1, \ldots, \underline{d}_k$ are conditionally normal. That is, each $\underline{d}_i$ given δ_i is approximately normal for the ith study. The distribution of δ_i values is often assumed to be normal, which implies that the unconditional distribution of $\underline{d}_i$ is also normal. The unconditional distribution of $\underline{d}_i$ is then:

$$\underline{d}_i \sim N(\mu_\delta, \sigma^2_\delta + \sigma^2(\underline{d}_i|\delta_i)), \quad (21)$$

where μ_δ is the expected value of the population effect-size values, σ^2_δ is the variance of the population distribution of effect sizes, and $\sigma^2(\underline{d}_i|\delta_i)$ is the variance of the conditional distribution of $\underline{d}_i$ given δ_i, and is described in

formula (5) and (6).

Hypotheses. The steps in testing for a random-effects model are, first, to estimate the mean effect size μ_δ (the population mean of the δs) and the variance $\sigma^2{}_\delta$ and, then, to test the hypothesis that $\sigma^2{}_\delta$ is zero. If $\sigma^2{}_\delta = 0$, then no variation exists among the δ_is, that is, the conditional variance of $\underline{d}_i$, $\sigma^2(\underline{d}_i|\delta_i)$, equals the unconditional variance of d_i, $\sigma^2(\underline{d}_i)$ in the fixed-effects model. A test of $\sigma^2{}_\delta = 0$ in the random-effects model corresponds to a test for homogeneity of effect sizes in the fixed-effects model. Hence, the following two hypotheses are the same:

$$H_0: \quad \sigma^2{}_\delta = 0, \text{ and} \tag{22}$$

$$H_0: \quad \delta_1 = \delta_2 = \ldots = \delta_k = \delta, \text{ for some } \delta.$$

Homogeneity test statistic. Under the above null hypothesis that the population effect sizes have no variation, the homogeneity test statistic is

$$\underline{H}_+ = \sum_{i=1}^{k} \frac{(\underline{d}_i - \underline{d}_+)^2}{\hat{\sigma}^2(\underline{d}_i|\delta_i)} \overset{A}{\sim} \chi^2{}_{(k-1)}, \tag{23}$$

where

$$\underline{d}_+ = \frac{\sum_{i=1}^{k} \hat{\sigma}^{-2}(\underline{d}_i|\delta_i)\ \underline{d}_i}{\sum_{i=1}^{k} \hat{\sigma}^{-2}(\underline{d}_i|\delta_i)}. \tag{24}$$

The estimate of the variance is obtained by substituting $\underline{d}_i$

for δ_i in the asymptotic variance in (5).

Distribution of the homogeneity test for random-effects models. The statistical power of the homogeneity test is the probability of rejecting a null hypothesis when the alternative hypothesis is true, that is, when the true variance of the δ_is is not zero. The distribution of the H_+ statistic under the alternative hypothesis is no longer a central χ^2, as under the null hypothesis that δ_is have no variation. However, it is not a simple noncentral χ^2 distribution either. It is a combination of many noncentral χ^2 distributions.

Theorem. Let $d_1, \ldots, d_k$ be defined as in (21) and the homogeneity test H_+ defined as in (23). Then when H_0: $\sigma_\delta^2 = 0$ is true, $H_+ \sim \chi^2_{k-1}$, and when H_0 is false H_+ is a combination of many $\chi^2_{k-1}(\lambda.)$ variates where $\lambda.$ is defined in (12).

Proof: Let $\sigma_{\theta i} = \sqrt{\sigma_\delta^2 + \sigma^2(d_i|\delta_i)}$, let $v_i = d_i/\sigma_{\theta i}$ denote d_i weighted by the square root of its precision, and let vector $\underset{\sim}{\mu_v} = \mu_\delta(1/\sigma_{\theta 1}, \ldots, 1/\sigma_{\theta k})'$ denote μ_δ weighted similarly, so that the vector $\underset{\sim}{v}$ of v_is is normally distributed with a mean vector $\underset{\sim}{\mu_v}$, and with a variance equal to the identity matrix, I_k. In matrix form,

$$\underset{\sim}{v} = \begin{bmatrix} d_1/\sigma_{\theta 1} \\ \vdots \\ \vdots \\ \vdots \\ d_k/\sigma_{\theta k} \end{bmatrix} \qquad (25)$$

$$= \begin{bmatrix} 1/\sigma_{\theta 1} & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 1/\sigma_{\theta k} \end{bmatrix} \underset{\sim}{d} \qquad \sim N [\underset{\sim}{\mu_v}, \underset{\sim}{I_k}],$$

where $\underset{\sim}{I_k}$ is an identity of dimension k. Let vector $\underset{\sim}{t_0} = (1/\sigma_{\theta 1}, \ldots, 1/\sigma_{\theta k})'$. Under the null hypothesis that $\sigma_\delta^2 = 0$, vector $\underset{\sim}{t_0}$ equals vector $\underset{\sim}{x_0}$ (as defined in the proof for the fixed-effects model), and vector $\underset{\sim}{v}$ is vector $\underset{\sim}{w}$ in formula (16) for the fixed-effects model. Thus, under the null hypothesis, the projection of vector $\underset{\sim}{v}$ on $\underset{\sim}{t_0}$ in the random-effects model equals the projection of vector $\underset{\sim}{w}$ on $\underset{\sim}{x_0}$ in fixed-effects:

$$p(\underset{\sim}{v}|\underset{\sim}{t_0}) = p(\underset{\sim}{w}|\underset{\sim}{x_0}) = \underset{\sim}{d}.\underset{\sim}{x_0}, \tag{26}$$

and

$$\underset{\sim}{v} - p(\underset{\sim}{v}|\underset{\sim}{t_0}) = \underset{\sim}{w} - p(\underset{\sim}{w}|\underset{\sim}{x_0}). \tag{27}$$

The squared length of the difference between vector $\underset{\sim}{v}$ and its projection on $\underset{\sim}{t_0}$ is thus distributed as a central chi squared with $(k - 1)$ degrees of freedom under H_0 as was H in the fixed-effects case.‖

However, the nonnull distributions of H_+ for random-effects models differ from that of H for fixed-effects. For fixed-effects models, the distribution of H under the alternative hypothesis is a noncentral chi-squared distribution. In random-effects models, the probability that $H_+ \leq h$ given the δ_is is an average over k dimensions:

$$E_{\underset{\sim}{\delta}}[P(\underline{H}_+ \leq \underline{h} | \delta_1, \ldots, \delta_{\underline{k}})] = P(\underline{H}_+ \leq \underline{h}), \tag{28}$$

for $\underset{\sim}{\delta} = (\delta_1, \ldots, \delta_{\underline{k}})$. For each possible $\underset{\sim}{\delta}$ vector from the population of $\delta_{\underline{i}}$s, $\underline{H}_+$ has a $\chi^2_{\underline{k}-1}(\lambda.)$ distribution with noncentrality parameter $\lambda.$ as in (12):

$$P(\underline{H}_+ \leq \underline{h} | \delta_{\underline{1}}, \ldots, \delta_{\underline{k}}) = P(\chi^2_{\underline{k}-1}(\lambda.) \leq \underline{h})] = F(\underline{h}; g(\underset{\sim}{\delta})), \tag{29}$$

where F is the cumulative density function of $\underline{H}_+$, and $g(\underset{\sim}{\delta})$ is the noncentrality parameter $\lambda.$ for the noncentral $\chi^2_{\underline{k}-1}$ distribution. Thus

$$E_{\underset{\sim}{\delta}}[P(\underline{H}_+ \leq \underline{h} | \delta_{\underline{1}}, \ldots, \delta_{\underline{k}})] = E_{\underset{\sim}{\delta}}[F(\underline{h}; g(\underset{\sim}{\delta}))]. \tag{30}$$

We can also write:

$$E_{\underset{\sim}{\delta}}[F(\underline{h}; g(\underset{\sim}{\delta}))] = \int \cdots \int F(\underline{h}; g(\underset{\sim}{\delta})) \, \underline{f}(\underset{\sim}{\delta}) \, d\underset{\sim}{\delta}, \tag{31}$$

where $\underline{f}(\underset{\sim}{\delta})$ is the normal density function of the $\delta_{\underline{i}}$s. The power of the random-effects homogeneity test is

$$1 - P(\underline{H}_+ \leq \underline{h}) = 1 - \int \cdots \int F(\underline{h}; g(\underset{\sim}{\delta})) \, \underline{f}(\underset{\sim}{\delta}) \, d\underset{\sim}{\delta}. \tag{32}$$

No simple form of the distribution of $\underline{H}_+$ under the alternative in the random-effects case can be written.

CHAPTER IV

SIMULATION OF THE DISTRIBUTIONS OF THE STATISTICS FOR POWER UNDER FIXED- OR RANDOM-EFFECTS MODELS

In this Chapter the asymptotic distributions of the homogeneity statistics $\underline{H}$ and $\underline{H}_+$ (for fixed- and random-effects models) are compared to numerical simulations of those distributions. Specifically, differences between cumulative density functions of chi-squared distributions (with $\lambda_\cdot \geq 0$) and simulated cumulative density functions for $\underline{H}$ and $\underline{H}_+$ are examined. Confidence intervals are drawn for the differences at the 95% level. The parameters varied in the simulation include (1) the significance criterion (α level), (2) the noncentrality parameter of the chi-square density (the degree to which H_0 is false), (3) the number of effect sizes ($\underline{k}$), and (4) the sample sizes ($\underline{n}$). It is known that, other things being equal, power increases as sample size increases. The same relationship exists between the power and the effect size, and between power and α levels.

Parameters of the Simulation Study

An empirical study of published reviews suggested values for the parameters of the simulation study. Practical ranges for variables in the simulation were

designed by reviewing a random sample of twenty published meta-analyses (see Appendix E). Many of these twenty meta-analyses did not report sufficient information on the original studies they reviewed to inform the selection of variable values for the simulation. Therefore, I examined about 40 more reviews in Review of Educational Research from the middle of 1985 to the beginning of 1990 (volumes 55(2) through 59(3)).

Factors examined included the following: the number of studies (or number of independent effect sizes), k; the magnitude of effect sizes (d_i), the sample variance of simulated effect sizes (S^2_δ), the sample size of the experimental group for each study i, n_i^E; and the sample size of the control group for each study i, n_i^C. From these factors values of the population effect sizes, δ_i; the variances of population effects, σ^2_δ; and the significance level, α; were chosen for the simulation.

Number of Effect Sizes

In contrast to previously examined reviews (Becker, 1985), the reviews examined here tended to include more studies, that is, to have larger k values. Of reviews that reported information about individual studies, approximately one fourth included more than one hundred studies, and about one fourth analyzed fewer than twenty. One tenth of the reviews contained fewer than ten studies. Very rarely, the homogeneity test was applied to only two studies ($k = 2$).

Although the k values (numbers of studies) were generally quite large in this set of reviews, power studies have often been performed assuming small numbers of studies. For this reason, a broader range of k values (k = 2, 5, 10, and 30) was selected for this power study.

Sample Sizes

Based on the empirical study, study sample sizes ($n = \Sigma n_i/k$, i = 1, ..., k) of 20 (e.g., 10 in each experimental or control group), 60, 120 and 200 were selected. In empirical reviews, studies rarely have equal sample sizes. The sample-size values in the simulation were determined by the total sample size across studies (N), the total sample sizes of each study (n_i, i = 1, ..., k), the sampling fractions ($\pi_i = n_i/N$, i = 1, ..., k), and the ratio of the size of the experimental group over the total sample size of a study ($\phi_i = n_i^E/n_i$, i = 1, ..., k).

For example, in the case of k = 2, with a series sample size of N = 40, with sampling fractions (π_1, π_2) = (.5, .5) and (.3, .7), and within-study sampling fractions (ϕ_1, ϕ_2) = (.5, .5) and (.35, .35), the simulation will include the sets of parameters described below.

Sampling fraction (π_1, π_2) = (.5, .5) indicated that studies had equal sample sizes, that is, (n_1, n_2) = (20, 20). Two values of within-study sampling fractions determined the sample sizes for two sets of samples. For (ϕ_1, ϕ_2) = (.5, .5), samples were equal within studies. For

$(\phi_1, \phi_2) = (.35, .35)$, the ratio of the sample sizes of the experimental group over the total sample size within each study was 0.35 (and was the same across studies. Symbolically,

$$(\pi_1, \pi_2) = (.5, .5) \Rightarrow (\underline{n}_1, \underline{n}_2) = (20, 20)$$

Then,

$$(\phi_1, \phi_2) = (.5, .5) \Rightarrow \underline{n}_1^E = \underline{n}_1^C = 10 \text{ and } \underline{n}_2^E = \underline{n}_2^C = 10.$$

And for $(\phi_1, \phi_2) = (.35, .35)$, then

$$\underline{n}_1^E = 7, \underline{n}_1^C = 13 \text{ and } \underline{n}_2^E = 7, \underline{n}_2^C = 13.$$

Thus the combination of fixed values of $\underline{N}$ and (π_1, π_2), with the pair of (ϕ_1, ϕ_2) values produced two sets of sample sizes for the simulation.

Unequal sampling fractions such as $(\pi_1, \pi_2) = (.3, .7)$ indicated that some studies had larger sample sizes than others. In this example, the ratios of the study sample sizes over the total of the sample sizes for the two studies were 0.3 or 0.7. Thus for $\underline{N} = 40$, $(\underline{n}_1, \underline{n}_2) = (12, 28)$. The two values of within-study sampling fractions again determined the within-study sample sizes. Sampling fractions used within studies $(\phi_1, \phi_2) = (.5, .5)$ or $(.35, .35)$ were the same as outlined above. Thus

$$(\pi_1, \pi_2) = (.3, .7) \Rightarrow (\underline{n}_1, \underline{n}_2) = (12, 28)$$

Then,

$$(\phi_1, \phi_2) = (.5, .5) \Rightarrow \underline{n}_1^E = \underline{n}_1^C = 6 \text{ and } \underline{n}_2^E = \underline{n}_2^C = 14.$$

And for $(\phi_1, \phi_2) = (.35, .35)$, then

$$\underline{n}_1^E = 4,\ \underline{n}_1^C = 8 \text{ and } \underline{n}_2^E = 10,\ \underline{n}_2^C = 18.$$

The values of $\underline{N}$, $\underline{n}_i$, π_i, and ϕ_i were selected based on my empirical study of reviews. Total sample sizes across $\underline{k}$ studies with average sample size $\underline{n} = \Sigma\underline{n}_i/\underline{k}$ were $\underline{N} = \underline{n}*\underline{k}$, $20\underline{k}$, $60\underline{k}$, $120\underline{k}$, or $200\underline{k}$. Sampling fractions were the ratios of the sample sizes of each study to the total sample size across studies. Sampling fractions differed for each $\underline{k}$ and are listed in Table 1. Two values of the sampling fraction within studies were selected: 0.5, or 0.35. That is, experimental and control sample sizes were either balanced ($\phi_i = 0.5$) or unbalanced ($\phi_i = 0.35$) within studies. Specific numbers used for the simulation are listed in Table 62 in Appendix B.

Table 1

Sampling Fractions for Power Study

$\underline{k}$	$(\pi_1, \ldots, \pi_k)$
2	(.5 .5)
	(.3 .7)
5	(.2 .2 .2 .2 .2)
	(.15 .2 .2 .2 .25)
10	(.1 .1 .1 .1 .1 .1 .1 .1 .1 .1)
	(.05 .06 .07 .07 .08 .08 .09 .1 .15 .25)
30	(.03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03 .03)
	(.007 .01 .01 .01 .013 .02 .02 .02 .02 .02 .02 .023 .023 .023 .027 .027 .027 .027 .037 .037 .037 .04 .04 .047 .056 .056 .056 .067 .067 .113)

Population Effect Sizes

In the homogeneity test, the alternative hypothesis that "at least one population effect size differs" is a composite hypothesis. The number and complexity of possible alternative hypotheses makes the power study difficult. However, by examining past reviews, I have selected sets of typical values for δ_i. The conditions depicted include (1) the null hypothesis, where all the estimates of effect sizes share a common population parameter (δ), and (2) several alternative hypotheses, where at least one sampled effect size arises from a different population.

For example, the empirical reviews showed that effect sizes often vary from study to study. Thus, a typical pattern of the effect sizes shows a set of δ_i values that differ slightly from each other. Other possible sets of δ_i values are also suggested by the empirical study. One larger δ_i value with (k - 1) smaller δ_i values is of interest (Becker, 1985). The pattern of two larger δ_i values will also be studied when $k \geq 10$. Another pattern of interest is one in which the δ_i values are more evenly distributed, for example, having equal value within three or five equal subsets, but differing between subsets.

For the fixed-effects model, five patterns of δ_is were designed: (1) all equal to zero, (2) $\delta_1 = \ldots = \delta_{k-1} = 0$ and one nonzero value δ_k (taking values 0, 0.1, 0.25, 0.5, 0.75, and 1.0), (3) $\delta_1 = \ldots = \delta_{k-2} = 0$ and two nonzero δ_is

(δ_{k-1} and δ_k) (for $k \geq 10$), (4) three equal subsets of δ_is in which one subset contains zeros, and studies in the other two subsets share nonzero values δ and 2δ, respectively, and (5) five equal subsets of δ_is where, again, one subset contains zeros, and the other four subsets have nonzero values (of $\frac{1}{2}\delta$, δ, $1\frac{1}{2}\delta$, 2δ). The patterns of population effects used were:

(1) (0, ..., 0),

(2) (0, ..., 0, δ),

(3) (0, ..., 0, δ, δ),

(4) (0, ..., 0, δ, ..., δ, 2δ, ..., 2δ), and

(5) (0 ,..., 0, $\frac{1}{2}\delta$, ..., $\frac{1}{2}\delta$, δ,..., δ, $1\frac{1}{2}\delta$,..., $1\frac{1}{2}\delta$, 2δ,.., 2δ).

The population effect sizes used for the fixed-effects models are listed in Tables 63 to 66 in Appendix B.

Variances of Population Effects

Values of the variance of the population effect sizes (σ^2_δ) in the random-effects models were also suggested through the empirical study. Variance values selected for the random-effects models are 0.01, 0.03, 0.05, 0.07, 0.09, and 0.1.

Design of the Simulation Study

Combinations of the variables outlined above formed 992 patterns of simulation parameter values for fixed-effects models and about 2400 combinations for random-effects

models. The probability distribution of the homogeneity statistic was simulated for each combination of variables. Simulated distributions were compared with the corresponding asymptotic distribution at fifteen percentile points ($1-\alpha$): 0.05, 0.10(0.10)0.90 (i.e., from 0.10 to 0.90 with increment of 0.10), 0.95, 0.975, 0.99, 0.995, and 0.999. That is, 14880 simulated and theoretical power values were obtained from 992 combinations of parameters for fixed-effects models.

The simulation followed these procedures:

Case I. $\sigma^2_\delta = 0$, for fixed-effects models:

A. Generate 2000 replications (see rationale in Appendix A) of normal and chi-square deviates and compute $\underline{k}$ effect sizes ($\underline{d}_1$, ..., $\underline{d}_k$) for each combination of the parameters presented in Table 1.

B. Calculate the homogeneity statistic $\underline{H}$ from the $\underline{k}$ generated effect sizes for each of the 2000 replications. Computations for steps A, and B were done for each replication.

C. Compute proportions of $\underline{H}$ values (from the 2000 replications) that fall beyond central χ^2 critical values at fifteen significance levels (α).

D. Compare proportions of significant $\underline{H}$ statistics at 15 α levels from step C to the

probabilities based on the approximate noncentral chi-squared distribution at each significance (α) level.

Case II. σ^2_δ = 0.01(0.02)0.09, 0.10, for random-effects models:

A. Generate 2000 replications of δ_is ($\underline{i}$ = 1, ..., $\underline{k}$) from normal deviates and given sets of (μ_δ, σ^2_δ) values.

B. Calculate the noncentrality parameter λ. from each vector of δ_is. Randomly select a value of $\underline{H}_+$ from the noncentral chi-squared distribution based on λ.. As in Case I, computations for steps A, and B were done for each replication.

C. Compute proportions of $\underline{H}_+$ values (from the 2000 replications) beyond central chi-squared critical values (χ^2_α).

D. Compare proportions of significant $\underline{H}_+$ statistics at various significance levels (α) to the probabilities based on the calculated power values from formula (29) in Chapter III at page 24.

Attention is drawn below to the difference between simulated and theoretical power in cases involving extreme values, especially small values of δ_is, $\underline{k}$s, and $\underline{n}$s. The strength and nature of the relationships between power and

the simulation parameters are examined.

Computation for Simulated Distributions

Simulations were conducted using FORTRAN programs and the resulting data were analyzed through the SPSS-X and SAS statistical packages. FORTRAN programs were written by the author. The accuracy of the programs and subroutines was assumed by inspection of initial detailed printouts of results on individual iterations. For small numbers of iterations, results of the simulation were listed and checked by hand calculation.

Fixed-effects Models

Sample effect-sizes were obtained from noncentral t statistics, computed using normal deviates and chi-squared random numbers generated by IMSL subroutines DRNNOR and RNCHI. Note that d is exactly a noncentral t statistic even though its asymptotically normal. Glass's estimator of the effect size has a t distribution. The formula used for the unbiased effect size estimator was $d_i = \{1 - [3/(4(n_i^E+n_i^C) - 9)]\} * t_i$, where $t_i = \{\delta_i + [(\sqrt{(n_i^E + n_i^C)/n_i^E*n_i^C}) * Z_i]\}/(\sqrt{C_i/df})$, Z_i is a normal deviate, and C_i is a chi-squared random value. H statistics were calculated from those effect sizes using FORTRAN programs. For each given set of population effect sizes (δ_is) and a combination of other simulation parameters, 2000 replications of H statistics formed a simulated distribution. Upper tail

probability values from the simulated distributions were compared with upper tail probabilities of noncentral chi-squared distributions (provided by IMSL subroutine CSNDF) at 15 percentile points. Power values were calculated as the proportions of $\underline{H}$ statistics exceeding critical values at the 15 significance levels.

Random-effects Models

In random-effects models, population effect sizes (δ_is) were not fixed values; rather, they were assumed to vary randomly around one grand mean μ_δ. In the simulation, sets of population effect sizes δ_is were generated from normal distributions through IMSL subroutine DRNNOR with a given mean, μ_δ, and variance σ^2_δ. From one set of means and variances, 2000 replications of δ_is were generated. For each set of δ_is, a noncentrality parameter $\lambda.$ was calculated to obtain probability values from a noncentral chi-squared distribution using IMSL subroutine CSNDF. A homogeneity test statistic ($\underline{H}_+$) was drawn randomly from each noncentral chi-squared distribution to form a set of 2000 $\underline{H}_+$s. I did not generate $\underline{d}_i$, ..., $\underline{d}_k$ to calculate $\underline{H}_+$ because results of $\underline{H}$ from the fixed-effects models showed that noncentral $\chi^2(\lambda.)$ based on the asymptotic theory approximates well for the distributions of $\underline{H}$ for large sample sizes. Simulated power values were calculated as the proportions of $\underline{H}_+$ values exceeding various percentile points from the central chi-squared distribution (null distribution) through subroutine

CHIIN. Simulated power values were compared with these obtained from an average of 2000 noncentral chi-squared probabilities.

Test for the goodness of fit was used to examine the accuracy of the theoretical distributions to the simulated distributions. Patterns of power of homogeneity test were studied. Power values were tabulated.

Test for Goodness of Fit

A slight modification of the Kolmogorov-Smirnov one-sample test (Massey, 1956) was used to test the goodness of fit between the asymptotic distribution and the simulated distribution of $\underline{H}$. The Kolmogorov-Smirnov test focuses on the largest of the deviations between two distributions one of which is an empirical distribution based on $\underline{R}$ observations. The maximum deviation, denoted as $\underline{D}$:

$$\underline{D} = \text{maximum } |\underline{F}_0(\underline{X}) - \underline{S}_{\underline{R}}(\underline{X})| \qquad (33)$$

where

$\underline{F}_0$ = the theoretical cumulative distribution,

$\underline{F}_0(\underline{X})$ = the proportion of values equal to or less than $\underline{X}$, and

$\underline{S}_{\underline{R}}(\underline{X})$ = the observed cumulative step-function of $\underline{R}$ observations, $\underline{r}/\underline{R}$, where $\underline{r}$ is the number of observations equal to or less than $\underline{X}$.

An approximate critical value for $\underline{D}$ at the 0.05 level is $1.36/\sqrt{\underline{R}}$ if $\underline{R} > 35$ (Massey, 1956).

For each combination of various values of N, k, and the pattern of effect-sizes of the simulated distribution, fifteen proportions (simulated power values) were compared to fifteen noncentral chi-squared tail areas. Thus the empirical power function could be considered to have been observed on $R = 15$ occasions. Since the 15 measured proportions slightly differed from the $S_R(X)$ in the formula for D, the statistic can be called D^*. When $R = 15$, the Kolmogorov-Smirnov critical value for goodness of fit is 0.338 at $\alpha = 0.05$ (Massey, 1956). The critical value of 0.338 was lenient, and no significant differences were found for $R = 15$. However, since there were 2000 H statistics and sets of probability values ($R = 2000$), the critical value for D^* to reject the goodness of fit was revised to 0.030. Though only 15 differences (out of a possible of 2000 based on all available probabilities) were observed, the use of $R = 2000$ should provide a more conservative measure of differences between the two functions than the critical value for $R = 15$.

Results

Power Discrepancies for Fixed-effects Models

For fixed-effects models the simulated power values generally tended to be greater than theoretical power values. The averages of differences between the theoretical and simulated power values at $\alpha = 0.05$ for each k and N were

computed. Table 2 shows the results of paired *t* tests on the difference (theoretical power - simulated power) for each total sample size (*N*) and number of effect sizes (*k*). These tests of the mean differences gave general information about the two power values for each sample group within *k*. Both values of *t* and the mean differences indicated that the discrepancy between theoretical and simulated power values increased as *k* increased or *N* decreased.

Table 2

Paired *t* Test between Theoretical and Simulated Power for Fixed-effects Model ($\alpha = 0.05$)

k	*N*	Mean Diff.*	Sd	Se	Paired *t*	*df*	*p*
2	20*k*	0.0001	0.008	0.002	0.06	23	0.950
	60*k*	-0.0004	0.008	0.002	-0.24	23	0.816
	120*k*	-0.0010	0.010	0.002	-0.48	23	0.632
	200*k*	0.0013	0.007	0.001	0.94	23	0.359
5	20*k*	-0.0163	0.008	0.001	-14.57	47	0.000*
	60*k*	-0.0062	0.010	0.002	-4.13	47	0.000*
	120*k*	-0.0040	0.008	0.001	-3.67	47	0.001*
	200*k*	-0.0011	0.008	0.001	-0.88	47	0.382
10	20*k*	-0.0277	0.010	0.001	-26.04	87	0.000*
	60*k*	-0.0091	0.010	0.001	-8.39	87	0.000*
	120*k*	-0.0043	0.008	0.001	-4.89	87	0.000*
	200*k*	-0.0013	0.008	0.001	-1.49	87	0.141
30	20*k*	-0.0592	0.021	0.002	-26.90	87	0.000*
	60*k*	-0.0139	0.015	0.002	-8.82	87	0.000*
	120*k*	-0.0060	0.011	0.001	-4.94	87	0.000*
	200*k*	-0.0035	0.009	0.001	-3.82	87	0.000*

Note: * $p \leq 0.001$, positive mean difference indicates theoretical power > simulated power.

Data was further examined using the modified Kolmogorov-Smirnov test to detect significant discrepancies between theoretical and simulated power functions. The criterion for a "significant discrepancy" is 0.030, derived from formula (33). Again, significant discrepancies increased as the number of effect sizes (k) increased. A frequency table of the significant discrepancies crosstabulated by k is in Table 3, where the difference D stands for theoretical power values minus simulated power values.

Table 3

Crosstabulation of Discrepancies by k

	Number of effect sizes (k)				
Discrepancy	2	5	10	30	Total
D < -0.030	0 0%	19 10%	81 23%	125 36%	225 23%
-0.030 ≤ D ≤ 0.030	94	171	268	227	760
D > 0.030	2	2	3	0	7 1%
Total	96	192	352	352	992

$\chi^2 = 82.9909$ (df = 6, p < 0.00001)

Since only 7 of 992 (less than 0.7%) distributions had higher theoretical power values, the following analyses will ignore the sign and focus on the frequency of the significant discrepancies. More detailed information on

differences between simulated and theoretical power values is summarized below according to the following factors: total sample size (N), number of effect-sizes (k), sampling fractions (π_i), sample ratios (ϕ_i), patterns of effect-sizes (four patterns of fixed effect-size parameters), variation in effect-sizes.

Number of effect sizes (k). The chi-squared test for independence between "number of effect-sizes k (2, 5, 10, 30)" and the "significant discrepancy (yes or no)" was significant (69.8485, df = 3, p < .00001). Data in Table 4 indicated that discrepancies occurred the most for k = 30, and the least (or almost never) for k = 2. However, as shown in Tables 63 to 66 in the Appendix B, the values of the effect-size parameters differ for different k values.

Table 4

Crosstabulation of Significant Discrepancies by k

Significant Discrepancy	Number of Effect-sizes (k)				
	k = 2	k = 5	k = 10	k = 30	Total
Yes	2 0%	21 11%	84 24%	125 36%	232 23%
No	94	171	268	227	760
Total	96	192	352	352	992

χ^2 = 69.8485 (df = 3, p < 0.00001)

For k = 2, possible conditions were the null case (all δs were zeros) and one extreme value case. For k = 5, one additional condition showed three equal subsets of parameter effects. Only k = 10, and k = 30 contained all possible conditions: the null case, the one-extreme-value case, the two-extreme-values case, three equal subsets of parameter effects, and five equal subsets of effects. Comparisons of results for different k values overlook other important factors such as pattern of δ_is. Further analysis for each k value was necessary and is described below.

Sample sizes (N). Discrepancies between simulated and asymptotic distributions happened more often for small sample sizes (N) with larger k values. The chi-squared value to test for the dependence between "total sample size N (with values $20k$, $60k$, $120k$, $200k$)" and "significant discrepancy (yes or no)" is 260.7375 (df = 3, p < 0.00001). Data in Table 5 indicated that the discrepancies occurred the most for the smallest N and the least for the largest N. In other words, when total sample sizes were small, especially for $N = 20k$, simulated distributions showed higher power values than theoretical distributions. The asymptotic power fitted much better with effect size calculated from samples of 120 (60 in each experimental or control group) or greater.

For each value of k, the discrepancies between the simulated and asymptotic distributions were consistently

smaller for larger Ns. For k = 2, simulated and theoretical distributions fitted well. Only 2 out of 96 combinations had significant discrepancies and they are not mentioned further. Chi-square tests for the independence of "total sample sizes" and a "significant discrepancies" within each k were as follows: for k = 5, χ^2 = 0.16 (df = 3, p = 0.984); for k = 10, χ^2 = 100.95 (df = 3, p < 0.00001); and for k = 30, χ^2= 223.43 (df = 3, p < 0.00001). Complete information is listed in Table 6.

Table 5

Crosstabulation of Significant Discrepancies by Sample Size

Significant Discrepancy	Total Sample Size 20k	60k	120k	200k	Total
Yes	149 60%	46 18%	23 9%	14 7%	232 23%
No	99	202	225	234	760 77%
Total	248	248	248	248	992

χ^2 = 260.7375 (df = 3, p < 0.00001)

These results suggested that simulated distributions with large sample sizes (N) fitted better with the calculated noncentral chi-squared distributions which demonstrated the concept of the "asymptotic" distributions (for large samples). Discrepancies occurred more with small

samples. Results for each k showed that the differences among sample sizes were stronger as k increased. When k increased, small total sample sizes N were composed of more small (within-study) samples.

Table 6

Crosstabulation of Significant Discrepancies by N and k

Significant Discrepancy	Total Sample Size N= 20k	N= 60k	N=120k	N=200k	Total
k = 5					
Yes	5	6	5	5	21
	10%	13%	10%	10%	11%
No	43	42	43	43	171
χ^2_3 = 0.16 (p = 0.984)					192
k = 10					
Yes	55	16	9	4	84
	63%	18%	10%	5%	24%
No	33	72	79	84	268
χ^2_3 = 100.95 (p < 0.00001)					352
k = 30					
Yes	88	24	8	5	125
	100%	27%	9%	6%	36%
No	0	64	80	83	277
χ^2_3 = 223.43 (p < 0.00001)					352

Sampling fractions (π_i). Discrepancies between simulated and calculated power values did not depend on the "pattern of sample sizes" designated by sampling fractions

(π_i). The sets of sampling fractions included were either balanced or unbalanced. When sample sizes were the same for all effect sizes, sample sizes were considered balanced. Unbalanced sample sizes were designed according to the sampling fractions obtained from the empirical study discussed in the beginning of Chapter IV and listed in Table 62 in Appendix B.

Discrepancies between simulated and theoretical power values did not depend on sampling fractions. The test of independence chi-squared value between "significant discrepancy", and "sampling fraction" was 3.80 (df = 1, p = 0.051). Frequencies of discrepancies are listed in Table 7.

Table 7

Crosstabulation of Significant Discrepancies by π_i

Significant Discrepancy	Sampling Fraction π_i		
	Balanced	Unbalanced	Total
Yes	129 (26%)	103 (21%)	232 (23%)
No	367	393	760 (77%)
Total	496	496	992

$\chi^2_1 = 3.80$ (p = 0.051)

However, as noted in the description of the unbalanced sample sizes pattern, large effects were only accompanied with large samples. Results were not completely independent

(as also indicated by the observed significant level of 0.05 for the chi-square test); simulated values for unbalanced samples across studies tended to be higher than the theoretical values. Detailed information for each value of k is listed in Table 8.

Table 8

Crosstabulation of Significant Discrepancies by π_i and k

	Significant Discrepancy	Sampling Fraction π_i Balanced	Unbalanced	Total
$k = 5$				
	Yes	11 (12%)	10 (10%)	21 (11%)
	No	85	86	171
	$\chi^2_1 = 0.054$	($p = 0.817$)		192
$k = 10$				
	Yes	50 (28%)	34 (19%)	84 (24%)
	No	126	142	268
	$\chi^2_1 = 4.003$	($p = 0.045$)		352
$k = 30$				
	Yes	68 (39%)	57 (32%)	125 (36%)
	No	108	119	227
	$\chi^2_1 = 1.501$	($p = 0.22$)		352

Sample ratios (ϕ_i). Discrepancies between theoretical and simulated power did not depend on the ratios ϕ_i of n^E to the total sample size within a study. The chi-squared value for "significant discrepancy" and "sample ratio (0.5 or

0.35)" was 0.563 (df = 1, p = 0.453). Results are listed in Table 9. This result was consistent within each k value. Proportions of the significant discrepancies for each k are listed in Table 10.

Table 9

Crosstabulation of Significant Discrepancies by ϕ_i

Significant Discrepancy	Sample Ratio ϕ_i $n_i^E/n_i = 0.5$	$n_i^E/n_i = 0.35$	Total
Yes	121 (24%)	111 (22%)	232 (23%)
No	375	385	760 (77%)
Total	496	496	992

$\chi^2_1 = 0.56$ (p = 0.453)

Table 10

Crosstabulation of Significant Discrepancies by ϕ_i and k

Significant Discrepancy	k	Sample Ratio ϕ_i $n_i^E/n_i = 0.5$	$n_i^E/n_i = 0.35$	Total
Yes	5	13 (14%)	8 (8%)	21 (11%)
	10	41 (23%)	43 (24%)	84 (24%)
	30	65 (37%)	60 (34%)	125 (36%)

Patterns of effect-size parameters. In the simulation, the non-null effect-size parameters were designed with four patterns: (1) one distinct value with other values being zero, (2) two distinct values with others being zero, (3)

three subsets with values equal within each subset but different across subsets, and one subset contained zeros, and (4) five subsets with values equal within but different across subsets, and one subset contained zeros. Significant discrepancies between simulated and theoretical values depended on the pattern of effect sizes.

The chi-square test for the independence of "significant discrepancy (yes or no)" and "pattern of effect-sizes" was 24.03 (df = 4, p < .0001). As listed in Table 11, discrepancies occurred more when population effects had one or two extreme values. Simulated values were higher than theoretical power values when one or two extreme parameter values existed.

Table 11

Crosstabulation of Significant Discrepancies by Pattern of δ_is

	Pattern of Effect-Size Parameters					
Significant Discrepancy	Zero Effects	One Extreme	Two Extremes	Three Subsets	Five Subsets	Total
Yes	8	86	54	47	37	232
	13%	27%	34%	16%	23%	23%
No	56	234	106	241	123	760
Total	64	320	160	288	160	992

$\chi^2_4 = 24.031$ $(p < 0.0001)$

As was true in the context of other factors, when total sample size N increased, the pattern of effect-sizes was

less relevant in introducing discrepancies. However, significant discrepancies still occurred more when extreme population effects existed than when effects sizes were more evenly distributed even with large sample sizes. When the number of effects k increased, the discrepancies between sample sizes or patterns of effect sizes also increased. Results for each k value are listed in Table 12. Detailed information on power discrepancies and pattern of effect-sizes for each N by k combination is listed in Table 13.

Table 12

Crosstabulation of Significant Discrepancies by Pattern of δ_is and k

Significant Discrepancy	Zero Effects	One Extreme	Two Extremes	Three Subsets	Five Subsets	Total
k = 5						
Yes	0	17	-	4	-	21
		21%		4%		11%
No	16	63	-	92	-	171
χ^2_3 = 15.217		(p < 0.001)				192
k = 10						
Yes	3	28	21	16	16	84
	19%	35%	26%	17%	20%	24%
No	13	52	59	80	64	268
χ^2_4 = 9.336		(p = 0.053)				352
k = 30						
Yes	5	39	33	27	21	125
	31%	49%	41%	28%	26%	36%
No	11	41	47	69	59	227
χ^2_4 = 12.683		(p < 0.013)				352

Table 13

Crosstabulation of Significant Discrepancies by Pattern of δ_is, N, and k

	Pattern of Effect-size Parameters						
k*n	Zero Effects	One Extreme	Two Extremes	Three Subsets	Five Subsets	%	Total Count
5(20)	0	15%	-	8%	-	10%	(5)
5(60)	0	20%	-	8%	-	13%	(6)
5(120)	0	25%	-	0	-	10%	(5)
5(200)	0	25%	-	0	-	10%	(5)
						21/192 =	11%
10(20)	50%	65%	65%	58%	65%	63%	(55)
10(60)	25%	35%	25%	4%	10%	18%	(16)
10(120)	0	25%	15%	0	5%	10%	(9)
10(200)	0	15%	0	4%	0	5%	(4)
						84/352 =	24%
30(20)	100%	100%	100%	100%	100%	100%	(88)
30(60)	25%	55%	40%	13%	5%	27%	(24)
30(120)	0	20%	20%	0	0	9%	(8)
30(200)	0	20%	5%	0	0	6%	(5)
						125/352 =	36%
Total						232/992 =	23%

When there were many studies with small sample sizes, discrepancies between the asymptotic and the simulated distributions increased. As described above, discrepancies occurred most often when the set of parameters had one extreme value. In fact, that when k = 30 and n = 10, almost half of the measured percentile points of each simulated

distribution were significantly higher than those of the theoretical distribution (these values are not tabled). Simulation data repeatedly indicated that when effect-sizes were from many studies (e.g., $\underline{k}$ = 30) all having small sample sizes (e.g., $\underline{n}$ = 20), the homogeneity test produced greater simulated power values than the asymptotic theory. The discrepancies between the asymptotic and simulated distributions became insignificant as sample sizes increased.

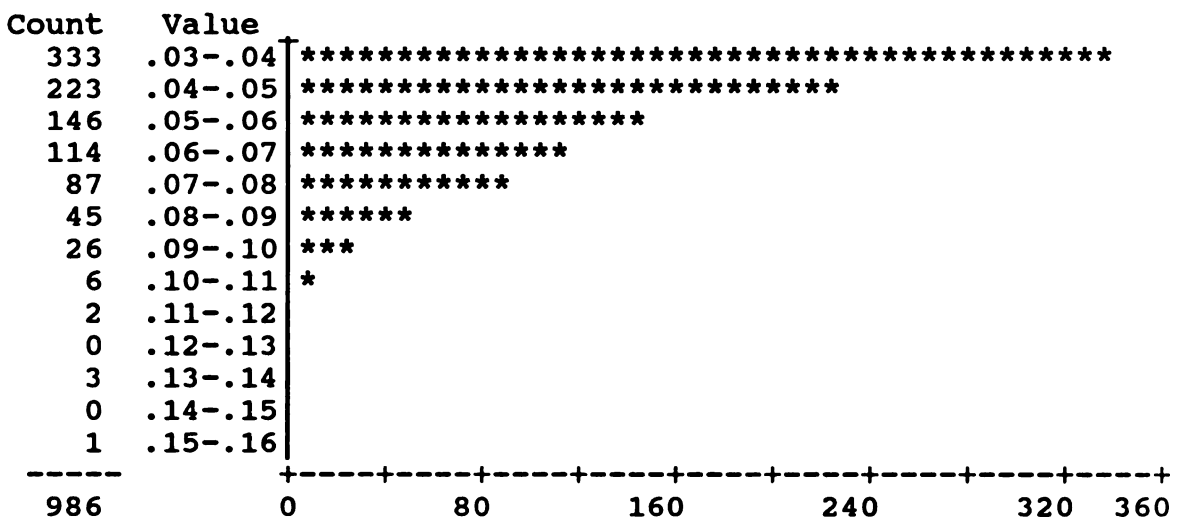
Further analyses of power discrepancies examined the magnitudes of the discrepancies. Of the 14880 measures (992 combinations × 15 percentiles) 986 had significant discrepancies: 978 were negative, where theoretical values were lower than simulated values; and 8 theoretical values were higher than the simulated values. The frequency distribution of the 986 significant differences (theoretical values - simulated values) was negatively skewed in a range from -0.15 to 0.04 with a mean of -0.051, a mode of -0.035 (333 cases, or 33.8% showed this modal discrepancy), and a standard deviation of 0.02. Figure 4.1.0 is a frequency table showing the absolute values of these discrepancies.

A paired $\underline{t}$ test showed that overall theoretical values were lower than simulated power by an average of -0.008 ($\underline{t}$ = -46.40, $\underline{p} < 0.0001$, for 14,880 records). For the 986 absolute values of significant discrepancies, about one third (34%) ranged from 0.03 to 0.04, more than one half

(56%) had values less than 0.05, and almost all (99%) had values less than 0.10.

Figure 4.1.0

Frequencies of Absolute Significant Discrepancies



As discussed above, discrepancies occurred the most often for large $\underline{k}$, small $\underline{n}$, and extreme parameter effect sizes. The magnitudes of the discrepancies also appeared to be greater for these described conditions. Mean discrepancies for pattern of population effects, number of effects $\underline{k}$, and sample sizes are listed in Table 14. The mean significant discrepancy for $\underline{k}$ = 30 and $\underline{n}$ = 20 was around 0.058 (for 594 records).

Table 14

Means of Significant Discrepancies by Pattern of δ_is, N, and k

k*n_i	Pattern of Effect-size Parameters: Zero Effects	One Extreme	Two Extremes	Three Subsets	Five Subsets	Total
2(20)	-	.031(1)	-	-	-	.031(1)
2(60)	-	- (0)	-	-	-	- (0)
2(120)	-	.031(1)	-	-	-	.031(1)
2(200)	-	- (0)	-	-	-	- (0)
5(20)	-	.037(6)	-	.037(5)	-	.037(11)
5(60)	-	.044(15)	-	.033(2)	-	.035(17)
5(120)	-	.042(6)	-	- (0)	-	.042(6)
5(200)	-	.036(8)	-	- (0)	-	.036(8)
10(20)	.038(8)	.040(47)	.036(34)	.036(33)	.038(38)	.038(160)
10(60)	.031(1)	.044(24)	.043(15)	.036(1)	.038(3)	.043(44)
10(120)	- (0)	.046(10)	.041(4)	- (0)	.033(1)	.040(15)
10(200)	- (0)	.035(8)	- (0)	.033(1)	- (0)	.028(9)
30(20)	.052(28)	.058(134)	.058(140)	.058(151)	.057(141)	.057(594)
30(60)	.036(2)	.048(43)	.054(28)	.033(4)	.038(1)	.049(78)
30(120)	- (0)	.063(15)	.049(10)	- (0)	- (0)	.058(25)
30(200)	- (0)	.050(15)	.033(2)	- (0)	- (0)	.048(17)

* Underlining indicates average theoretical power was higher. Numbers in parentheses are counts pf dofferemces.

Summary. Simulated distributions tended to have fatter upper tails than noncentral chi-squared distributions. Simulated distributions fitted quite well to noncentral chi-squared distributions when studies had large sample sizes or evenly distributed effects. Discrepancies occurred the most often and were largest when a review included many studies (large k) with small sample sizes, or when studies had extreme parameter effects.

In other words, homogeneity tests were more sensitive

than indicated by theory for data with small sample sizes or with extreme parameter effects. The non-central chi-squared distributions based on the asymptotic theory were useful for data with large samples and evenly distributed parameter effects. 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.

In his paper, Bangert-Drowns (1986) questioned the use of the homogeneity test due to the lack of understanding of the behavior of statistics for small or nonnormal samples. Simulation data indicated that simulated α values for homogeneous population effects approximately equaled the preset significance levels. Only for large collections of small samples was the size of the test significantly greater than 0.05. As shown in Table 11, when $\underline{k}$ = 30 and average within-study sample size $\underline{n}$ = 20, simulated sizes and power values were consistently higher than theoretical values. Also simulated sizes were around 0.10 (0.05 higher than the nominal level) for $\underline{n}$ = 20 and $\underline{k}$ = 30 (Table 11). Under the null hypothesis, these higher values indicate an inflated rate of false rejections (type I error).

When effects were not homogeneous (i.e., under alternative hypotheses), higher simulated power for small samples and extreme parameter effects was not problematic. In these cases (1) heterogeneity should be detected (since H_0 is false), and (2) simulated power values were not much

higher than the asymptotic power values. Asymptotic power underestimated the power of the homogeneity test for extreme parameter effects and small samples.

Power Discrepancies in Random-effects Models

Results (patterns of discrepancies) were similar across different population effect-size means, μ_δ, of 0, 0.1, 0.25, or 0.5. Table 15 demonstrates the results of paired *t* tests for each *k*, showing the differences in theoretical and simulated power values.

Table 15

Paired *t* Test between Theoretical and Simulated Power for Random-effects Model

k	μ_δ	Mean Diff.*	Sd	Se	Paired *t*	*df*	*p*
2	0.00	0.0003	0.008	0.000	1.64	2399	0.102
	0.10	0.0002	0.008	0.000	1.07	2399	0.287
	0.25	0.0000	0.008	0.000	0.09	2399	0.931
	0.50	-0.0000	0.008	0.000	-0.06	2399	0.950
5	0.00	-0.0002	0.007	0.000	-1.14	2399	0.253
	0.10	-0.0004	0.007	0.000	-2.53	2399	0.012*
	0.25	0.0001	0.007	0.000	0.83	2399	0.405
	0.50	0.0003	0.008	0.000	1.90	2399	0.058
10	0.00	-0.0000	0.006	0.000	-0.08	2399	0.933
	0.10	0.0002	0.007	0.000	1.37	2399	0.172
	0.25	-0.0001	0.007	0.000	-0.58	2399	0.559
	0.50	-0.0006	0.006	0.000	-4.75	2399	0.000#
30	0.00	-0.0003	0.007	0.000	-1.90	1919	0.057
	0.10	0.0006	0.007	0.000	3.76	1919	0.000#
	0.25	0.0002	0.007	0.000	1.28	1919	0.202
	0.50	-0.0003	0.007	0.000	-2.00	1919	0.046*

Note: * $p < 0.05$, # $p < 0.001$, positive mean difference indicates theoretical power > simulated power.

For k = 2, none of the paired t tests showed significance at the 0.05 level. For k = 5 and 10, one average difference was significant. For k = 30, one group was found significant and another barely significant. The mean differences were very small. Statistical significance was largely due to the large degrees of freedom and small standard error values. These average differences would not be consequential for our interpretation of theoretical power values.

The modified Kolmogorov-Smirnov one-sample test was again used to determine the goodness of fit between the asymptotic power functions and the simulated power functions. As in the fixed-effects case, the number of replications used in random-effects was 2000. Thus the maximum deviation, D, from formula (30) was again 0.030.

Only 20 of 2688 (0.07%) distributions had significant discrepancies. For k = 2, 9 of 640 (1.4%) distributions had significant discrepancies. For k = 5, 6 of 640 (1%) distributions had significant discrepancies. For k = 10, 2 of 640 (0.3%) combinations had significant discrepancies. And for k = 30, 4 of 512 (0.8%) combinations had significant discrepancies. Frequencies of power discrepancies are listed in Table 16.

Significant discrepancies occurred less than 1 out of 100 times. Their occurrence was dependent upon k, the number of effect-sizes (test of association, $\chi^2_3 = 14.898$, p

< 0.005). Significant discrepancies did not depend on sample sizes N ($\chi^2_3 = 4.861$, $p < 0.25$). No significant association ($\chi^2_9 = 10.18$, $p < 0.40$) was found between the number of effect-sizes (k) and sample sizes (N) and occurrence of significant power discrepancies. In other words, the dependence of power discrepancies on sample sizes N did not vary with k. Significant discrepancies occurred the most often for k = 2; however, the occurrence rate was still less than 1.5%.

Table 16

Frequency Table for Significant Discrepancies for Random-effects Model

k	N	μ_δ				Total for N	Total for k
		0.00	0.10	0.25	0.50		
2	1	-	-	2	1	3	
	2	-	-	2	-	2	
	3	1	1	1	-	3	
	4	-	-	1	-	1	9
5	1	1	-	1	1	3	
	2	-	-	-	-	0	
	3	-	-	-	1	1	
	4	-	1	1	-	2	6
10	1	-	1	-	-	1	
	2	-	-	-	-	0	
	3	-	-	-	-	0	
	4	-	1	-	-	1	2
30	1	-	1	-	-	1	
	2	-	1	-	-	1	
	3	-	-	-	-	0	
	4	-	1	-	-	1	3
	Total	2	7	8	3	20	20

The magnitude of significant power discrepancies for random-effects models was examined. Significant discrepancies occurred for 28 out of 36,480 (0.8%) measures. 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. The mean of the 28 significant values was 0.009 ($\underline{t}$ = 1.48, $\underline{p}$ > .05). The $\underline{t}$ statistic indicated that the mean did not differ significantly from zero. In other words, theoretical power values were not consistently either higher or lower on average than simulated power values.

The occurrence rates as well as the magnitudes of significant discrepancies differed for random- and fixed-effects models. The dissimilarity may have resulted partially from the fact that population effect-sizes in the random-effects models were all normally distributed, unlike the cases examined for fixed-effects models. Also, the $\underline{H}_+$ statistics were generated from asymptotic noncentral chi-squared distributions in the random-effects models. Results from the fixed-effects case had indicated that the theoretical power functions approximated well the simulated functions when sample size was large. However, simulated power values in the random-effects simulations may still be underestimating the true power for small samples and large $\underline{k}$ values. Simulation data did not indicate a many differences

between simulated and theoretical power values; therefore, the analysis of power for random-effects models will focus on the theoretical values.

Power Analysis

Power values at $\alpha = 0.05$ were selected for analysis. Factors for the power analysis included: the number of effect sizes k (2, 5, 10, 30), total sample sizes N ($20k$, $60k$, $120k$, $200k$), sampling fractions π_i (balanced vs. unbalanced sample sizes between studies), sample ratios ϕ_i (balanced vs. unbalanced sample sizes within studies), and patterns of effect size parameters. Relations between power and these factors were studied through analysis of variance, regression, correlation and curve fitting.

Fixed-effects model. Power values for the homogeneity test were positively related to the variance of simulated effects, sample sizes N, and number of effects k. However, since these variables were not directly (or linearly) related to power, correlation coefficients representing the relationships appeared weak. For the fixed-effects model, the correlation coefficient r(power, k) was 0.15 (p = 0.001), and r(power, $\sqrt{k}$) was 0.16. Between power and total sample size N, the correlation coefficient r(power, N) was 0.38 ($p < 0.001$), and r(power, $\sqrt{N}$) was 0.43. The relationship between power and the spread among population effects was greater, r(power, S^2_δ) was 0.47 ($p < 0.001$), and

$\underline{r}$(power, $\underline{S}_\delta$) was 0.64[1]. The relations between sampling fraction or sample ratio and power were not significant. The correlation coefficient $\underline{r}$(power, π) was 0.05 ($\underline{p}$ = 0.14) and $\underline{r}$(power, ϕ) was -0.02 ($\underline{p}$ = 0.44).

A regression analysis of the power values used a stepwise procedure. The particular stepwise procedure selected predictor variables in the order of the amount of the variation (change in $\underline{R}^2$) in power values being explained by the predictor. The variable representing the pattern of δ_is was not continuous thus was not entered as a predictor variable. As expected, the weighted average of parameter effects, $\delta_\cdot$ (as in formula (9) in Chapter III, page 17), and the spread of δ_is, $\underline{S}_\delta$, increased linearly within each pattern of δ_is. The combination of $\delta_\cdot$ and $\underline{S}_\delta$ contained information about the pattern of δ_is. Therefore, $\delta_\cdot$ and $\underline{S}_\delta$ were entered into the regression as predictor variables instead. The association between the pattern of δ_is and power was also studied below via analysis of variance.

The predictor variable first selected in the regression model was the index of spread among parameter effects $\underline{S}_\delta$ (multiple $\underline{R}$ = 0.64, $\underline{R}^2$ = 0.41, $\underline{F}_{1,990}$ = 678.49, $\underline{p}$ < 0.0001, for 992 cases). Total sample size with square root $\sqrt{\underline{N}}$ was next to be included in the model with $\underline{R}$ increased to 0.83, $\underline{R}^2$ = 0.69, $\underline{R}^2$ change = 0.28, and $\underline{F}_{2,989}$ = 1105.84 ($\underline{p}$ <

---

[1] In the fixed-effects case, S^2_δ represents the distance between fixed δ_i values.

0.0001). The third predictor included in the regression model was the square root of $\underline{k}$, $\sqrt{\underline{k}}$ ($\underline{R}$ = 0.84, $\underline{R}^2$ = 0.70, $\underline{R}^2$ change = 0.01, $\underline{F}_{3,988}$ = 765.44, $\underline{p}$ < 0.0001). Sampling fraction between studies (π: 1 = balanced, 2 = unbalanced) had a very small effect, however, was also selected into the model last ($\underline{R}$ = 0.84, $\underline{R}^2$ = 0.70, $\underline{R}^2$ change = .002). The final regression model for combination of parameters j for the fixed-effects model is listed below:

$$\underline{\text{Power}}_j = -0.252 + 1.839\ (\underline{S}_\delta)_j + 0.012\ \sqrt{\underline{N}_j} + (-0.032)\ \sqrt{\underline{k}_j} + 0.035\ \pi_j. \qquad (34)$$

As predicted the spread in δ_is explained much variation in power. Total sample size was also important. Number of effects $\underline{k}$ had a smaller effect, since $\underline{N} = \underline{k}*\underline{n}$ had already partially taken into account the effect of $\underline{k}$.

Analysis of variance was also conducted for power with number of effects, sample sizes, sampling fraction, sample ratio, and pattern of parameters as factors for the fixed-effects model. Results are listed in Table 17.

The power of $\underline{H}$ was explained most by sample size and the pattern of δ_is. Sampling fraction and sample ratio were again not influential on the power of $\underline{H}$. This result seems reasonable since effect sizes for homogeneity test were weighted by their precision which is nearly proportional to the sample sizes (see formula (5), and (9) in Chapter III at page 15, and 16). And the power of homogeneity test should

depend on whether effect sizes were similar. Changes in sample sizes combined with values of δ_is should affect the power of the homogeneity test. Thus the total sample sizes increased differences among the effects sizes were also emphasized. However, the regression model seemed better than the ANOVA model in explaining the variation of power. The amount explained by the ANOVA model was around 31% which was much less than the amount explained by the regression model (70%).

Table 17

Analysis of Variance for Power of H

Source of Variation	Sum of Squares	Amt. Exp.	df	Mean Squares	F	p
Main Effect	39.031	(31%)	11	3.548	38.255	.000
k	1.642	(1%)	3	.547	5.900	.001
Sample size	33.706	(27%)	3	11.235	121.131	.000
Sampling fraction	.316	(0%)	1	.316	3.408	.065
Sample ratio	.088	(0%)	1	.088	.944	.331
Pattern of δ_is	2.250	(2%)	3	.750	8.087	.000
Residual	84.963	(69%)	916	.093		
Total	123.994		927	.134		

Average power values at $\alpha = 0.05$ were calculated. For fixed-effects, the grand mean power was 0.44 (across 992 cases) with a standard deviation of 0.37. Too much information is aggregated in the grand mean; thus this value has little practical meaning. Further categorization of the data was necessary. Mean power values for each pattern of

δ_is and total sample size are listed in Table 18.

Table 18

Means of Theoretical Power of H by Pattern of δ_is, N, and k (α = 0.05)

k*n	Pattern of Effect-size Parameters					
	Zero Effects	One Extreme	Two Extremes	Three Subsets	Five Subsets	Total
2(20)	.050(4)	.146(20)	-	-	-	.1300(24)
2(60)	.050(4)	.315(20)	-	-	-	.2708(24)
2(120)	.050(4)	.470(20)	-	-	-	.4003(24)
2(200)	.050(4)	.571(20)	-	-	-	.4844(24)
5(20)	.050(4)	.143(20)	-	.125(24)	-	.1259(48)
5(60)	.050(4)	.337(20)	-	.301(24)	-	.2951(48)
5(120)	.050(4)	.500(20)	-	.496(24)	-	.4606(48)
5(200)	.050(4)	.596(20)	-	.641(24)	-	.5732(48)
10(20)	.050(4)	.154(20)	.189(20)	.171(24)	.158(20)	.1630(88)
10(60)	.050(4)	.360(20)	.460(20)	.432(24)	.409(20)	.3992(88)
10(120)	.050(4)	.515(20)	.595(20)	.636(24)	.606(20)	.5657(88)
10(200)	.050(4)	.607(20)	.669(20)	.765(24)	.718(20)	.6640(88)
30(20)	.050(4)	.134(20)	.196(20)	.296(24)	.256(20)	.2140(88)
30(60)	.050(4)	.315(20)	.430(20)	.624(24)	.566(20)	.4705(88)
30(120)	.050(4)	.462(20)	.573(20)	.797(24)	.723(20)	.6191(88)
30(200)	.050(4)	.564(20)	.651(20)	.870(24)	.808(20)	.6992(88)

Simulated power values were slightly higher, the grand mean was 0.45 (992 cases) with a standard deviation of 0.36. Means of simulated power values for each pattern of δ_is by total sample size are listed in Table 19.

Table 19

Means of Simulated Power of H by Pattern of δ_is, N, and k (α = 0.05)

	Pattern of Effect-size Parameters					
k*n	Zero Effects	One Extreme	Two Extremes	Three Subsets	Five Subsets	Total
2(20)	.049(4)	.146(20)	-	-	-	.1300(24)
2(60)	.053(4)	.315(20)	-	-	-	.2711(24)
2(120)	.050(4)	.471(20)	-	-	-	.4012(24)
2(200)	.045(4)	.571(20)	-	-	-	.4830(24)
5(20)	.058(4)	.159(20)	-	.142(24)	-	.1421(48)
5(60)	.054(4)	.348(20)	-	.304(24)	-	.3013(48)
5(120)	.051(4)	.504(20)	-	.501(24)	-	.4646(48)
5(200)	.055(4)	.598(20)	-	.641(24)	-	.5742(48)
10(20)	.078(4)	.184(20)	.217(20)	.196(24)	.186(20)	.1908(88)
10(60)	.058(4)	.373(20)	.470(20)	.439(24)	.415(20)	.4082(88)
10(120)	.055(4)	.521(20)	.601(20)	.638(24)	.609(20)	.5699(88)
10(200)	.054(4)	.610(20)	.671(20)	.764(24)	.718(20)	.6651(88)
30(20)	.100(4)	.194(20)	.257(20)	.354(24)	.306(20)	.2732(88)
30(60)	.062(4)	.337(20)	.449(20)	.631(24)	.576(20)	.4844(88)
30(120)	.055(4)	.477(20)	.582(20)	.796(24)	.726(20)	.6251(88)
30(200)	.051(4)	.574(20)	.651(20)	.871(24)	.811(20)	.7027(88)

When all effects were zero (homogeneous), the simulated power was higher than expected α levels, especially for small samples (e.g., n = 20). Simulated power values for α = 0.10, 0.025, and 0.01 are also listed in Table 20.

Table 20

Means of Simulated Power of H by Homogeneous δ_is, N, and k (δ = 0)

k*n	α = 0.10	0.05	0.025	0.01
2(20)	.092 (4)	.049 (4)	.026 (4)	.014 (4)
2(60)	.104 (4)	.053 (4)	.025 (4)	.009 (4)
2(120)	.100 (4)	.050 (4)	.026 (4)	.011 (4)
2(200)	.096 (4)	.045 (4)	.023 (4)	.009 (4)
5(20)	.113 (4)	.058 (4)	.033 (4)	.015 (4)
5(60)	.103 (4)	.054 (4)	.027 (4)	.011 (4)
5(120)	.101 (4)	.051 (4)	.025 (4)	.010 (4)
5(200)	.101 (4)	.055 (4)	.028 (4)	.012 (4)
10(20)	.132 (4)	.078 (4)	.047 (4)	.025 (4)
10(60)	.110 (4)	.058 (4)	.030 (4)	.013 (4)
10(120)	.103 (4)	.055 (4)	.029 (4)	.014 (4)
10(200)	.105 (4)	.054 (4)	.030 (4)	.012 (4)
30(20)	.160 (4)	.100 (4)	.066 (4)	.038 (4)
30(60)	.118 (4)	.062 (4)	.034 (4)	.017 (4)
30(120)	.105 (4)	.055 (4)	.027 (4)	.012 (4)
30(200)	.103 (4)	.051 (4)	.025 (4)	.010 (4)

Analysis of variance (ANOVA) was applied to power values for each pattern of δ_is. Within each pattern, as the mean value or the spread of the population effects increased, the power of the homogeneity test increased. Sample size was again a significant factor. Tables 21 to 28 list the ANOVA results and mean power values for each pattern of δ_is.

Table 21

ANOVA on Power of H for δ_is with One Extreme Value

Source of Variation	Sum of Squares	df	Mean Squares	F	p
Main Effect	31.780	10	3.178	307.305	.000
k	.079	3	.026	2.529	.058
N	8.874	3	2.958	285.771	.000
Magnitude of δ_is	22.828	4	5.707	551.361	.000
Two-way Interactions	4.658	33	.141	13.637	.000
k × N	.014	9	.002	.153	.998
k × δ	.044	12	.004	.353	.978
N × δ	4.600	12	.383	37.043	.000
Three-way Interactions	.051	36	.001	.137	1.00
Residual	2.484	240	.010		
Total	38.973	319	.133		

Table 22

Means of Power of H for δ_is with One Extreme Value by N and k (α = 0.05)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
2(20)	.053(4)	.066(4)	.114(4)	.195(4)	.304(4)	.1300(20)
2(60)	.058(4)	.097(4)	.246(4)	.472(4)	.702(4)	.2708(20)
2(120)	.065(4)	.148(4)	.436(4)	.762(4)	.940(4)	.4003(20)
2(200)	.075(4)	.215(4)	.640(4)	.930(4)	.995(4)	.4844(20)
5(20)	.052(4)	.063(4)	.106(4)	.187(4)	.305(4)	.1426(20)
5(60)	.056(4)	.092(4)	.247(4)	.515(4)	.775(4)	.3370(20)
5(120)	.063(4)	.141(4)	.475(4)	.843(4)	.979(4)	.5000(20)
5(200)	.071(4)	.213(4)	.720(4)	.976(4)	.999(4)	.5960(20)
10(20)	.052(4)	.063(4)	.110(4)	.203(4)	.342(4)	.1540(20)
10(60)	.056(4)	.094(4)	.276(4)	.572(4)	.801(4)	.3597(20)
10(120)	.063(4)	.149(4)	.532(4)	.856(4)	.973(4)	.5146(20)
10(200)	.072(4)	.235(4)	.759(4)	.970(4)	.999(4)	.6070(20)
30(20)	.051(4)	.060(4)	.096(4)	.172(4)	.294(4)	.1346(20)
30(60)	.055(4)	.083(4)	.236(4)	.503(4)	.698(4)	.3148(20)
30(120)	.059(4)	.127(4)	.468(4)	.749(4)	.906(4)	.4621(20)
30(200)	.066(4)	.199(4)	.663(4)	.901(4)	.990(4)	.5639(20)

Table 22.a

Means of Simulated Power of H for δ_is with One Extreme Value by N and k (α = 0.05)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
2(20)	.056(4)	.074(4)	.115(4)	.189(4)	.297(4)	.1461(20)
2(60)	.056(4)	.091(4)	.248(4)	.472(4)	.707(4)	.3147(20)
2(120)	.066(4)	.150(4)	.438(4)	.759(4)	.945(4)	.4714(20)
2(200)	.075(4)	.214(4)	.638(4)	.933(4)	.993(4)	.5706(20)
5(20)	.068(4)	.075(4)	.120(4)	.202(4)	.330(4)	.1589(20)
5(60)	.060(4)	.097(4)	.247(4)	.534(4)	.801(4)	.3477(20)
5(120)	.066(4)	.143(4)	.477(4)	.850(4)	.985(4)	.5041(20)
5(200)	.067(4)	.209(4)	.733(4)	.979(4)	1.000(4)	.5978(20)
10(20)	.078(4)	.089(4)	.138(4)	.232(4)	.383(4)	.1843(20)
10(60)	.061(4)	.103(4)	.286(4)	.588(4)	.827(4)	.3732(20)
10(120)	.065(4)	.154(4)	.540(4)	.864(4)	.983(4)	.5210(20)
10(200)	.070(4)	.235(4)	.771(4)	.976(4)	.999(4)	.6104(20)
30(20)	.111(4)	.113(4)	.152(4)	.234(4)	.366(4)	.1950(20)
30(60)	.069(4)	.091(4)	.258(4)	.533(4)	.735(4)	.3370(20)
30(120)	.060(4)	.139(4)	.486(4)	.767(4)	.934(4)	.4771(20)
30(200)	.070(4)	.211(4)	.676(4)	.919(4)	.994(4)	.5742(20)

Table 23

ANOVA on Power of H for δ_is with Two Extreme Values

Source of Variation	Sum of Squares	df	Mean Squares	F	p
Main Effect	19.994	8	2.499	236.143	.000
k	.010	1	.010	.948	.332
N	5.067	3	1.689	159.576	.000
Magnitude of δ_is	14.918	4	3.729	352.367	.000
Two-way Interactions	2.419	19	.127	12.028	.000
k × N	.008	3	.003	.238	.869
k × δ	.009	4	.002	.223	.925
N × δ	2.402	12	.200	18.910	.000
Three-way Interactions	.016	12	.001	.126	1.00
Residual	1.270	120	.011		
Total	23.699	159	.149		

Table 24

Means of Power of H for δ_is with Two Extreme Values by N and k (α = 0.05)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
10(20)	.053(4)	.069(4)	.142(4)	.290(4)	.393(4)	.1892(20)
10(60)	.059(4)	.116(4)	.397(4)	.773(4)	.954(4)	.4598(20)
10(120)	.069(4)	.203(4)	.729(4)	.977(4)	.999(4)	.5954(20)
10(200)	.083(4)	.336(4)	.928(4)	.999(4)	1.000(4)	.6690(20)
30(20)	.052(4)	.066(4)	.129(4)	.269(4)	.465(4)	.1961(20)
30(60)	.057(4)	.106(4)	.375(4)	.715(4)	.897(4)	.4301(20)
30(120)	.065(4)	.187(4)	.679(4)	.937(4)	.997(4)	.5731(20)
30(200)	.077(4)	.315(4)	.867(4)	.996(4)	1.000(4)	.6509(20)

Note: The pattern of δ_i values with two extreme values was $(0, \ldots, 0, \delta, \delta)$.

Table 24.a

Means of Simulated Power of H for δ_is with Two Extreme Values by N and k (α = 0.05)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
10(20)	.080(4)	.098(4)	.166(4)	.320(4)	.423(4)	.2175(20)
10(60)	.063(4)	.125(4)	.412(4)	.783(4)	.966(4)	.4698(20)
10(120)	.074(4)	.210(4)	.738(4)	.984(4)	1.000(4)	.6009(20)
10(200)	.084(4)	.350(4)	.924(4)	1.000(4)	1.000(4)	.6714(20)
30(20)	.105(4)	.122(4)	.188(4)	.332(4)	.536(4)	.2567(20)
30(60)	.068(4)	.119(4)	.400(4)	.733(4)	.924(4)	.4488(20)
30(120)	.072(4)	.201(4)	.688(4)	.949(4)	.998(4)	.5816(20)
30(200)	.075(4)	.307(4)	.877(4)	.996(4)	1.000(4)	.6512(20)

Note: The pattern of δ_i values with two extreme values was $(0, \ldots, 0, \delta, \delta)$.

Table 25

ANOVA on Power of H for Three Equal Subsets of δ_is

Source of Variation	Sum of Squares	df	Mean Squares	F	p
Main Effect	32.018	10	3.202	4347.081	.000
k	3.160	2	1.580	2145.030	.000
N	13.005	3	4.335	5885.632	.000
Magnitude of δs	15.853	5	3.171	4304.771	.000
Two-way Interactions	3.023	31	.098	132.419	.000
k × N	.195	6	.032	44.052	.000
k × δ	.450	10	.045	61.113	.000
N × δ	2.379	15	.159	215.303	.000
Three-way Interactions	1.221	30	.041	55.255	.000
Residual	.159	216	.001		
Total	36.421	287	.127		

Table 26

Means of Power for Three Equal Subsets of δ_is by N and k (α = 0.05)

k*n	δ = 0.10	0.20	0.25	0.30	0.40	0.50	Total
5(20)	.056	.076	.092	.112	.168	.243	.1245(24)
5(60)	.070	.138	.141	.272	.463	.666	.3010(24)
5(120)	.091	.247	.376	.524	.795	.944	.4962(24)
5(200)	.122	.401	.596	.772	.960	.997	.6414(24)
10(20)	.059	.088	.114	.147	.244	.377	.1714(24)
10(60)	.078	.190	.293	.423	.705	.902	.4319(24)
10(120)	.112	.379	.584	.774	.968	.998	.6359(24)
10(200)	.163	.621	.846	.960	.999	1.000	.6691(24)
30(20)	.064	.120	.174	.249	.463	.704	.2958(24)
30(60)	.100	.347	.559	.766	.973	.999	.6240(24)
30(120)	.169	.705	.917	.988	1.000	1.000	.7965(24)
30(200)	.285	.938	.996	1.000	1.000	1.000	.8699(24)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots, 0, \delta,\ldots, \delta, 2\delta,\ldots, 2\delta)$.

Table 26.a

Means of Simulated Power of H for Three Equal Subsets of δ_is by N and k (α = 0.05)

k*n	δ = 0.10	0.20	0.25	0.30	0.40	0.50	Total
5(20)	.069	.092	.106	.136	.184	.266	.1421(24)
5(60)	.076	.144	.199	.279	.460	.665	.3038(24)
5(120)	.092	.252	.381	.529	.801	.949	.5006(24)
5(200)	.120	.404	.595	.770	.962	.996	.6412(24)
10(20)	.080	.118	.138	.169	.269	.405	.1964(24)
10(60)	.080	.202	.307	.405	.705	.907	.4388(24)
10(120)	.121	.379	.590	.769	.970	.998	.6381(24)
10(200)	.153	.627	.843	.960	.999	1.000	.7637(24)
30(20)	.123	.189	.248	.324	.511	.728	.3541(24)
30(60)	.115	.360	.573	.767	.973	.998	.6311(24)
30(120)	.171	.701	.920	.985	1.000	1.000	.7961(24)
30(200)	.285	.938	.996	1.000	1.000	1.000	.8699(24)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots, 0, \delta,\ldots, \delta, 2\delta,\ldots, 2\delta)$.

Table 27

ANOVA on Power of H for Five Equal Subsets of δ_is

Source of Variation	Sum of Squares	df	Mean Squares	F	p
Main Effect	18.832	8	2.354	2997.240	.000
k	.437	1	.437	556.972	.000
N	7.590	3	2.530	3221.503	.000
Magnitude of δ_is	10.804	4	2.701	3439.109	.000
Two-way Interactions	1.994	19	.105	133.602	.000
k × N	.057	3	.019	23.998	.000
k × δ	.162	4	.040	51.549	.000
N × δ	1.775	12	.148	188.353	.000
Three-way Interactions	.291	12	.024	30.846	.000
Residual	.094	120	.001		
Total	21.210	159	.133		

Table 28

Means of Power of H for Five Equal Subsets of δ_is by N and k (α = 0.05)

k*n	½δ = 0.10	0.20	0.30	0.40	0.50	Total
10(20)	.057(4)	.082(4)	.129(4)	.207(4)	.317(4)	.1585(20)
10(60)	.073(4)	.164(4)	.355(4)	.614(4)	.835(4)	.4086(20)
10(120)	.101(4)	.318(4)	.686(4)	.932(4)	.994(4)	.6060(20)
10(200)	.313(4)	.532(4)	.917(4)	.997(4)	1.000(4)	.7519(20)
30(20)	.061(4)	.100(4)	.187(4)	.339(4)	.542(4)	.2458(20)
30(60)	.086(4)	.254(4)	.604(4)	.899(4)	.990(4)	.5664(20)
30(120)	.133(4)	.542(4)	.941(4)	.999(4)	1.000(4)	.7231(20)
30(200)	.211(4)	.830(4)	.998(4)	1.000(4)	1.000(4)	.8078(20)

Note: The pattern of five equal subsets of δ_i values was $(0,\ldots,0,\ \frac{1}{2}\delta,\ldots,\frac{1}{2}\delta,\ \delta,\ldots,\delta,\ 1\frac{1}{2}\delta,\ldots,1\frac{1}{2}\delta,\ 2\delta,\ldots,2\delta)$.

Table 28.a

Means of Simulated Power for δ_is with Five Equal Subsets by N by k (α = 0.05)

k*n	½δ = 0.10	0.20	0.30	0.40	0.50	Total
10(20)	.081(4)	.108(4)	.162(4)	.237(4)	.343(4)	.1863(20)
10(60)	.078(4)	.177(4)	.357(4)	.623(4)	.842(4)	.4154(20)
10(120)	.106(4)	.321(4)	.691(4)	.933(4)	.994(4)	.6090(20)
10(200)	.318(4)	.531(4)	.915(4)	.997(4)	1.000(4)	.7523(20)
30(20)	.120(4)	.167(4)	.251(4)	.400(4)	.589(4)	.3056(20)
30(60)	.105(4)	.267(4)	.613(4)	.903(4)	.989(4)	.5757(20)
30(120)	.140(4)	.547(4)	.942(4)	.999(4)	1.000(4)	.7256(20)
30(200)	.221(4)	.837(4)	.999(4)	1.000(4)	1.000(4)	.8113(20)

Note: The pattern of five equal subsets of δ_i values was $(0,\ldots,0,\ \frac{1}{2}\delta,\ldots,\frac{1}{2}\delta,\ \delta,\ldots,\delta,\ 1\frac{1}{2}\delta,\ldots,1\frac{1}{2}\delta,\ 2\delta,\ldots,2\delta)$.

The main effect of k and the two-way interaction effects of k by N, and k by δ were not significant for the one-extreme-value case or the two-extreme-values case. Power values did not vary with k when population effects had extreme values. However these effects were significant for the three-equal-subsets and the five-equal-subsets patters. Power values increased faster with large k.

Random-effects model. Correlation coefficients were also obtained for power of H_+ and number of effects, total sample sizes, variance of parameter effects, sampling fraction, and sample ratio for the random-effects model. In comparison to the fixed-effects model, the relationships between power and the first three variables were stronger for the random-effects model; r(power, k) was 0.29 ($p < 0.001$), r(power, $\sqrt{k}$) was 0.34, r(power, N) was 0.43 ($p < 0.001$), r(power, $\sqrt{N}$) was 0.53, r(power, σ^2_δ) = 0.48 ($p < 0.001$), and r(power, σ_δ) was 0.55 for random-effects. Correlations were not significant between power and the sampling fraction (r(power, π) was -0.24, $p = 0.54$), or between power and the sample ratio ϕ (r(power, ϕ) was 0.02, $p = 0.64$).

Regression analysis with a stepwise procedure was also applied to the power of H_+. For random-effects, instead of the predicted δ. (weighted average of δ_is) and S_δ (the index of spread among the fixed δ_is), the standard variation of parameter effects (σ_δ) was included in the regression

analysis. For $\mu_\delta = 0.00$, the stepwise procedure also selected the standard deviation of parameter effects σ_δ as the most important predictor for power of $\underline{H}_+$ ($\underline{R} = 0.55$, $\underline{R}^2 = 0.30$, $\underline{F}_{1,990} = 261.97$, $\underline{p} < 0.0001$). The second predictor included in the regression was the square root of the total sample size $\sqrt{\underline{N}}$ ($\underline{R} = 0.87$, $\underline{R}^2 = 0.76$, $\underline{R}^2$ change $= 0.46$, $\underline{F}_{2,989} = 943.18$, $\underline{p} < 0.0001$). Only two predictors were selected for the random-effects model, however, the variation explained by the model reached 76%. For $\mu_\delta = 0.10$, 0.25, and 0.50 results were similar to the case of $\mu_\delta = 0.00$.

For $\mu_\delta = 0.00$, the final regression model was:

$$\underline{\text{Power}}_j = (-0.326) + 1.557\ (\sigma_\delta)_j + 0.013\ \sqrt{\underline{N}}_j. \qquad (35)$$

Results indicated that the power of $\underline{H}_+$ depended upon the variation of effects σ_δ and the total sample size in the random-effects model. It appeared that $\underline{k}$ had no effect, however, since $\underline{N} = \underline{k}*\underline{n}$, the total sample size had already taken into account the effect of $\underline{k}$.

The grand mean power value for $\mu_\delta = 0.00$ was 0.41 with a standard deviation of 0.31. Mean power values for random-effects increased as the variance of population effects or the sample sizes increased. Mean power values according to the variance of parameter effects for random-effects with $\mu_\delta = 0.00$ are listed in Table 29.

Asymptotic and simulated power values were calculated

and power curves drawn for α = .05; $\underline{k}$ = 2, 5, 10, 30; and $\underline{N}$ = 20$\underline{k}$, 60$\underline{k}$, 120$\underline{k}$, 200$\underline{k}$ for fixed-effects models in Figures 4.1.1 to 4.4.2 in the Appendix D. For random-effects models, power values were calculated with μ_δ = 0, 0.10, 0.25, 0.50; and σ_δ^2 = 0.01(0.02)0.9, 0.10. See Figures 4.5.1 to 4.8.4 in the Appendix D. Power tables for other α levels are also listed in the Appendix C.

Table 29

Mean Power of $\underline{H}_+$ at α = 0.05 for μ_δ = 0 for the Random-effects Model

σ^2_δ	$\underline{N}$ = 20$\underline{k}$	60$\underline{k}$	120$\underline{k}$	200$\underline{k}$	Total
.00	0.05(16)	0.05(16)	0.05(16)	0.05(16)	0.05(64)
.00-.02	0.06(16)	0.13(20)	0.23(20)	0.35(20)	0.20(76)
.02-.04	0.09(16)	0.29(20)	0.50(20)	0.63(20)	0.39(76)
.04-.06	0.13(16)	0.42(20)	0.54(16)	0.67(16)	0.44(68)
.06-.08	0.17(16)	0.53(20)	0.51(12)	0.64(12)	0.45(60)
.08-.10	0.23(32)	0.47(28)	0.59(24)	0.71(24)	0.48(108)
.15	0.34(16)	0.52(12)	0.70(12)	0.78(12)	0.57(52)
.20	0.42(16)	0.60(12)	0.75(12)	0.81(12)	0.63(52)
.25	0.48(16)	0.65(12)	0.78(12)	0.76(8)	0.64(48)
Total	0.22(160)	0.37(160)	0.48(144)	0.56(140)	0.41(604)

CHAPTER V

THE INFLUENCE OF THE SIGNIFICANCE LEVEL AND POWER OF THE FIRST STAGE TEST ON THE SECOND STAGE TEST -- A SEQUENTIALLY RELATED TESTING PROCEDURE --

In this section, I will first distinguish among several similar terms: "sequential analysis" (Wald, 1952), "sequential decision" (Sobel & Wald, 1949), and "sequentially related testing procedure". Use of these terms in the literature suggests that "sequential analysis" defines the sampling procedure, "sequential decision" relates to the selection of the hypothesis, and "sequentially related testing procedure" refers to the ordering of testing in a multi-stage testing process.

Wald (1952) defined sequential analysis as "a method of statistical inference whose characteristic feature is that the number of observations required by the procedure is not determined in advance of the experiment. The decision to terminate the experiment depends, at each stage, on the results of the observations previously made" (p. 1). Sequential analysis is often used in medical research (e.g., Anscombe, 1963; Armitage, 1960; Whitehead, 1983, 1987; etc.), probably because fewer subjects are required in sequential trials than in fixed trials (Lewis, 1990).

A sequential decision involves the sequential examination of hypotheses. Sobel and Wald (1949) discussed a sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. Consider a variable $\underline{x}$ which is normally distributed with known variance σ^2, but with an unknown mean μ. Given two real numbers $a_1 < a_2$ and a set of hypotheses to be examined, say, H_1: $\mu < a_1$, H_2: $a_1 \leq \mu \leq a_2$, and H_3: $\mu > a_2$, the problem is to choose one of these three mutually exclusive and exhaustive hypotheses. This is a process of making decisions about a sequence of hypotheses.

The third term, to be used in this study, is "sequentially related testing procedure." Such a procedure does not draw observations sequentially, nor does it involve sequential decisions about several alternative hypotheses. It involves testing more than one hypothesis in sequence for one set of data. The sequentially related hypotheses tested imply that one will test a second qualitatively different hypothesis only after a specific decision is made at stage 1.

When tests are sequentially related, it is natural to consider the relationship of the testing errors among the tests. Will the testing error in the first test influence errors made in conducting the next test? Does the impact involve either one of, or both, type I and type II errors? Effect-size meta-analysis involves the process of

sequentially related testing, since many effect-size meta-analyses involve the two-stage testing procedure outlined above in (7) and (7a) in Chapter IV. Therefore, in studying the power of the homogeneity test in effect-size meta-analysis, the sequential impact of testing errors is a concern.

In this chapter, I will discuss the influence of sequentially related hypothesis test, and I will examine the impact of the first-stage decisions on the second-stage statistical errors.

Two-Stage Testing

Effect-size meta-analyses involve at least two tests in sequence: the homogeneity test for the consistency of the effect sizes and the test for the magnitude of the common effect. When the study effects are determined to be homogeneous, one further estimates the value of the probable common population effect and tests whether the common value is zero.

For example, consider a review of sex differences on science achievement for grade-school students. After computing effect sizes from a series of studies, the reviewer first tests the homogeneity of all effects to decide whether they are consistent. If the homogeneity of effects is accepted, the reviewer then tests to determine whether gender has an effect on science achievement. If the

homogeneity test for the full set of effects is rejected, one decides that the magnitudes of sex differences on science achievement may vary. To proceed with the analysis, one either considers effects to be random, or seeks homogeneity within smaller groups of effects. For instance, effects may vary with grade levels, such that girls perform better than boys only in certain grade levels. The homogeneity test would then be performed on the effects for each grade level. If homogeneity of effects is accepted within a subgroup or grade level, the second-stage test measuring the magnitude of the average sex differences will be conducted for that subgroup.

Influence of Sequentially Related Hypothesis Testing on Statistical Errors

The role of sequentially related hypothesis testing in determining statistical errors is observed below in two situations: acceptance or rejection of the overall homogeneity test at the first stage.

Acceptance of the Overall Homogeneity Test

Since the test for homogeneity and the test for the common population effect are sequentially related, the validity of the former test can affect the validity of the latter. If at stage one, the analyst made a type II error in the homogeneity test, the second stage test for the common effect is misleading. Precisely, when population

effects are heterogeneous, the estimate of the effect-size in the second stage test is an estimate of an "average" effect (μ_δ) from a set of random effects rather than of the "common" effect (δ) representing a set of equal effects. The interpretation of the test for the "average" effect should differ from the interpretation of the test for the "common" effect. As in the case of a random-effects analysis-of-variance model, in the heterogeneous case population effects are random numbers with some distribution (i.e., $\sigma^2_\delta \neq 0$). Sampled effect sizes do not share one population effect. Wrongly accepting the homogeneity of effects will treat an average effect as the common effect.

The variance used for calculating the $\underline{z}$ statistic for testing the hypothesis H_0: $\delta = 0$ under the assumption of homogeneity will not reflect the variation of population effects. The estimate of the variance used for the test statistic for the hypothesis in (7a) (on p. 16) at the second stage will be too small. Instead of using the estimate of $(\sigma^2_\delta + \sigma^2(\underline{d}_i|\delta_i))$ for the variance of the $\underline{i}$th effect size, calculation of the $\underline{z}$ statistic (say, $\underline{z}_F$) under the decision of homogeneity would use the estimate of $\sigma^2(\underline{d}_i|\delta_i)$. Therefore, when the effects are heterogeneous (i.e., $\sigma^2_\delta > 0$), the test statistic $\underline{z}_F$ tends to be too large, which likely results in a greater chance of type I error (false rejection) or "too much power" in the second stage test.

Rejection of the Overall Homogeneity Test

When the overall homogeneity test is rejected, one assumes that several "true" effects may exist. One common approach to further study of these effects is to divide the collection of effect sizes into subgroups by certain factors and repeat the homogeneity test for each subgroup. Another approach to analyzing these effects is applying a random-effects model and testing for the average population effect.

As mentioned above, errors at the first stage will impact the validity of tests at the second stage. When a false rejection is made at the first stage, dividing effects into small groups can lead to more errors. First, because the population effects are truly homogeneous, classifying the effects from the same population into subclasses and conducting separate analyses is unnecessary. Second, the effective sample sizes for z tests for each subgroup are obviously reduced from the total sample size used for the z test for the whole group. Therefore, when population effect-sizes are homogeneous, tests of homogeneity for smaller subgroups are conservative or less powerful relative to the one test for the whole group.

Applying random-effects tests at the second stage is sometimes considered after rejection of the homogeneity test. In a random-effects test, the variance used for calculating the z statistic (denoted z_R here) will include an estimate of the variation in population effects (σ^2_δ).

Including the estimated variance of population effects (σ^2_δ) rather then using $\sigma^2(\underline{d}_i|\delta_i)$ alone would overestimate the variance when population effects are actually consistent. The $\underline{z}_R$ test statistic will then be too small and become less powerful than tests using $\underline{z}_F$ under the fixed-effects model (which should be applied when effects are truly homogeneous).

The additional simulation in this Chapter will examine the statistical errors and the appropriateness of tests using fixed- versus random-effects models. The simulation addresses the following questions: When the homogeneity test at the first stage is rightly rejected or wrongly rejected will the statistical error rates of the $\underline{z}$ tests ($\underline{z}_F$ and $\underline{z}_R$) at the second stage be similar? Specifically, when the homogeneity test is wrongly rejected (a type I error occurs at stage one), how much is the power of the $\underline{z}_R$ test (i.e., assuming random effects at the second stage) decreased? And, when the homogeneity test is wrongly accepted (a type II error occurs at stage one), how much is the power of $\underline{z}_F$ increased?

Summary

In conclusion, when the overall homogeneity test is wrongly accepted (a type II error) at the first stage, the fixed-effects model test $\underline{z}_F$ would be wrongly applied at stage two. Two errors will be made: the test is (1) conceptually invalid, and (2) subject to type I error. When

the overall homogeneity test is wrongly rejected (a type I error) at the first stage, the test at the second stage should be less powerful when the random-effects test (Z_R) is wrongly applied. Table 30 illustrates the relationship among two-stage sequential testing errors.

Table 30

Two-stage Testing Errors

True State

$\delta_1 = \delta_2 = \ldots = \delta_k = \delta$ At least one δ_i differs

Decision

Homo

A		B
True State: $\delta = 0$ / $\delta \neq 0$		Type II; True State: $\mu_\delta = 0$ / $\mu_\delta \neq 0$
$\delta = 0$: — / β		$\delta = 0$: — / β
$\delta \neq 0$: α / $1 - \beta$		$\delta \neq 0$: α / $1 - \beta$

Not Homo

C	D
Type I; True State: $\delta = 0$ / $\delta \neq 0$	True State: $\mu_\delta = 0$ / $\mu_\delta \neq 0$
$\mu_\delta = 0$: — / β	$\mu_\delta = 0$: — / β
$\mu_\delta \neq 0$: α / $1 - \beta$	$\mu_\delta \neq 0$: α / $1 - \beta$

In Table 30, the four main cells represent the first-stage test. For convenience, these cells are named A, B, C, and D (marked at their upper right corners). The second-stage tests and their statistical errors are illustrated by small tables within each cell of the large table. Population effects for the first stage are denoted by δ_is. The common population effect for the homogeneous effects is

δ. The average population effect for the heterogeneous effects is denoted as μ_δ. Cells marked "Type I" or "Type II" represent occurrences of the two types of statistical errors.

From the above summary, I predicted that the second-stage tests in cell C using the random-effects test ($\underline{z}_R$) may have higher type II error rates than the correct fixed-effects test $\underline{z}_F$. And second-stage tests in cell B using the fixed-effects test $\underline{z}_F$ may have lower type II error rates than the correct random-effects test ($\underline{z}_R$), and may have higher type I error rates.

Second stage $\underline{z}$ tests to test the hypotheses H_0: $\mu_\delta = 0$ vs. H_1: $\mu_\delta \neq 0$ for fixed-effects and random-effects models are:

$$\underline{z}_F = \frac{\underline{d}.}{1 \sqrt{\Sigma(1/\hat{\sigma}^2(\underline{d}_i|\delta_i))}}, \tag{36}$$

and

$$\underline{z}_R = \frac{\underline{d}.}{1 \sqrt{\Sigma(1/[\hat{\sigma}^2_\delta+\hat{\sigma}^2(\underline{d}_i|\delta_i)])}}, \tag{37}$$

where $\underline{d}.$ is the average effect weighted by precision,

$$\underline{d}. = \underline{w}_i\underline{d}_i, \tag{38}$$

$$\underline{w}_i = \frac{1/\underline{S}_i^2}{\sqrt{\Sigma(1/\underline{S}_i^2)}}. \quad (39)$$

The estimators of the variances $\underline{S}^2_i$ for fixed- and random-effects differ:

For fixed-effects,

$$\underline{S}_i^2 = \hat{\sigma}^2(\underline{d}_i|\delta_i). \quad (40)$$

For random-effects,

$$\underline{S}_i^2 = \hat{\sigma}^2_\delta + \hat{\sigma}^2(\underline{d}_i|\delta_i). \quad (41)$$

The estimator of the variance of population effects was an estimate developed by Hedges and Olkin (Hedges & Olkin, 1985), specifically:

$$\hat{\sigma}^2_\delta = \underline{S}^2(\underline{d}_i) - (1/\underline{k}) \Sigma \hat{\sigma}^2(\underline{d}_i|\delta_i), \quad (42)$$

where $\underline{S}^2(\underline{d}_i)$ is the usual sample variance computed using the $\underline{d}_i$ values as data.

Simulation of Power for Sequential Tests

Power values for the $\underline{z}$ tests were constructed through further simulation. Counts of both type I and type II errors for the second stage $\underline{z}$ tests were noted. Simulation will allow me to determine (1) whether or not the preset significance level of the $\underline{z}$ test is maintained, and (2) whether or not the second-stage $\underline{z}$ test given errors at the

first stage is as powerful as it is following correct decisions.

Factors that produced high or low power in the homogeneity tests are crucial in studying errors of subsequent z tests. The power simulation in Chapter IV indicated that for certain non-normal distributions of δ values and for effects with small sample sizes the actual power of homogeneity tests was greater than power based on the asymptotic theory . The primary goal of this Chapter is to examine the statistical errors of the second-stage based upon the decision at the first stage. Extra focus was on the subsequent level of errors at the second stage in conditions that showed higher power for the homogeneity test at stage one. Results from "non-normal" sets of δs (or "sets of δs with extreme values") or small sample sizes were compared to those from more evenly distributed sets of δs or large samples.

Factors for Simulation of Subsequent z Tests

Factors from previous simulations were chosen for the simulation of z-test behavior. The fixed-effects models were used to fully demonstrate the subsequent impact of the power of the first-stage test on the power of the second-stage test. Those combinations of factors that had resulted in differences between the simulated and asymptotic power values of homogeneity tests were closely examined. Other factors used in the additional simulation were the same as

those for the simulation in Chapter IV, with the elimination of (1) cases where $\underline{k} = 2$, and (2) patterns of population effects with two extreme values.

The simulation procedure for the power of the second-stage $\underline{z}$ tests followed the simulation for homogeneity tests in Chapter IV:

A. Test significance of the homogeneity test (at $\alpha = 0.05$). Consider the second-stage test to occur in one of the four decision categories based on the homogeneity test and the known pattern of δ values. The four categories (shown as A through D in Table 30) are rightly accepting homogeneity, wrongly accepting homogeneity, rightly rejecting homogeneity, or wrongly rejecting homogeneity.

B. Calculate two $\underline{z}$ statistics using the two estimates of variance, and note which would be used based on the decision about homogeneity. (using $\underline{z}_R$ if homogeneity is rejected, or $\underline{z}_F$ if homogeneity is accepted) for each of 2000 sets of generated effects.

C. Continue to replicate until the count of $\underline{z}$ tests in each category of decision based on the homogeneity test reaches 2000 replications.

D. Compute proportions of $\underline{z}$ statistics (across the 2000 replications) exceeding normal critical values at various significance levels separately for the

above four decision categories.

E. Calculate theoretical power values for both fixed- and random-effects tests (z_F and z_R) based on the known parameters δ_i, i = 1 to k.

F. Compare proportions of the significant z statistics (as power values) with the theoretical power values.

G. Determine if z tests were more powerful for cell B (z_F vs. z_R) than for cell A or less powerful for cell C (z_R vs. z_F) than for cell D. (Note that in cell A and cell D, the z tests used would have been computed with the correct estimate of variance.)

Results

Simulated power values for z tests from the second stage of effect-size meta-analysis were compared to theoretical power values. Analysis of power for z were carried out for each of the four decision categories for the homogeneity test at the first stage: (A) rightly accept homogeneity test, (B) wrongly accept homogeneity test, (C) wrongly reject homogeneity test, or (D) rightly accept homogeneity test.

Simulated vs. Theoretical Power Values

Simulated power values based on tests with fixed- or random-effects variance estimates were compared with the corresponding theoretical power, based on either the fixed-

or random-effects variance parameters.

Under the true state of homogeneity, theoretical power of both fixed- and random-effects tests were equal since the variance of population effects (σ^2_δ) was zero. Under heterogeneity, theoretical power values for random-effects tests were less than values for fixed-effects tests because the random-effects test z_R used a larger variance value (in its denominator).

Fixed-effects tests. Theoretical power values were calculated with the fixed-effects variance $\sigma^2(d_i|\delta_i)$. The simulated power values were obtained by computing z_F with the estimated fixed-effects variance (using d_i for δ_i in formula (5)).

When effects were homogeneous, and the stage-one decision about homogeneity was correct (in cell A), theoretical power values for z_F were slightly greater than simulated power values. The difference decreased as sample sizes increased. At $\alpha = 0.05$, for common effect $\delta = 0$, the mean difference across all homogeneous groups was .003 (.050-.047). A paired t test for the equality of simulated and theoretical power means was 4.36 ($df = 47$, $p < .001$). When power was analyzed according to sample size and k the mean difference in theoretical versus simulated power was significant only for sample sizes $n = 20$. Paired t tests on mean theoretical and simulated power values for z_F for homogeneous groups with different sample sizes and $\delta = 0$ are

listed in Table 31.

Table 31

Paired *t* Tests on Mean Theoretical and Simulated z_F Power for Homogeneous Effects with $\delta = 0$ ($\alpha = 0.05$)

k	*N*	Mean Diff.*	Sd	Se	Paired *t*	*df*	*p*
5	20*k*	.0069	.005	.002	2.85	3	.065
	60*k*	.0043	.009	.004	.99	3	.395
	120*k*	.0040	.006	.003	1.31	3	.281
	200*k*	.0023	.003	.002	1.41	3	.254
10	20*k*	.0070	.004	.002	3.20	3	.049*
	60*k*	-.0016	.003	.002	-1.02	3	.385
	120*k*	.0033	.005	.003	1.28	3	.290
	200*k*	.0003	.006	.003	.09	3	.934
30	20*k*	.0100	.003	.002	5.73	3	.011*
	60*k*	-.0029	.002	.001	-3.05	3	.056
	120*k*	.0030	.003	.001	2.10	3	.127
	200*k*	.0048	.004	.002	2.66	3	.076

Note: * $p < .05$, positive mean difference indicates theoretical power > simulated power.

For homogeneous effects with $\delta > 0$, the mean difference was .01 (.719-.709). The paired *t* test value was 7.56 (*df* = 143, *p* < .001). Like the case in which $\delta = 0$, the difference also decreased as sample size increased. Results were similar for $\delta = 0.1$, 0.2, or 0.3. Paired *t* tests on theoretical minus simulated power values for fixed-effects tests (z_F) for homogeneous groups with different sample sizes and $\delta > 0$ are listed in Table 32.

Table 32

Paired *t* Tests on Mean Theoretical and Simulated z_F Power for Homogeneous Effects with $\delta > 0$ ($\alpha = 0.05$)

k	*N*	Mean Diff.*	Sd	Se	Paired *t*	df	p
5	20*k*	.0237	.012	.003	7.04	11	.000#
	60*k*	.0073	.011	.003	2.30	11	.042*
	120*k*	.0032	.009	.003	1.26	11	.234
	200*k*	.0014	.007	.002	.72	11	.484
10	20*k*	.0364	.012	.003	10.87	11	.000#
	60*k*	.0084	.011	.003	2.54	11	.028*
	120*k*	.0002	.006	.002	.13	11	.899
	200*k*	.0002	.004	.001	.21	11	.838
30	20*k*	.0296	.020	.006	5.14	11	.000#
	60*k*	.0044	.008	.002	1.99	11	.072
	120*k*	.0017	.005	.001	1.27	11	.229
	200*k*	.0005	.002	.001	.85	11	.412

Note: * $p < .05$, positive mean difference indicates theoretical power > simulated power.
# $p < .001$.

Next I applied the modified Kolmogorov-Smirnov test, with critical value $D^* = 0.030$, to the distribution of (theoretical power - simulated power) values. For homogeneous population effects with $\delta = 0$, only 1 of 48 combinations showed a significant difference between the theoretical and simulated z_F power functions. When $\delta > 0$, 22% (32/144) had significant discrepancies in which. The theoretical power values were greater than the simulated ones. Discrepancies increased as the sample size decreased. Discrepancies were independent of the number of effect sizes *k*, the value of δ, and the sampling fractions between or

within studies. Frequencies are listed by numbers of effects k, total sample sizes N, equal vs. unequal sample sizes between studies (π_i), within study sample-size balance (ϕ_i), or the value of the common effect δ in Tables 33 to 37.

Table 33

Frequencies of Significant Discrepancies for Power of z_F by k for Homogeneous Effects with $\delta > 0$

Significant Discrepancy	Number of effect-sizes (k) 5	10	30	Total
Yes	8 (17%)	11 (23%)	13 (27%)	110 (22%)
No	40	37	35	112 (78%)
Total	48	48	48	144

$\chi^2 = 1.527$ ($df = 2$, $p = .466$)

Table 34

Frequencies of Significant Discrepancies for Power of z_F by N for Homogeneous Effects with $\delta > 0$

Significant Discrepancy	Total sample sizes (N) 20k	60k	120k	200k	Total
Yes	26(72%)	6(17%)	0	0	32 (22%)
No	10	30	36	36	112 (78%)
Total	36	36	36	36	144

$\chi^2 = 73.286$ ($df = 3$, $p < .001$)

Table 35

Frequencies of Significant Discrepancies for Power of z_r by π_i for Homogeneous Effects with $\delta > 0$

Significant Discrepancy	Sampling fraction between studies (π_i) Balanced	Unbalanced	Total
Yes	18 (25%)	14 (19%)	32 (22%)
No	54	58	112 (78%)
Total	72	72	144

$\chi^2 = 0.643$ ($df = 1$, $p = .423$)

Table 36

Frequencies of Significant Discrepancies for Power of z_r by ϕ_i for Homogeneous Effects with $\delta > 0$

Significant Discrepancy	Sampling fraction within studies (ϕ_i) Balanced	Unbalanced	Total
Yes	18 (25%)	14 (19%)	32 (22%)
No	54	58	112 (78%)
Total	72	72	144

$\chi^2 = 0.643$ ($df = 1$, $p = .423$)

Table 37

Frequencies of Significant Discrepancies for Power of z_F by δ for Homogeneous Effects with $\delta > 0$

Discrepancy	Common population effect (δ) 0.1	0.2	0.3	Total
Yes	7 (15%)	12 (25%)	13 (27%)	32 (22%)
No	41	36	35	112 (78%)
Total	48	48	48	144

$\chi^2 = 2.491$ ($df = 2$, $p = .288$)

When population effects are truly heterogeneous, fixed-effects tests are not appropriate (in cell B and D). However, the simulated power values for z_F were also compared with the theoretical power values calculated with the fixed-effects variances in cell B because in this case the stage-one decision implies that z_F should be used. At $\alpha = 0.05$, theoretical power values were significantly less than simulated power values, with a mean difference of -.040, and the paired t-test value was -8.13 ($df = 375$, $p < 0.001$). Results were similar across sample sizes. Paired t tests on theoretical and simulated power values for z_F for heterogeneous groups with different sample sizes are listed in Table 38.

Table 38

Paired *t* Tests on Theoretical and Simulated Power of z_F for Heterogeneous Effects ($\alpha = 0.05$)

k	*N*	Mean Diff.*	Sd	Se	Paired *t*	*df*	*p*
5	20*k*	.0113	.021	.004	2.62	23	.015*
	60*k*	-.0138	.024	.005	-2.79	23	.010*
	120*k*	-.0157	.034	.007	-2.27	23	.033*
	200*k*	-.0219	.041	.008	-2.64	23	.015*
10	20*k*	-.0361	.082	.014	-2.65	35	.012*
	60*k*	-.0590	.084	.014	-4.22	35	.000*
	120*k*	-.0626	.121	.020	-3.09	35	.004*
	200*k*	-.0665	.146	.024	-2.73	35	.010*
30	20*k*	-.0419	.082	.014	-3.07	35	.004*
	60*k*	-.0396	.086	.014	-2.77	35	.009*
	120*k*	-.0368	.090	.015	-2.47	35	.019*
	200*k*	-.0554	.143	.027	-2.04	27	.051

Note: * $p < 0.05$, positive mean difference indicates theoretical value > simulated value.

Results of the modified Kolmogorov-Smirnov test for heterogeneous population effects with fixed-effects tests showed that 51% of 376 combinations showed a significant difference between the theoretical and simulated z_F power. Most significant discrepancies were negative, that is, simulated values were higher than the theoretical values. Positive discrepancies were more common for smaller sample sizes. That is, when sample sizes were small, some theoretical values were higher than simulated power values.

Discrepancies were not associated with patterns of population effects. Discrepancies were independent of the sampling ratio within studies, but were associated with sampling fraction between studies. When studies with large effects had large sample sizes, the simulated values were consistently higher than theoretical values. When sample sizes across studies were equal, the simulated values were consistently lower than the theoretical values. Crosstabulation of significant discrepancies are listed in Tables 39 to 43.

Table 39

Frequencies of Significant Discrepancies for Power of z_r by k for Heterogeneous Effects

Significant Discrepancy	Number of effect-sizes (k) 5	10	30	Total
Yes	47 (49%)	80 (56%)	66 (49%)	193 (51%)
No	49	64	70	183 (49%)
Total	96	144	136	376

$\chi^2 = 1.672$ ($df = 2$, $p = .433$)

Table 40

Frequencies of Significant Discrepancies for Power of z_r by by N for Heterogeneous Effects

Significant Discrepancy	Total sample sizes (N) 20k	60k	120k	200k	Total
Yes	67 (70%)	49 (51%)	42 (44%)	35 (40%)	193 (51%)
No	29	47	54	53	183
Total	96	96	96	88	376

$\chi^2 = 20.0133$ ($df = 3$, $p < .001$)

Table 41

Significant Discrepancies for Power of z_r by Pattern of δ_i for Heterogeneous Effects

Significant Discrepancy	Pattern of population effects: One Extreme	Three Subsets	Five Subsets	Total
Yes	70 (49%)	74 (53%)	49 (53%)	193 (51%)
No	74	66	43	183 (49%)
Total	144	140	92	376

$\chi^2 = 0.694$ ($df = 2$, $p = .707$

Table 42

Frequencies of Significant Discrepancies for Power of z_F by π_i for Heterogeneous Effects

Significant Discrepancy	Sampling fraction between studies (π_i) Balanced	Unbalanced	Total
Yes	34 (18%)	159 (85%)	193 (51%)
No	154	29	183 (49%)
Total	188	188	376

$\chi^2 = 166.341$ ($df = 2$, $p < .001$)

Table 43

Frequencies of Significant Discrepancies for Power of z_F by ϕ_i for Heterogeneous Effects

Significant Discrepancy	Sampling fraction within studies (ϕ_i) Balanced	Unbalanced	Total
Yes	97	96	193 (51%)
No	91 (48%)	92 (49%)	183 (49%)
Total	188	188	376

$\chi^2 = 0.0107$ ($df = 2$, $p = .918$)

Random-effects tests. Theoretical power values for z_R were calculated with the random-effects variance $\sigma^2_\delta + \sigma^2(d_i|\delta_i)$. The simulated power values were obtained with the estimate of the random-effects variance (see formulas (40) and (41)).

When population effects are truly homogeneous, random-effects tests are not appropriate (in cell A and C). However, in cell C the decision made at stage one is to reject H_0, thus this decision would lead (incorrectly) to the use of z_R at stage two. At $\alpha = 0.05$, the discrepancy between the theoretical and simulated power values for z_R was large (in comparison to that for z_F, the fixed-effects test). When $\delta = 0$ the mean difference across all sample groups was .041 (.050-.009). The paired t-test value was 49.92 ($df = 47$, $p < 0.001$), showing that the theoretical values were significantly greater than the simulated power values. Paired t tests on theoretical and simulated power values of z_R for homogeneous groups with $\delta = 0$ and for different sample sizes are listed in Table 44.

For $\delta > 0$, at $\alpha = 0.05$, the mean difference across all sample sizes was .1832 (.7187-.5355). The paired t test was 15.90 ($df = 143$, $p < .001$) which showed that theoretical values were significantly greater than simulated power values. Results were similar across sample sizes. However, as power values approached 1 for some large samples the differences were forced to decrease. Paired t tests on mean theoretical and simulated power values for z_R for homogeneous groups with $\delta > 0$ for different sample sizes are listed in Table 45.

Table 44

Paired *t* Tests on Theoretical and Simulated Power of z_R for Homogeneous Effects with $\delta = 0$ ($\alpha = 0.05$)

k	*N*	Mean Diff.*	Sd	Se	Paired *t*	*df*	*p*
5	20*k*	.0450	.002	.001	48.11	3	.000
	60*k*	.0474	.002	.001	55.68	3	.000
	120*k*	.0485	.001	.000	168.01	3	.000
	200*k*	.0453	.001	.001	62.70	3	.000
10	20*k*	.0421	.001	.001	64.07	3	.000
	60*k*	.0401	.001	.001	84.79	3	.000
	120*k*	.0425	.002	.001	38.66	3	.000
	200*k*	.0419	.001	.001	58.32	3	.000
30	20*k*	.0356	.007	.004	10.16	3	.002
	60*k*	.0351	.005	.002	14.27	3	.001
	120*k*	.0338	.002	.001	43.42	3	.000
	200*k*	.0346	.005	.002	14.53	3	.001

Note: * positive mean difference indicates that theoretical power > simulated power.

Table 45

Paired *t* Tests on Theoretical and Simulated Power of z_R for Homogeneous Effects with $\delta > 0$ ($\alpha = 0.05$)

k	N	Mean Diff.*	Sd	Se	Paired t	df	p
5	20k	.2089	.091	.026	8.00	11	.000
	60k	.3228	.106	.030	10.58	11	.000
	120k	.3035	.100	.029	10.47	11	.000
	200k	.2433	.153	.044	5.51	11	.000
10	20k	.2380	.089	.026	9.29	11	.000
	60k	.2285	.085	.024	9.35	11	.000
	120k	.1627	.127	.037	4.43	11	.001
	200k	.1265	.157	.045	2.79	11	.018
30	20k	.2072	.101	.029	7.09	11	.000
	60k	.0887	.107	.031	2.87	11	.015
	120k	.0491	.080	.023	2.31	11	.056
	200k	.0192	.034	.010	1.98	11	.073

Note: * positive mean difference indicates that theoretical power > simulated power.

Applying the modified Kolmogorov-Smirnov test to difference based on homogeneous population effects, when $\delta = 0$, all 48 combinations showed significant difference between the theoretical and simulated *z* power. One half of the significant discrepancies was positive and the other half was negative. Significant discrepancies for $\delta = 0$ was not associated with any simulation factors.

When $\delta > 0$, 89% of 144 combinations had significant discrepancies, most of which were positive. The theoretical power values were greater than the simulated ones. Discrepancies increased when the number of effects *k* and the sample size *N* decreased. When the value of δ decreased

discrepancies also increased. Discrepancies were independent of the sampling fraction either between or within studies. Frequencies are listed by number of effects k, total sample sizes N, and equal vs. unequal sample sizes between study sample sizes (π_i), within study sample size balance (ϕ_i), and value of the common effect δ in Tables 46 to 50.

Table 46

Frequencies of Significant Discrepancies for Power of z_R by k for Homogeneous Effects ($\delta > 0$)

Significant Discrepancy	Number of effect-sizes (k) 5	10	30	Total
Yes	48 (100%)	47 (98%)	33 (69%)	128 (89%)
No	0	1	15	16 (11%)
Total	48	48	48	144

$\chi^2 = 29.672$ (df = 2, p < .001)

Table 47

Frequencies of Significant Discrepancies for Power of z_R by N for Homogeneous Effects ($\delta > 0$)

Significant Discrepancy	Total sample sizes (N) 20k	60k	120k	200k	Total
Yes	36 (100%)	34 (94%)	31 (86%)	27 (75%)	128 (89%)
No	0	2	5	9	16
Total	36	36	36	36	144

$\chi^2 = 12.938$ (df = 3, $p < .01$)

Table 48

Frequencies of Significant Discrepancies for Power of z_R by π_i for Homogeneous Effects ($\delta > 0$)

Significant Discrepancy	Sampling fraction between studies (π_i) Balanced	Unbalanced	Total
Yes	62 (86%)	66 (92%)	128 (89%)
No	10	6	16 (11%)
Total	72	72	144

$\chi^2 = 1.125$ (df = 1, $p = .289$)

Table 49

Frequencies of Significant Discrepancies for Power of z_R by ϕ_i for Homogeneous Effects ($\delta > 0$)

Significant Discrepancy	Sampling fraction within studies (ϕ_i) Balanced	Unbalanced	Total
Yes	63 (88%)	65 (90%)	128 (89%)
No	9	7	16 (11%)
Total	72	72	144

$\chi^2 = 0.281$ ($df = 1$, $p = .596$)

Table 50

Frequencies of Significant Discrepancies for Power of z_R by δ for Homogeneous Effects ($\delta > 0$)

Significant Discrepancy	Common population effect (δ) 0.1	0.2	0.3	Total
Yes	48 (100%)	43 (90%)	37(77%)	128 (89%)
No	0	5	11	16 (11%)
Total	48	48	48	144

$\chi^2 = 12.79752$ ($df = 2$, $p < .01$)

When the population effects were heterogeneous and the first stage hypothesis is rejected (in cell D), the random-effects test z_R was the correct test. At $\alpha = 0.05$,

theoretical power values for z_R were greater than the simulated values. The mean difference across all sample sizes of 0.10 was significant (.463-.363), with a t = 18.67 (df = 375, $p < 0.001$). Results were similar across sample sizes. As above, when power values approached 1 for some large samples, discrepancies were limited and reduced. Paired t tests on theoretical and simulated z_R power values for heterogeneous groups and different sample sizes are listed in Table 51.

Table 51

Paired t Tests on Theoretical and Simulated Power of z_R for Heterogeneous Effects (α = 0.05)

k	N	Mean Diff.*	Sd	Se	Paired t	df	p
5	20k	.1467	.089	.018	8.06	23	.000
	60k	.2021	.116	.024	8.56	23	.000
	120k	.1776	.088	.018	9.90	23	.000
	200k	.1445	.107	.022	6.60	23	.000
10	20k	.1572	.087	.015	10.84	35	.000
	60k	.1454	.073	.012	12.01	35	.000
	120k	.0871	.105	.017	4.98	35	.000
	200k	.0447	.097	.016	2.75	35	.009
30	20k	.1057	.053	.009	11.87	35	.000
	60k	.0380	.061	.010	3.75	35	.001
	120k	.0141	.045	.007	1.88	35	.068
	200k	-.0012	.062	.012	- .10	27	.919

Note: * positive mean difference indicates theoretical power > simulated power.

Applying the modified Kolmogorov-Smirnov test to power functions for z_R for heterogeneous effects, almost all (96%)

of the 376 combinations showed significant differences between the theoretical and simulated power. Most of the significant differences were negative. About 33% of the measures (376 × 15 = 5640 measures) showed that theoretical power values were less than the simulated ones, and 13% showed that theoretical values were less than the simulated values. Significant discrepancies decreased as sample size N or the number of effects k increased. Discrepancies occurred more when population effects had extreme values than when population effects were more evenly dispersed. Frequencies are listed by number of effects k, total sample sizes N, and equal vs. unequal sample sizes between study sample sizes (π_i) and within study sample-size balance (ϕ_i) in Tables 52 to 56.

Table 52

Frequencies of Significant Discrepancies for Power of z_R by k for Heterogeneous Effects

Significant Discrepancy	Number of effect-sizes (k) 5	10	30	Total
Yes	96(100%)	144(100%)	120 (88%)	360 (96%)
No	0	0	16	16 (4%)
Total	96	144	136	376

$\chi^2 = 29.490$ (df = 2, p < .001)

Table 53

Frequencies of Significant Discrepancies for Power of z_R by N for Heterogeneous Effects

Significant Discrepancy	Total sample sizes (N) 20k	60k	120k	200k	Total
Yes	96 (100%)	96 (100%)	88 (92%)	80 (91%)	360 (96%)
No	0	0	8	8	16
Total	96	96	96	88	376

$\chi^2 = 17.502$ (df = 3, p < .001)

Table 54

Frequencies of Significant Discrepancies for Power of z_R by π_i for Heterogeneous Effects

Significant Discrepancy	Sampling fraction between studies (π_i) Balanced	Unbalanced	Total
Yes	180 (96%)	180 (96%)	360 (96%)
No	8	8	16 (4%)
Total	188	188	376

$\chi^2 = 0.000$ (df = 1, p = 1.000)

Table 55

Frequencies of Significant Discrepancies for Power of z_R by ϕ_i for Heterogeneous Effects

Significant Discrepancy	Sampling fraction within studies (ϕ_i) Balanced	Unbalanced	Total
Yes	180 (96%)	180 (96%)	360 (96%)
No	8	8	16 (4%)
Total	188	188	376

$\chi^2 = 0.000$ ($df = 1$, $p = 1.000$)

Table 56

Significant Discrepancies for Power of z_R by Pattern of δ_i for Heterogeneous Effects

Significant Discrepancy	Pattern of Population Effects: One Extreme	Three Subsets	Five Subsets	Total
Yes	144(100%)	132(94%)	84 (91%)	360 (96%)
No	0	8	8	16 (4%)
Total	144	140	92	376

$\chi^2 = 11.584$ ($df = 2$, $p < .01$)

Summary. In general, theoretical and simulated values matched better for large samples than small samples. Because they are based on asymptotic theory, the theoretical values should fit better for large samples. However, since

both power values had an upper limit, and both power values increased as the sample size increased, the discrepancies also tend to decrease as sample size increases because both power functions tend more quickly to one.

Theoretical values for z_F power fitted the best when homogeneity tests at the first stage were correctly accepted (in cell A). For homogeneous effects with $\delta = 0$, almost no significant discrepancies between simulated and theoretical power functions were found. When $\delta > 0$, most discrepancies occurred when sample size was small (e.g., $n_i = 20$), where theoretical values were significantly greater than the simulated values.

About half of the distributions showed significant discrepancies between theoretical and simulated power values for z_F when homogeneity was falsely accepted (in cell B). Discrepancies increased as sample sizes decreased. When studies had equal sample sizes (equal π_is), theoretical values were closer to the simulated values then when studies had unequal samples. When large effects were combined with large samples, the theoretical values were lower than the simulated values.

Power functions for random-effects tests (z_R) did not fit as well as those for fixed-effects tests. When homogeneity was falsely rejected (in cell C), for $\delta = 0$, all combinations had significant discrepancies (half were positive values, and the other half were negative values).

Significant discrepancies were not clearly associated with any other simulation factors. When $\delta > 0$, about nine tenth of the distributions had higher theoretical values. Discrepancies decreased as the number of effect sizes, the sample size, or the value of δ increased.

When homogeneity was correctly rejected (in cell D), almost all theoretical power values (96%) for $\underline{z}_R$ were significantly different from the simulated values. When population effects were fairly evenly distributed, theoretical values were higher than simulated values. When one population had one extreme effect-size value, theoretical values could be either higher or lower than the simulated values. Also discrepancies decreased as the number of effects $\underline{k}$ increased.

Results showed that overall theoretical power values did not fit well with the simulated values for random-effects tests ($\underline{z}_R$). Theoretical values were sometimes greater and sometimes less than simulated values. This result leads to a question about the precision of the estimate of the variance of population effects (σ^2_δ).

Hedges and Olkin (1985) gave an approximation to the distribution of the effect-size parameter-variance estimator. As they indicated, the estimator of the variance of population effects has an asymptotic normal distribution, however, the large sample normal approximation to the distribution of the estimate of σ^2_δ is probably not very

good unless the number of effects $\underline{k}$ is quite large. More needs to be known about the accuracy of the large sample approximation to the distribution of the estimate of the variance of population effects.

When effects were homogeneous, the power of the random-effects test $\underline{z}_R$ seemed excessively low. One possibility is that the variance of the population effects σ^2_δ for homogeneous effects ($\sigma^2_\delta = 0$) may be systematically overestimated (biased). When effects were heterogeneous, the estimate of the population variance seemed appropriate and may be more accurate.

The behavior of the estimator of the population variance based on different homogeneity decisions at stage one was studied via further simulation. Two sample sizes $\underline{n}_i$ of 20 and 60 were selected and two sets of effect-size parameters were set for the case where $\underline{k} = 5$. The average effect size was the same for both homogeneous and heterogeneous effects: the δ values for $\sigma^2_\delta = 0$ were (0.2, 0.2, 0.2, 0.2, 0.2), and for $\sigma^2_\delta > 0$ the effects were (0, 0.2, 0.2, 0.2, 0.4). 2000 replications were generated for both correct and incorrect decisions about homogeneity. When homogeneity was accepted values of the variance estimates were close to zero and were less dispersed for both homogeneous and heterogeneous effects. As predicted the bias of the estimate was greater when effects were homogenous than when effects were heterogeneous.

Power of z Based on Decisions about Homogeneity

Power values for $\underline{z}_F$ and $\underline{z}_R$ were compared at $\alpha = 0.05$. If the homogeneity was rightly accepted (in cell A) or rightly rejected (in cell D), the second stage $\underline{z}$ tests which follow from the stage-one decision are tests with correct variance components. No comparison was necessary when the correct $\underline{z}$ test was applied. When homogeneity was falsely accepted (cell B) or falsely rejected (cell C), the subsequent $\underline{z}$ test (suggested by the stage-one test) would use the estimate of the wrong variance and be incorrect. Since population effects were known values in the simulation, both $\underline{z}$ tests were calculated for cells B and C. Simulated power values were compared for the two tests (i.e., for tests using the correct versus incorrect variance).

Homogeneous population effects. When effects were homogeneous and the homogeneity was rejected (in cell C), the recommended $\underline{z}$ test on the average effect would be calculated as $\underline{z}_R$, that is, using the estimate of the random-effects variance $\sigma^2_\delta + \sigma^2(\underline{d}_i | \delta_i)$. However, the correct $\underline{z}$ test ($\underline{z}_F$) should use the estimate of the fixed-effects variance $\sigma^2(\underline{d}_i | \delta_i)$. Since the estimate of σ^2_δ must be greater than or equal to 0, power values based on $\underline{z}_R$ and the random-effects variance should always be less than values based on the fixed-effects test ($\underline{z}_F$).

For homogeneous effects with $\delta = 0$, at $\alpha = 0.05$, across

all sample groups the mean power difference between z_F and z_R was .0387 (.0477-.0090), with a paired t = 31.33 (df = 47, p < .001). When the common effect $\delta = 0$, the probability of falsely rejecting the z test is the type I error rate. Mean simulated power values showed that both z_F and z_R had smaller type I error rates (0.0477 and 0.009) than the preset α level (0.05). However, the size of z_R is much lower than either the α level or the size of z_F. When the number of effects k increased, mean differences between the power of z_F and z_R slightly decreased. Paired t tests on homogeneous effects with $\delta = 0$ for each sample-size group are listed in Table 57.

Table 57

Paired t Tests on Power (size) of z_F versus z_R for Homogeneous Effects with $\delta = 0$ ($\alpha = 0.05$) and Homogeneity Was Rejected

k	N	Mean Diff.*	Sd	Se	Paired t	df	p
5	20k	.0434	.004	.002	23.41	3	.000
	60k	.0432	.005	.003	16.42	3	.000
	120k	.0465	.008	.004	11.61	3	.000
	200k	.0522	.006	.003	16.37	3	.000
10	20k	.0327	.004	.002	15.26	3	.001
	60k	.0396	.002	.001	41.99	3	.000
	120k	.0417	.003	.001	30.66	3	.000
	200k	.0419	.001	.001	56.09	3	.000
30	20k	.0275	.009	.004	6.44	3	.008
	60k	.0327	.008	.004	8.14	3	.004
	120k	.0305	.003	.001	20.92	3	.000
	200k	.0320	.001	.001	52.26	3	.000

Note: * positive mean difference indicates power of z_F > power of z_R.

For homogeneous effects with $\delta > 0$, across all sample groups the mean power difference between z_F and z_R was .1751 (.7107-.5355), with a paired t = 15.65 (df = 143, $p < .001$). Power values increased as either the value of δ or the sample size increased. However, power values for fixed-effects tests (z_F) increased faster than those for random-effects tests (z_R) as either the value of δ or the sample size increase. When δ or the sample sizes were large, both power values approached 1. Mean power values for both tests for different sample sizes and δ values are listed in Table 58. Since population effects were homogeneous, the z_F test should still be the correct test here.

Table 58

Mean z Power Values of z_F versus z_R for Homogeneous Effects with $\delta > 0$ ($\alpha = 0.05$) and Homogeneity Was Rejected

		$\delta = 0.10$		$\delta = 0.20$		$\delta = 0.30$	
k	N	z_F	z_R	z_F	z_R	z_F	z_R
5	20k	.1081	.0186	.2296	.0445	.4036	.1087
	60k	.2142	.0299	.5015	.1391	.8014	.4012
	120k	.3246	.0605	.7630	.3372	.9735	.7619
	200k	.4495	.1002	.9200	.5912	.9990	.9609
10	20k	.1502	.0415	.3599	.1352	.6292	.3351
	60k	.3166	.1066	.7669	.4449	.9672	.8331
	120k	.5117	.2170	.9561	.7831	.9994	.9876
	200k	.6996	.3690	.9971	.9529	1.0000	.9999
30	20k	.2939	.1331	.7437	.4971	.9656	.8187
	60k	.6567	.4396	.9922	.9535	1.0000	.9992
	120k	.9061	.7530	1.0000	.9994	1.0000	1.0000
	200k	.9830	.9256	1.0000	1.0000	1.0000	1.0000

Heterogeneous population effects. When effects were heterogeneous but homogeneity was accepted (in cell B), the z test which follows from the stage-one decision whould typically be calculated as z_F, using the estimate of the fixed-effects variance $\sigma^2(d_i|\delta_i)$. The correct z test, however, is z_R, which should use the estimate of the random-effects variance $\sigma^2_\delta+\sigma^2(d_i|\delta_i)$. Here the power of the incorrect test (z_F) would be expected to be greater than the power of the correct test. The mean power difference between z_R minus z_F across all sample groups was -0.330 (.5122-.5452), paired t = - 20.05 (df = 375, p < .001). Mean power values for each sample group and patterns of δ_is are listed in Table 59.

Table 59

Mean z Power Values of z_F versus z_R for Heterogeneous Effects (α = 0.05) and Homogeneity Was Accepted

		One Extreme		Three Subsets		Five Subsets	
k	N	z_F	z_R	z_F	z_R	z_F	z_R
5	20k	.0792	.0654	.2808	.2389	-	-
	60k	.1387	.1077	.5922	.5214	-	-
	120k	.1997	.1437	.7745	.7188	-	-
	200k	.2796	.1955	.8638	.8261	-	-
10	20k	.0895	.0731	.4099	.3591	.4116	.3628
	60k	.1608	.1284	.7256	.6693	.7258	.6744
	120k	.2418	.1986	.8637	.8294	.8596	.8258
	200k	.3118	.2623	.9302	.9039	.9683	.9498
30	20k	.0747	.0641	.7292	.6989	.7203	.6947
	60k	.1233	.1053	.9198	.9020	.9224	.9055
	120k	.1711	.1490	.9802	.9750	.9806	.9745
	200k	.2272	.2018	.9958	.9938	.9960	.9941

The population effects for the pattern with one extreme case were (0, ..., 0, δ_k) and δ_k = 0.1, 0.2, or 0.25. The value of δ for three or five subsets, varied as 0.1, 0.2, or 0.25, and can also be viewed as the average effect. At α = 0.05, when the average effect was small, the differences between power values of the fixed- and random-effects tests increased as sample sizes increased, as was true for the one-extreme-value case. When the average effect was large, power values reached 1, and the differences between power values for the fixed- and random-effects tests were forced to diminish.

Summary. The power difference between the fixed- and random-effects tests at α = 0.05 increased as the value of the average effect or sample size increased. As the average effect or sample size became large, power approached 1 and the differences diminished. Power differences were smaller when the homogeneity of effects was falsely accepted (cell B) than when the homogeneity of effects was falsely rejected (cell C). The fixed-effects z test z_F was always the more powerful test.

Adjustment to Maintain the Desired Testing Error Rates

Caution needs to be taken in any sequentially related testing procedure. To achieve the desired significance level, sometimes, the criteria for the choice of the significance level at each stage needs to be adjusted. At

other times, corrections need to be made for estimation and tests of hypotheses.

In Chapter IV, the actual size of the homogeneity test for large k with all small samples ($n_i = 20$) was found to be greater than the preset α value (see Table 2 and Figure 4.1.4). In other words, there was a slightly higher chance (up to about 0.05 more) that homogeneity of effects would be falsely rejected for large k with small samples than for smaller k with large samples. Results in Chapter V showed that the use of an incorrect z test (i.e., with an incorrect variance) was associated with greater type I and type II error rates when homogeneity of effects was falsely rejected than when homogeneity was falsely accepted.

Meta-analysts who encounter many studies all with small samples need to be aware that the homogeneity test has an inflated type I error rate. Also subsequent z tests, erroneously computed with random-effects variances, will be much less sensitive to the magnitude of the common effect. In order to maintain a desired statistical error rate for H, for example 0.05, one may want to lower the nominal α level to 0.025 (for which simulated power was around 0.066) for the homogeneity test with many studies all having sample sizes less than or equal 20.

Power values and the type I error rates for the second-stage z tests were computed for selected cases to examine the consequences of lowering the α level from 0.05 to 0.025

for the homogeneity test at the first stage. For $\underline{k}$ = 30, $\underline{n}$ = 20, and homogeneous effect sizes with common effect δ = 0, the actual rejection rate of $\underline{H}$ was 0.0780 for a nominal α = 0.05. The actual rejection rate for the $\underline{z}_R$ test was 0.0185 when homogeneity was falsely rejected, and was 0.0435 for $\underline{z}_R$ when homogeneity was correctly rejected. For the same values of $\underline{k}$ and $\underline{n}$ for the homogeneity test with α = 0.025, the rejection rate of $\underline{H}$ was 0.0465. And the rejection rate for $\underline{z}_R$ test was 0.020 when homogeneity was falsely rejected, and the chance of rejecting was 0.0425 when homogeneity was correctly rejected.

The total rejection rates for the second-stage $\underline{z}$ tests at the 0.05 level, $P(R_2)$, were compared under the first-stage α values of 0.05 and 0.025 and can be written as below:

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|R^c{}_1)[1 - P(R_1)], \qquad (43)$$

where

$P(R_1)$ = the rejection rate of $\underline{H}$ at stage one,

$P(R^c{}_1) = 1 - P(R_1)$,

$P(R_2|R_1)$ = the chance of rejecting H_0: $\mu_\delta = 0$, given that homogeneity has been rejected, and

$P(R_2|R^c{}_1)$ = the chance of rejecting H_0: $\mu_\delta = 0$, given that homogeneity has been accepted.

For α = 0.05 at stage one:

$$P(R_2) = (0.0185)(0.0780) + (0.0435)(0.9220) = 0.0416.$$

For $\alpha = 0.025$ at stage one:

$$P(R_2) = (0.0200)(0.0465) + (0.0425)(0.9535) = 0.0415.$$

Thus here reducing the first-stage α does not impact the size of the $\underline{z}$ test procedure at all. When effect sizes were homogeneous with common effect $\delta = 0.2$, the rejection rates at the second stage, for first-stage α values 0.05 and 0.025 are,

for $\alpha = 0.05$ at stage one,

$$P(R_2) = (0.6100)(0.0860) + (0.7565)(0.9140) = 0.7439,$$

and for $\alpha = 0.025$ at stage one,

$$P(R_2) = (0.6140)(0.0465) + (0.7690)(0.9535) = 0.7618.$$

The lower α value at stage one here is associated with a slight increase in power at stage 2, which is beneficial since the stage 2 hypothesis is false ($\delta = 0.2$). When effect sizes were heterogeneous with average effect $\mu_\delta = 0$, the rejection rates at the second stage under first-stage α values 0.05 and 0.025 are,

for $\alpha = 0.05$ at stage one,

$$P(R_2) = (0.0160)(0.1815) + (0.0420)(0.8185) = 0.0373,$$

and for $\alpha = 0.025$ at stage one,

$$P(R_2) = (0.0150)(0.1210) + (0.0400)(0.8790) = 0.0370.$$

Again the change in the type I error rate is minimal, thus the reduce of stage-one α does not naturally affect the

stage-two α value. When effect sizes were heterogeneous with average effect $\mu_\delta = 0.2$, the rejection rates at the second stage for first-stage α values 0.05 and 0.025 are,

for $\alpha = 0.05$ at stage one,

$$P(R_2) = (0.5865)(0.1890) + (0.7635)(0.8110) = 0.7300,$$

and for $\alpha = 0.025$ at stage one:

$$P(R_2) = (0.5775)(0.1235) + (0.7625)(0.8765) = 0.7397.$$

Again a slight power increase is seen, though it is only minimal. However, in none of these instances is a reduction in stage-one α associated with detrimental effects at stage two. From the above comparison, one can conclude that lowering the significant level for the homogeneity test at the first stage when $\underline{k} \geq 30$, and $\underline{n} \leq 20$, is appropriate. When the first-stage-test α was lowered from 0.05 to 0.025, the false rejection rates for the second-stage $\underline{z}$ tests were slightly decreased (for δ, or $\mu_\delta = 0$), and the total power of these $\underline{z}$ tests increased (for δ, or $\mu_\delta \neq 0$).

One can also consider other approaches such as categorizing the data into homogeneous subgroups instead of using the random-effects test after rejection of homogeneity, until more is learned about the estimate of the variance of the population effects.

CHAPTER VI

CONCLUSIONS AND IMPLICATIONS

This Chapter includes six sections. First I give an example with empirical data to illustrate how power of the homogeneity test can be useful to integrative reviewers. Second I summarize the simulation study. Then I discuss the results of the simulation, including the power of the homogeneity test, and the power of the sequential $\underline{z}$ testing procedure. Fifth, I present some practical implications for integrative reviews. And finally, I make suggestions for further research related to effect-size meta-analysis.

Example

The theoretical power of the homogeneity test was computed for a subset of data originally from the published reviews by Steinkamp and Maehr (1983, 1984) and reanalyzed by Becker (1989). Five studies with six samples on gender and Geology achievement were chosen. Power was computed for two sets of fixed-effects population effects: (0, 0, 0, 0, 0, 0.5), and (0, 0, 0.2, 0.2, 0.4, 0.4). The number of effects was $\underline{k} = 6$, and the sample sizes, conditional variances of effects $\sigma^2(\underline{d}_i|\delta_i)$, and noncentrality parameter λ. for the noncentral chi-square are listed in Tables 60 and 61.

Table 60

Computation of Noncentrality Parameter for the One-Extreme-Value Example

n^M	n^F	δ_i	$(\delta_i - \delta.)^2$	$\sigma^2(d_i \mid \delta_i)$	$(\delta_i - \delta.)^2/\sigma^2(d_i \mid \delta_i)$
52	54	0	.00694	.0378	.1839
46	47	0	.00694	.0430	.1614
458	430	0	.00694	.0045	1.5397
47	47	0	.00694	.0426	.1632
64	56	0	.00694	.0335	.2074
48	48	0.5	.24174	.0430	5.6258
					$\lambda. = 7.8814$

Table 61

Computation of Noncentrality Parameter for the Three-Equal-Values Example

n^M	n^F	δ_i	$(\delta_i - \delta.)^2$	$\sigma^2(d_i \mid \delta_i)$	$(\delta_i - \delta.)^2/\sigma^2(d_i \mid \delta_i)$
52	54	0	.04	.0378	1.0596
46	47	0	.04	.0430	.9298
458	430	0.2	.00	.0045	.0000
47	47	0.2	.00	.0428	.0000
64	56	0.4	.04	.0341	1.1714
48	48	0.4	.04	.0425	.9412
					$\lambda. = 4.1020$

For the given samples, power to detect the "true" heterogeneity for population effects including only one distinct value of 0.5 was about 0.55 (λ. = 7.8814, df = 5). With the given set of samples, the homogeneity test can detect true differences (with the single distinct value being 0.5) more than half of the time. Power decreases as the one extreme value decreases. In other words, if the extreme value was less than 0.5, the homogeneity test would be less likely to reject the homogeneity of effects.

Power for population effects with three equal values (with an average of 0.2) was about 0.42 (λ. 4.1020, df = 5). With the given set of data, homogeneity would be rejected slightly less than half of the time. Again, when the values of effects decrease or increase, the power of the homogeneity test will decrease or increase accordingly.

The homogeneity test is also sensitive to the dispersion of effects. Even though the mean effect of the three-equal-values set (0.2) was greater than the mean effect of the one-extreme-value set (0.0833), power of the homogeneity was higher for the sets of effects that contained one extreme values.

Summary

Effect-size meta-analysis has enabled research syntheses to become quantitatively more precise through analyses of standardized effect sizes from primary studies.

Hedges & Olkin (1985) present both an unbiased estimator of effect size and a homogeneity test for effect-size data. They recommend examining the consistency of the effect sizes before applying any test for the magnitude of the common or average effect across studies. In this research, I derived an approximate distribution for the homogeneity test under alternative models, and then studied the power of the homogeneity test through numerical simulation. I also explored the impact of decisions about homogeneity of effect sizes on subsequent tests of effect magnitude. Suggestions were made to assist meta-analysts in maintaining desirable statistical error rates.

The Power of the Homogeneity Test

The H statistic or homogeneity test had an asymptotic central chi-squared distribution when effect sizes were homogeneous, that is, under the null hypothesis. In the fixed-effects case, when alternative hypotheses were true, the distribution of the H statistic was well approximated by a noncentral chi-squared distribution. These theoretical distributions fit quite well with the simulated distributions for effect sizes based on large samples. The asymptotic distributions tended to underestimate power when some effects had extreme values or when large numbers of effects were based on small samples (e.g., total within-study sample sizes of $n_i = 20$).

When effects are homogeneous, the power of $\underline{H}$ should equal the α level or size of the test. In most cases the nominal and simulated significance levels were quite close. However, simulation data indicated that for a nominal α level of 0.05, the proportion of false rejections approached 0.10 for situations in which $\underline{k} = 30$ and $\underline{n_i} = 20$. Simulated significance levels were close to the nominal α level when sample sizes were larger ($\underline{n_i} \geq 60$). When encountering many studies (for $\underline{k} \geq 30$) all or many of which have small samples (e.g., $\underline{n_i} \leq 20$) meta-analysts may wish to lower the nominal α level of the homogeneity test to 0.025 to achieve an actual α nearer to 0.05.

In the random-effects case, under alternative hypotheses, the distribution of $\underline{H}$ could not be presented in a simple form. The nonnull distribution of $\underline{H}$ is a combination of many noncentral chi-squared distributions. Theoretical power values based on the combination of noncentral chi-squares corresponded closely to the simulated values for the random-effects case.

The Power of the z Tests

Based on the particular decision about homogeneity from the $\underline{H}$ test, a "second-stage" $\underline{z}$ test of effect magnitude can be calculated. If homogeneity is accepted, the estimate of the fixed-effects within-study variance is applied in the $\underline{z}_F$ test. When homogeneity is rejected, the estimate of the

random-effects variance would be used to compute z_R. The power functions of z_F and z_R were examined in this dissertation. In general, the theoretical power values were lower than the simulated values for the fixed-effects tests, and higher for the random-effects tests.

Power values were also compared for z tests calculated with the fixed-effects variance (z_F) versus tests with the random-effects variance (z_R), i.e., tests calculated in the presence of a statistical error at stage one of testing. Power values were always higher for the fixed-effects tests (z_F) than for the random-effects tests (z_R) in these cases. When homogeneity was falsely accepted, the more powerful fixed-effects tests would be applied. When homogeneity was falsely rejected, the much less powerful random-effects tests would be applied.

To prevent the z_R test from having excessively low power for homogeneous effects, the Type I error rate (the rate of false rejection) of the homogeneity test should be limited. This recommendation is consistent with the recommendation based on the simulation study of the homogeneity tests above. In order to maintain, if not to reduce, the rate of false rejection, the α level of 0.05 for the homogeneity test may be lowered for effect sizes based on many small samples.

Practical Implications

The study of the power of the homogeneity test and the power of the subsequent *z* test was useful theoretically in understanding the distributions of both statistics. Practically, these distributions enable reviewers to estimate the power of the homogeneity tests and to adjust for possible inflation of statistical errors. Studying the sequential process in meta-analysis gives a sense of the impact of the first-stage homogeneity test on the second-stage *z* test.

Simulation results showed that when many studies have small samples homogeneity tests were likely to be falsely rejected and thus cause the subsequent *z* test to lack power. Classifying effects into homogenous subgroups, or applying more complicated linear models are alternative approaches in which the reviewer explains variation among the effects. Meta-analysts were advised to adjust the significance level of the homogeneity test. However, a more general suggestion to researchers should be to include more subjects (i.e., large samples) in primary studies. It is always better to integrate studies of higher quality or with stronger evidence.

Suggestions for Further Research

More needs to be learned about the estimator of the population variance component, which figures in random-

effects *z* tests. The estimator proposed by Hedges & Olkin (1985) had an asymptotic normal distribution but the small-sample behavior of the estimator is unexplored. The variance of the estimator as well as the behavior of the estimator for different numbers of studies or sample sizes should be further studied.

APPENDICES

APPENDIX A

CHOOSING THE NUMBER OF REPLICATIONS FOR SIMULATION

Simulated power values are measured by the proportion of replications. We want to be able to draw a 95% confidence intervals for these proportions. With $\underline{R}$ replications, the proportions are approximately normally distributed with an expected value π, and a variance of $\pi(1-\pi)/\underline{R}$. We can write:

$$\underline{p} \sim N\left(\pi, \frac{\pi(1-\pi)}{\underline{R}}\right).$$

Let $\pi = .95$, and let the desired 95% confidence interval for the proportion be $\underline{p} \pm .01$. That is,

$$(1.96)\sqrt{\frac{.95\ (1-.95)}{\underline{R}}} = .01$$

The solution of this equation gives $\underline{R} = 1827$. Thus, I choose $\underline{R} = 2000$ as the number of replications for the simulation.

APPENDIX B

Table 62

Values of Sample Sizes Used in Simulation Study						
k	π	ϕ	N =	$20k$	$60k$	$120k$	$200k$
2	1	.5	n_1 =	10, 10	30, 30	60, 60	100, 100
			n_2 =	10, 10	30, 30	60, 60	100, 100
		.35		7, 13	20, 40	42, 78	70, 130
				7, 13	20, 40	42, 78	70, 130
	2	.5		6, 6	18, 18	36, 36	60, 60
				14, 14	42, 42	84, 84	140, 140
		.35		4, 8	12, 24	24, 48	40, 80
				10, 18	30, 54	60, 108	100, 180
5	1	.5	n_1 =	10, 10	30, 30	60, 60	100, 100
			n_2 =	10, 10	30, 30	60, 60	100, 100
			n_3 =	10, 10	30, 30	60, 60	100, 100
			n_4 =	10, 10	30, 30	60, 60	100, 100
			n_5 =	10, 10	30, 30	60, 60	100, 100
		.35		7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
	2	.5		7, 8	22, 23	45, 45	75, 75
				10, 10	30, 30	60, 60	100, 100
				10, 10	30, 30	60, 60	100, 100
				10, 10	30, 30	60, 60	100, 100
				12, 13	37, 38	75, 75	125, 125
		.35		4, 11	15, 30	32, 58	52, 98
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				9, 16	26, 49	52, 98	87, 163

Table 62 --- Continued

Values of Sample Sizes Used in Simulation Study

k	π	φ	N =	20k	60k	120k	200k
10	1	.5	n_1 =	10, 10	30, 30	60, 60	100, 100
			n_2 =	10, 10	30, 30	60, 60	100, 100
			n_3 =	10, 10	30, 30	60, 60	100, 100
			n_4 =	10, 10	30, 30	60, 60	100, 100
			n_5 =	10, 10	30, 30	60, 60	100, 100
			n_6 =	10, 10	30, 30	60, 60	100, 100
			n_7 =	10, 10	30, 30	60, 60	100, 100
			n_8 =	10, 10	30, 30	60, 60	100, 100
			n_9 =	10, 10	30, 30	60, 60	100, 100
			n_{10}=	10, 10	30, 30	60, 60	100, 100
		.35		7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
				7, 13	21, 39	42, 78	70, 130
	2	.5		5, 5	15, 15	30, 30	50, 50
				6, 6	18, 18	36, 36	60, 60
				7, 7	21, 21	42, 42	70, 70
				7, 7	21, 21	42, 42	70, 70
				8, 8	24, 24	48, 48	80, 80
				8, 8	24, 24	48, 48	80, 80
				9, 9	27, 27	54, 54	90, 90
				10, 10	30, 30	60, 60	100, 100
				15, 15	45, 45	90, 90	150, 150
				25, 25	75, 75	150, 150	250, 250
		.35		3, 7	10, 20	21, 39	35, 65
				4, 8	13, 23	25, 47	42, 78
				5, 9	15, 27	29, 55	49, 91
				5, 9	15, 27	29, 55	49, 91
				6, 10	17, 31	34, 62	56, 104
				6, 10	17, 31	34, 62	56, 104
				6, 12	19, 35	38, 70	63, 117
				7, 13	21, 39	42, 78	70, 130
				11, 19	31, 59	63, 117	105, 195
				17, 33	52, 98	105, 195	175, 325

Table 62 --- Continued

Values of Sample Sizes Used in Simulation Study						
k	π	ϕ	N =	$20k$	$60k$	$120k$	$200k$
30	1	.5	n_1 =	10, 10	30, 30	60, 60	100, 100
			n_2 =	10, 10	30, 30	60, 60	100, 100
			n_3 =	10, 10	30, 30	60, 60	100, 100
			n_4 =	10, 10	30, 30	60, 60	100, 100
			n_5 =	10, 10	30, 30	60, 60	100, 100
			n_6 =	10, 10	30, 30	60, 60	100, 100
			n_7 =	10, 10	30, 30	60, 60	100, 100
			n_8 =	10, 10	30, 30	60, 60	100, 100
			n_9 =	10, 10	30, 30	60, 60	100, 100
			n_{10} =	10, 10	30, 30	60, 60	100, 100
			n_{11} =	10, 10	30, 30	60, 60	100, 100
			n_{12} =	10, 10	30, 30	60, 60	100, 100
			n_{13} =	10, 10	30, 30	60, 60	100, 100
			n_{14} =	10, 10	30, 30	60, 60	100, 100
			n_{15} =	10, 10	30, 30	60, 60	100, 100
			n_{16} =	10, 10	30, 30	60, 60	100, 100
			n_{17} =	10, 10	30, 30	60, 60	100, 100
			n_{18} =	10, 10	30, 30	60, 60	100, 100
			n_{19} =	10, 10	30, 30	60, 60	100, 100
			n_{20} =	10, 10	30, 30	60, 60	100, 100
			n_{21} =	10, 10	30, 30	60, 60	100, 100
			n_{22} =	10, 10	30, 30	60, 60	100, 100
			n_{23} =	10, 10	30, 30	60, 60	100, 100
			n_{24} =	10, 10	30, 30	60, 60	100, 100
			n_{25} =	10, 10	30, 30	60, 60	100, 100
			n_{26} =	10, 10	30, 30	60, 60	100, 100
			n_{27} =	10, 10	30, 30	60, 60	100, 100
			n_{28} =	10, 10	30, 30	60, 60	100, 100
			n_{29} =	10, 10	30, 30	60, 60	100, 100
			n_{30} =	10, 10	30, 30	60, 60	100, 100

Table 62 --- Continued

Values of Sample Sizes Used in Simulation Study						
k	π	ϕ	N =	$20k$	$60k$	$120k$	$200k$
30	1	.35	n_1 =	7, 13	21, 39	42, 78	70, 130
			n_2 =	7, 13	21, 39	42, 78	70, 130
			n_3 =	7, 13	21, 39	42, 78	70, 130
			n_4 =	7, 13	21, 39	42, 78	70, 130
			n_5 =	7, 13	21, 39	42, 78	70, 130
			n_6 =	7, 13	21, 39	42, 78	70, 130
			n_7 =	7, 13	21, 39	42, 78	70, 130
			n_8 =	7, 13	21, 39	42, 78	70, 130
			n_9 =	7, 13	21, 39	42, 78	70, 130
			n_{10} =	7, 13	21, 39	42, 78	70, 130
			n_{11} =	7, 13	21, 39	42, 78	70, 130
			n_{12} =	7, 13	21, 39	42, 78	70, 130
			n_{13} =	7, 13	21, 39	42, 78	70, 130
			n_{14} =	7, 13	21, 39	42, 78	70, 130
			n_{15} =	7, 13	21, 39	42, 78	70, 130
			n_{16} =	7, 13	21, 39	42, 78	70, 130
			n_{17} =	7, 13	21, 39	42, 78	70, 130
			n_{18} =	7, 13	21, 39	42, 78	70, 130
			n_{19} =	7, 13	21, 39	42, 78	70, 130
			n_{20} =	7, 13	21, 39	42, 78	70, 130
			n_{21} =	7, 13	21, 39	42, 78	70, 130
			n_{22} =	7, 13	21, 39	42, 78	70, 130
			n_{23} =	7, 13	21, 39	42, 78	70, 130
			n_{24} =	7, 13	21, 39	42, 78	70, 130
			n_{25} =	7, 13	21, 39	42, 78	70, 130
			n_{26} =	7, 13	21, 39	42, 78	70, 130
			n_{27} =	7, 13	21, 39	42, 78	70, 130
			n_{28} =	7, 13	21, 39	42, 78	70, 130
			n_{29} =	7, 13	21, 39	42, 78	70, 130
			n_{30} =	7, 13	21, 39	42, 78	70, 130

Table 62 --- Continued

Values of Sample Sizes Used in Simulation Study						
k	π	ϕ	N =	$20k$	$60k$	$120k$	$200k$
30	2	.5	n_1 =	2, 2	6, 6	12, 12	20, 20
			n_2 =	3, 3	9, 9	18, 18	30, 30
			n_3 =	3, 3	9, 9	18, 18	30, 30
			n_4 =	3, 3	9, 9	18, 18	30, 30
			n_5 =	4, 4	12, 12	24, 24	40, 40
			n_6 =	6, 6	18, 18	36, 36	60, 60
			n_7 =	6, 6	18, 18	36, 36	60, 60
			n_8 =	6, 6	18, 18	36, 36	60, 60
			n_9 =	6, 6	18, 18	36, 36	60, 60
			n_{10} =	6, 6	18, 18	36, 36	60, 60
			n_{11} =	6, 6	18, 18	36, 36	60, 60
			n_{12} =	7, 7	21, 21	42, 42	70, 70
			n_{13} =	7, 7	21, 21	42, 42	70, 70
			n_{14} =	7, 7	21, 21	42, 42	70, 70
			n_{15} =	8, 8	24, 24	48, 48	80, 80
			n_{16} =	8, 8	24, 24	48, 48	80, 80
			n_{17} =	8, 8	24, 24	48, 48	80, 80
			n_{18} =	8, 8	24, 24	48, 48	80, 80
			n_{19} =	11, 11	33, 33	66, 66	110, 110
			n_{20} =	11, 11	33, 33	66, 66	110, 110
			n_{21} =	11, 11	33, 33	66, 66	110, 110
			n_{22} =	12, 12	36, 36	72, 72	120, 120
			n_{23} =	12, 12	36, 36	72, 72	120, 120
			n_{24} =	14, 14	42, 42	84, 84	140, 140
			n_{25} =	17, 17	51, 51	102, 102	170, 170
			n_{26} =	17, 17	51, 51	102, 102	170, 170
			n_{27} =	17, 17	51, 51	102, 102	170, 170
			n_{28} =	20, 20	60, 60	120, 120	200, 200
			n_{29} =	20, 20	60, 60	120, 120	200, 200
			n_{30} =	34, 34	102, 102	204, 204	340, 340

Table 62 --- Continued

Values of Sample Sizes Used in Simulation Study						
k	π	ϕ	N =	$20k$	$60k$	$120k$	$200k$
30	2	.35	n_1 =	2, 2	4, 8	8, 16	14, 26
			n_2 =	2, 4	6, 12	13, 23	21, 39
			n_3 =	2, 4	6, 12	13, 23	21, 39
			n_4 =	2, 4	6, 12	13, 23	21, 39
			n_5 =	3, 5	8, 16	17, 31	28, 52
			n_6 =	4, 8	13, 23	25, 47	42, 78
			n_7 =	4, 8	13, 23	25, 47	42, 78
			n_8 =	4, 8	13, 23	25, 47	42, 78
			n_9 =	4, 8	13, 23	25, 47	42, 78
			n_{10} =	4, 8	13, 23	25, 47	42, 78
			n_{11} =	4, 8	13, 23	25, 47	42, 78
			n_{12} =	5, 9	15, 27	29, 55	49, 91
			n_{13} =	5, 9	15, 27	29, 55	49, 91
			n_{14} =	5, 9	15, 27	29, 55	49, 91
			n_{15} =	6, 10	17, 31	33, 63	56, 104
			n_{16} =	6, 10	17, 31	33, 63	56, 104
			n_{17} =	6, 10	17, 31	33, 63	56, 104
			n_{18} =	6, 10	17, 31	33, 63	56, 104
			n_{19} =	8, 14	23, 43	46, 86	77, 143
			n_{20} =	8, 14	23, 43	46, 86	77, 143
			n_{21} =	8, 14	23, 43	46, 86	77, 143
			n_{22} =	8, 16	25, 47	50, 94	84, 156
			n_{23} =	8, 16	25, 47	50, 94	84, 156
			n_{24} =	10, 18	29, 55	59, 109	98, 182
			n_{25} =	12, 22	36, 66	71, 133	119, 221
			n_{26} =	12, 22	36, 66	71, 133	119, 221
			n_{27} =	12, 22	36, 66	71, 133	119, 221
			n_{28} =	14, 26	42, 78	84, 156	140, 260
			n_{29} =	14, 26	42, 78	84, 156	140, 260
			n_{30} =	24, 44	71, 133	143, 265	238, 422

Table 63

Values of δ_is Used in the Simulation for $\underline{k}$ = 2						
set =	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	.1	.25	.5	.75	1

Table 64

Values of δ_is Used in the Simulation for $\underline{k}$ = 5												
set =	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	.1	.2	.25	.3	.4	.5
3	0	0	0	0	0	0	.1	.2	.25	.3	.4	.5
4	0	0	0	0	0	0	.2	.4	.5	.6	.8	1
5	0	.1	.25	.5	.75	1	.2	.4	.5	.6	.8	1

Table 65

Values of δ_is Used in the Simulation for $\underline{k}$ = 10												
set =	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	.1
5	0	0	0	0	0	0	0	0	0	0	0	.1
6	0	0	0	0	0	0	0	0	0	0	0	.1
7	0	0	0	0	0	0	0	0	0	0	0	.1
8	0	0	0	0	0	0	0	0	0	0	0	.2
9	0	0	0	0	0	0	.1	.25	.5	.75	1	.2
10	0	.1	.25	.5	.75	1	.1	.25	.5	.75	1	.2

Table 65 --- Continued

Values of δ_is Used in the Simulation for $\underline{k}$ = 10										
set =	13	14	15	16	17	18	19	20	21	22
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	.05	.1	.15	.2	.25
4	.2	.25	.3	.4	.5	.05	.1	.15	.2	.25
5	.2	.25	.3	.4	.5	.1	.2	.3	.4	.5
6	.2	.25	.3	.4	.5	.1	.2	.3	.4	.5
7	.2	.25	.3	.4	.5	.15	.3	.45	.6	.75
8	.4	.5	.6	.8	1	.15	.3	.45	.6	.75
9	.4	.5	.6	.8	1	.2	.4	.6	.8	1
10	.4	.5	.6	.8	1	.2	.4	.6	.8	1

Table 66

Values of δ_is Used in the Simulation for $\underline{k}$ = 30											
set =	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	.1	.25	.5	.75	1
30	0	.1	.25	.5	.75	1	.1	.25	.5	.75	1

Table 66 --- Continued

Values of δ_is Used in the Simulation for $\underline{k}$ = 30											
set =	12	13	14	15	16	17	18	19	20	21	22
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	.05	.1	.15	.2	.25
8	0	0	0	0	0	0	.05	.1	.15	.2	.25
9	0	0	0	0	0	0	.05	.1	.15	.2	.25
10	0	0	0	0	0	0	.05	.1	.15	.2	.25
11	.1	.2	.25	.3	.4	.5	.05	.1	.15	.2	.25
12	.1	.2	.25	.3	.4	.5	.05	.1	.15	.2	.25
13	.1	.2	.25	.3	.4	.5	.1	.2	.3	.4	.5
14	.1	.2	.25	.3	.4	.5	.1	.2	.3	.4	.5
15	.1	.2	.25	.3	.4	.5	.1	.2	.3	.4	.5
16	.1	.2	.25	.3	.4	.5	.1	.2	.3	.4	.5
17	.1	.2	.25	.3	.4	.5	.1	.2	.3	.4	.5
18	.1	.2	.25	.3	.4	.5	.1	.2	.3	.4	.5
19	.1	.2	.25	.3	.4	.5	.15	.3	.45	.6	.75
20	.1	.2	.25	.3	.4	.5	.15	.3	.45	.6	.75
21	.2	.4	.5	.6	.8	1	.15	.3	.45	.6	.75
22	.2	.4	.5	.6	.8	1	.15	.3	.45	.6	.75
23	.2	.4	.5	.6	.8	1	.15	.3	.45	.6	.75
24	.2	.4	.5	.6	.8	1	.15	.3	.45	.6	.75
25	.2	.4	.5	.6	.8	1	.2	.45	.6	.8	1
26	.2	.4	.5	.6	.8	1	.2	.45	.6	.8	1
27	.2	.4	.5	.6	.8	1	.2	.45	.6	.8	1
28	.2	.4	.5	.6	.8	1	.2	.45	.6	.8	1
29	.2	.4	.5	.6	.8	1	.2	.45	.6	.8	1
30	.2	.4	.5	.6	.8	1	.2	.45	.6	.8	1

APPENDIX C

Table 67

Means of Power of H for δ_is with One Extreme Value by N and k (α = 0.10)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
2(20)	.104(4)	.123(4)	.190(4)	.294(4)	.421(4)	.2265(20)
2(60)	.111(4)	.167(4)	.355(4)	.596(4)	.800(4)	.4062(20)
2(120)	.122(4)	.235(4)	.561(4)	.847(4)	.969(4)	.5470(20)
2(200)	.137(4)	.319(4)	.750(4)	.963(4)	.998(4)	.6333(20)
5(20)	.103(4)	.120(4)	.183(4)	.289(4)	.426(4)	.2245(20)
5(60)	.110(4)	.163(4)	.361(4)	.639(4)	.857(4)	.4260(20)
5(120)	.120(4)	.230(4)	.612(4)	.906(4)	.990(4)	.5694(20)
5(200)	.133(4)	.321(4)	.815(4)	.989(4)	1.000(4)	.6515(20)
10(20)	.103(4)	.121(4)	.189(4)	.308(4)	.462(4)	.2365(20)
10(60)	.110(4)	.166(4)	.391(4)	.681(4)	.867(4)	.4431(20)
10(120)	.120(4)	.241(4)	.646(4)	.908(4)	.986(4)	.5803(20)
10(200)	.134(4)	.346(4)	.835(4)	.985(4)	1.000(4)	.6598(20)
30(20)	.102(4)	.116(4)	.169(4)	.270(4)	.406(4)	.2128(20)
30(60)	.107(4)	.151(4)	.344(4)	.605(4)	.773(4)	.3960(20)
30(120)	.115(4)	.213(4)	.575(4)	.817(4)	.944(4)	.5327(20)
30(200)	.126(4)	.302(4)	.743(4)	.940(4)	.996(4)	.6213(20)

Note: The pattern of δ_i values with one extreme value was $(0, \ldots, 0, \delta)$.

Table 67.a

Means of Simulated Power of H for δ_is with One Extreme Value by N and k (α = 0.10)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
2(20)	.103(4)	.131(4)	.183(4)	.283(4)	.415(4)	.2228(20)
2(60)	.109(4)	.163(4)	.350(4)	.597(4)	.800(4)	.4038(20)
2(120)	.123(4)	.233(4)	.560(4)	.849(4)	.974(4)	.5475(20)
2(200)	.141(4)	.320(4)	.745(4)	.966(4)	.996(4)	.6334(20)
5(20)	.118(4)	.130(4)	.196(4)	.302(4)	.447(4)	.2386(20)
5(60)	.113(4)	.166(4)	.362(4)	.656(4)	.871(4)	.4337(20)
5(120)	.120(4)	.235(4)	.602(4)	.913(4)	.993(4)	.5726(20)
5(200)	.123(4)	.319(4)	.830(4)	.990(4)	1.000(4)	.6524(20)
10(20)	.133(4)	.147(4)	.220(4)	.336(4)	.502(4)	.2677(20)
10(60)	.117(4)	.177(4)	.401(4)	.700(4)	.883(4)	.4556(20)
10(120)	.124(4)	.241(4)	.656(4)	.913(4)	.992(4)	.5853(20)
10(200)	.135(4)	.344(4)	.848(4)	.987(4)	1.000(4)	.6626(20)
30(20)	.168(4)	.173(4)	.229(4)	.328(4)	.466(4)	.2727(20)
30(60)	.121(4)	.166(4)	.369(4)	.625(4)	.801(4)	.4163(20)
30(120)	.116(4)	.226(4)	.595(4)	.831(4)	.961(4)	.5457(20)
30(200)	.131(4)	.312(4)	.756(4)	.951(4)	1.000(4)	.6297(20)

Note: The pattern of δ_i values with one extreme value was $(0, \ldots, 0, \delta)$.

Table 68

Means of Power of H for δ_is with Two Extreme Values by N and k (α = 0.10)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
10(20)	.105(4)	.130(4)	.232(4)	.410(4)	.510(4)	.2776(20)
10(60)	.114(4)	.198(4)	.524(4)	.853(4)	.976(4)	.5330(20)
10(120)	.129(4)	.310(4)	.819(4)	.989(4)	1.000(4)	.6495(20)
10(200)	.150(4)	.460(4)	.960(4)	1.000(4)	1.000(4)	.7141(20)
30(20)	.104(4)	.125(4)	.216(4)	.384(4)	.580(4)	.2819(20)
30(60)	.112(4)	.185(4)	.495(4)	.796(4)	.937(4)	.5049(20)
30(120)	.124(4)	.289(4)	.766(4)	.965(4)	.999(4)	.6288(20)
30(200)	.142(4)	.434(4)	.915(4)	.998(4)	1.000(4)	.6979(20)

Note: The pattern of δ_i values with two extreme values was $(0, \ldots, 0, \delta, \delta)$.

Table 68.a

Means of Simulated Power of H for δ_is with Two Extreme Values by N and k (α = 0.10)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
10(20)	.131(4)	.163(4)	.254(4)	.432(4)	.628(4)	.3015(20)
10(60)	.120(4)	.200(4)	.535(4)	.864(4)	.982(4)	.5398(20)
10(120)	.138(4)	.317(4)	.827(4)	.992(4)	1.000(4)	.6547(20)
10(200)	.156(4)	.472(4)	.956(4)	1.000(4)	1.000(4)	.7166(20)
30(20)	.164(4)	.185(4)	.274(4)	.435(4)	.639(4)	.3396(20)
30(60)	.129(4)	.199(4)	.516(4)	.810(4)	.959(4)	.5228(20)
30(120)	.133(4)	.302(4)	.775(4)	.973(4)	.999(4)	.6364(20)
30(200)	.138(4)	.424(4)	.921(4)	.998(4)	1.000(4)	.8957(20)

Note: The pattern of δ_i values with two extreme values was $(0, \ldots, 0, \delta, \delta)$.

Table 69

Means of Power of H for Three Equal Subsets of δ_is by N and k (α = 0.10)

k*n	δ = 0.10	0.20	0.25	0.30	0.40	0.50	Total
5(20)	.110	.140	.163	.191	.265	.356	.2042(24)
5(60)	.130	.227	.301	.391	.590	.772	.4019(24)
5(120)	.162	.362	.503	.648	.873	.971	.5864(24)
5(200)	.206	.529	.713	.855	.980	.999	.7135(24)
10(20)	.114	.159	.195	.240	.360	.506	.2622(24)
10(60)	.144	.296	.416	.553	.805	.946	.5264(24)
10(120)	.192	.508	.704	.858	.985	.999	.7076(24)
10(200)	.261	.736	.909	.980	1.000	1.000	.8143(24)
30(20)	.123	.206	.277	.369	.595	.805	.3957(24)
30(60)	.177	.477	.684	.853	.987	1.000	.6962(24)
30(120)	.271	.806	.956	.995	1.000	1.000	.8378(24)
30(200)	.410	.968	.999	1.000	1.000	1.000	.8960(24)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots, 0, \delta,\ldots, \delta, 2\delta,\ldots, 2\delta)$.

Table 69.a

Means of Simulated Power of H for Three Equal Subsets of δ_is by N and k (α = 0.10)

k*n	δ = 0.10	0.20	0.25	0.30	0.40	0.50	Total
5(20)	.122	.160	.173	.216	.273	.372	.2193(24)
5(60)	.134	.231	.306	.400	.584	.773	.4046(24)
5(120)	.164	.365	.511	.649	.874	.972	.5890(24)
5(200)	.206	.528	.710	.856	.981	.999	.7133(24)
10(20)	.137	.192	.217	.260	.372	.524	.2837(24)
10(60)	.144	.307	.425	.557	.801	.948	.5304(24)
10(120)	.195	.503	.709	.853	.987	1.000	.7077(24)
10(200)	.250	.740	.905	.980	1.000	1.000	.8125(24)
30(20)	.190	.272	.349	.426	.625	.819	.4468(24)
30(60)	.190	.487	.689	.851	.987	1.000	.7007(24)
30(120)	.269	.802	.959	.992	1.000	1.000	.8370(24)
30(200)	.407	.969	.998	1.000	1.000	1.000	.8957(24)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots, 0, \delta,\ldots, \delta, 2\delta,\ldots, 2\delta)$.

Table 70

Means of Power of H for Five Equal Subsets of δ_is by N and k (α = 0.10)

k*n	½δ = 0.10	0.20	0.30	0.40	0.50	Total
10(20)	.112	.149	.216	.316	.442	.2470(20)
10(60)	.136	.262	.484	.730	.902	.5029(20)
10(120)	.176	.444	.789	.964	.998	.6742(20)
10(200)	.234	.657	.955	1.000	1.000	.7690(20)
30(20)	.117	.177	.294	.469	.669	.3451(20)
30(60)	.156	.374	.723	.944	.996	.6386(20)
30(120)	.223	.669	.970	1.000	1.000	.7722(20)
30(200)	.323	.898	.999	1.000	1.000	.8442(20)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots,0,\frac{1}{2}\delta,\ldots,\frac{1}{2}\delta,\delta,\ldots,\delta,1\frac{1}{2}\delta,\ldots,1\frac{1}{2}\delta,2\delta,\ldots,2\delta)$.

Table 70.a

Means of Simulated Power of H for Five Equal Subsets of δ_is by N and k (α = 0.10)

k*n	½δ = 0.10	0.20	0.30	0.40	0.50	Total
10(20)	.139	.184	.245	.336	.460	.2727(20)
10(60)	.143	.270	.485	.737	.902	.5073(20)
10(120)	.181	.439	.791	.967	.997	.6750(20)
10(200)	.233	.654	.952	.997	1.000	.7675(20)
30(20)	.186	.253	.352	.515	.703	.4018(20)
30(60)	.185	.388	.729	.943	.995	.6481(20)
30(120)	.233	.671	.970	1.000	1.000	.7748(20)
30(200)	.324	.904	1.000	1.000	1.000	.8455(20)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots,0,\frac{1}{2}\delta,\ldots,\frac{1}{2}\delta,\delta,\ldots,\delta,1\frac{1}{2}\delta,\ldots,1\frac{1}{2}\delta,2\delta,\ldots,2\delta)$.

Table 71

Means of Power of H for δ_is with One Extreme Value by N and k (α = 0.025)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
2(20)	.027(4)	.035(4)	.068(4)	.126(4)	.213(4)	.0937(20)
2(60)	.030(4)	.056(4)	.166(4)	.362(4)	.599(4)	.2427(20)
2(120)	.035(4)	.092(4)	.330(4)	.668(4)	.899(4)	.4046(20)
2(200)	.014(4)	.142(4)	.532(4)	.885(4)	.988(4)	.5178(20)
5(20)	.026(4)	.033(4)	.061(4)	.119(4)	.213(4)	.0905(20)
5(60)	.029(4)	.052(4)	.166(4)	.404(4)	.685(4)	.2670(20)
5(120)	.033(4)	.085(4)	.366(4)	.768(4)	.962(4)	.4426(20)
5(200)	.038(4)	.139(4)	.620(4)	.956(4)	.999(4)	.5505(20)
10(20)	.026(4)	.033(4)	.063(4)	.132(4)	.249(4)	.1001(20)
10(60)	.027(4)	.053(4)	.191(4)	.472(4)	.732(4)	.2953(20)
10(120)	.033(4)	.091(4)	.431(4)	.799(4)	.954(4)	.4616(20)
10(200)	.038(4)	.156(4)	.680(4)	.950(4)	.998(4)	.5649(20)
30(20)	.026(4)	.031(4)	.054(4)	.109(4)	.210(4)	.0868(20)
30(60)	.028(4)	.045(4)	.160(4)	.414(4)	.632(4)	.2557(20)
30(120)	.031(4)	.076(4)	.378(4)	.687(4)	.863(4)	.4070(20)
30(200)	.035(4)	.130(4)	.592(4)	.857(4)	.981(4)	.5190(20)

Note: The pattern of δ_i values with one extreme value was $(0, \ldots, 0, \delta)$.

Table 71.a

Means of Simulated Power of H for δ_is with One Extreme Value by N and k (α = 0.025)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
2(20)	.031(4)	.041(4)	.068(4)	.126(4)	.211(4)	.0953(20)
2(60)	.027(4)	.052(4)	.168(4)	.358(4)	.602(4)	.2412(20)
2(120)	.036(4)	.097(4)	.328(4)	.664(4)	.902(4)	.4054(20)
2(200)	.044(4)	.142(4)	.525(4)	.885(4)	.986(4)	.5162(20)
5(20)	.041(4)	.044(4)	.079(4)	.141(4)	.241(4)	.1091(20)
5(60)	.032(4)	.056(4)	.175(4)	.419(4)	.713(4)	.2790(20)
5(120)	.035(4)	.084(4)	.369(4)	.780(4)	.971(4)	.4476(20)
5(200)	.035(4)	.138(4)	.633(4)	.963(4)	1.000(4)	.5537(20)
10(20)	.047(4)	.055(4)	.094(4)	.162(4)	.291(4)	.1298(20)
10(60)	.034(4)	.061(4)	.203(4)	.498(4)	.763(4)	.3117(20)
10(120)	.031(4)	.097(4)	.440(4)	.809(4)	.970(4)	.4692(20)
10(200)	.039(4)	.160(4)	.695(4)	.961(4)	1.000(4)	.5706(20)
30(20)	.072(4)	.071(4)	.103(4)	.168(4)	.284(4)	.1395(20)
30(60)	.035(4)	.053(4)	.182(4)	.451(4)	.671(4)	.2783(20)
30(120)	.033(4)	.087(4)	.393(4)	.709(4)	.896(4)	.4235(20)
30(200)	.039(4)	.141(4)	.607(4)	.880(4)	.989(4)	.5308(20)

Note: The pattern of δ_i values with one extreme value was (0, ..., 0, δ).

Table 72

Means of Power of H for δ_is with Two Extreme Values by N and k (α = 0.025)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
10(20)	.027(4)	.037(4)	.085(4)	.201(4)	.297(4)	.1294(20)
10(60)	.030(4)	.068(4)	.294(4)	.687(4)	.924(4)	.4006(20)
10(120)	.036(4)	.131(4)	.635(4)	.960(4)	.999(4)	.5521(20)
10(200)	.045(4)	.240(4)	.886(4)	.999(4)	1.000(4)	.6339(20)
30(20)	.026(4)	.035(4)	.076(4)	.185(4)	.367(4)	.1379(20)
30(60)	.029(4)	.061(4)	.280(4)	.637(4)	.852(4)	.3717(20)
30(120)	.034(4)	.119(4)	.596(4)	.904(4)	.993(4)	.5293(20)
30(200)	.041(4)	.225(4)	.814(4)	.992(4)	1.000(4)	.6145(20)

Note: The pattern of δ_i values with two extreme values was $(0, \ldots, 0, \delta, \delta)$.

Table 72.a

Means of Simulated Power of H for δ_is with Two Extreme Values by N and k (α = 0.025)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
10(20)	.047(4)	.064(4)	.107(4)	.237(4)	.336(4)	.1581(20)
10(60)	.036(4)	.075(4)	.312(4)	.707(4)	.940(4)	.4141(20)
10(120)	.039(4)	.130(4)	.643(4)	.967(4)	.999(4)	.5556(20)
10(200)	.046(4)	.249(4)	.883(4)	.999(4)	1.000(4)	.6354(20)
30(20)	.068(4)	.081(4)	.130(4)	.245(4)	.448(4)	.1943(20)
30(60)	.037(4)	.072(4)	.310(4)	.662(4)	.886(4)	.3932(20)
30(120)	.037(4)	.133(4)	.603(4)	.920(4)	.997(4)	.5381(20)
30(200)	.040(4)	.224(4)	.825(4)	.993(4)	1.000(4)	.6164(20)

Note: The pattern of δ_i values with two extreme values was $(0, \ldots, 0, \delta, \delta)$.

Table 73

Means of Power of H for Three Equal Subsets of δ_is by N and k (α = 0.025)

k*n	δ = 0.10	0.20	0.25	0.30	0.40	0.50	Total
5(20)	.029	.041	.052	.065	.105	.162	.0757(24)
5(60)	.037	.084	.127	.186	.354	.560	.2246(24)
5(120)	.051	.166	.274	.413	.709	.906	.4198(24)
5(200)	.072	.297	.486	.681	.931	.993	.5766(24)
10(20)	.030	.049	.066	.089	.162	.274	.1117(24)
10(60)	.042	.121	.202	.315	.603	.846	.3548(24)
10(120)	.064	.276	.472	.682	.944	.997	.5725(24)
10(200)	.101	.511	.771	.930	.998	1.000	.7185(24)
30(20)	.034	.070	.108	.165	.352	.600	.2213(24)
30(60)	.056	.246	.446	.673	.951	.998	.5615(24)
30(120)	.104	.602	.868	.976	1.000	1.000	.7583(24)
30(200)	.194	.898	.992	1.000	1.000	1.000	.8472(24)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots, 0, \delta,\ldots, \delta, 2\delta,\ldots, 2\delta)$.

Table 73.a

Means of Simulated Power of H for Three Equal Subsets of δ_is by N and k (α = 0.025)

k*n	δ = 0.10	0.20	0.25	0.30	0.40	0.50	Total
5(20)	.044	.059	.066	.085	.123	.192	.0947(24)
5(60)	.041	.089	.130	.193	.355	.561	.2282(24)
5(120)	.055	.171	.274	.421	.718	.910	.4248(24)
5(200)	.071	.294	.485	.679	.939	.993	.5770(24)
10(20)	.048	.073	.093	.108	.188	.308	.1362(24)
10(60)	.044	.129	.211	.325	.604	.852	.3606(24)
10(120)	.070	.281	.481	.682	.946	.996	.5761(24)
10(200)	.095	.513	.771	.936	.998	1.000	.7187(24)
30(20)	.084	.130	.172	.241	.407	.644	.2797(24)
30(60)	.065	.260	.465	.679	.948	.997	.5688(24)
30(120)	.111	.595	.869	.973	1.000	1.000	.7580(24)
30(200)	.196	.901	.993	1.000	1.000	1.000	.8482(24)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots, 0, \delta,\ldots, \delta, 2\delta,\ldots, 2\delta)$.

Table 74

Means of Power of H for Five Equal Subsets of δ_is by N and k (α = 0.025)

k*n	½δ = 0.10	0.20	0.30	0.40	0.50	Total
10(20)	.029	.045	.077	.134	.222	.1012(20)
10(60)	.039	.101	.225	.504	.758	.3316(20)
10(120)	.057	.223	.581	.888	.987	.5472(20)
10(200)	.086	.420	.867	.993	1.000	.6731(20)
30(20)	.031	.056	.117	.240	.428	.1746(20)
30(60)	.047	.169	.491	.842	.979	.5057(20)
30(120)	.078	.429	.902	.997	1.000	.6813(20)
30(200)	.135	.750	.996	1.000	1.000	.7764(20)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots,0,\frac{1}{2}\delta,\ldots,\frac{1}{2}\delta,\delta,\ldots,\delta,1\frac{1}{2}\delta,\ldots,1\frac{1}{2}\delta,2\delta,\ldots,2\delta)$.

Table 74.a

Means of Simulated Power of H for Five Equal Subsets of δ_is with N and k (α = 0.025)

k*n	½δ = 0.10	0.20	0.30	0.40	0.50	Total
10(20)	.049	.068	.104	.163	.261	.1289(20)
10(60)	.043	.110	.257	.517	.767	.3387(20)
10(120)	.063	.232	.590	.887	.988	.5519(20)
10(200)	.086	.424	.863	.994	1.000	.6733(20)
30(20)	.082	.113	.181	.306	.484	.2332(20)
30(60)	.057	.178	.502	.846	.977	.5119(20)
30(120)	.083	.437	.903	.998	1.000	.6841(20)
30(200)	.140	.757	.997	1.000	1.000	.7787(20)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots,0,\frac{1}{2}\delta,\ldots,\frac{1}{2}\delta,\delta,\ldots,\delta,1\frac{1}{2}\delta,\ldots,1\frac{1}{2}\delta,2\delta,\ldots,2\delta)$.

Table 75

Means of Power of $\underline{H}$ for δ_is with One Extreme Value by $\underline{N}$ and $\underline{k}$ (α = 0.01)

$\underline{k}$*$\underline{n}$	δ = 0.10	0.25	0.50	0.75	1.00	Total
2(20)	.011(4)	.015(4)	.033(4)	.070(4)	.129(4)	.0517(20)
2(60)	.012(4)	.027(4)	.096(4)	.247(4)	.467(4)	.1699(20)
2(120)	.015(4)	.048(4)	.220(4)	.541(4)	.828(4)	.3303(20)
2(200)	.019(4)	.080(4)	.401(4)	.808(4)	.974(4)	.4563(20)
5(20)	.011(4)	.014(4)	.029(4)	.064(4)	.129(4)	.0495(20)
5(60)	.012(4)	.024(4)	.096(4)	.284(4)	.563(4)	.1956(20)
5(120)	.014(4)	.043(4)	.251(4)	.659(4)	.927(4)	.3788(20)
5(200)	.017(4)	.077(4)	.493(4)	.918(4)	.997(4)	.5005(20)
10(20)	.011(4)	.014(4)	.030(4)	.074(4)	.160(4)	.0577(20)
10(60)	.012(4)	.024(4)	.115(4)	.359(4)	.639(4)	.2299(20)
10(120)	.014(4)	.047(4)	.319(4)	.719(4)	.922(4)	.4041(20)
10(200)	.017(4)	.091(4)	.579(4)	.916(4)	.995(4)	.5196(20)
30(20)	.010(4)	.013(4)	.025(4)	.059(4)	.132(4)	.0478(20)
30(60)	.011(4)	.020(4)	.094(4)	.317(4)	.555(4)	.1994(20)
30(120)	.013(4)	.038(4)	.281(4)	.616(4)	.804(4)	.3502(20)
30(200)	.015(4)	.073(4)	.509(4)	.797(4)	.964(4)	.4713(20)

Note: The pattern of δ_i values with one extreme value was (0, ..., 0, δ).

Table 75.a

Means of Simulated Power of H for δ_is with One Extreme Value by N and k (α = 0.01)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
2(20)	.014(4)	.020(4)	.036(4)	.074(4)	.133(4)	.0553(20)
2(60)	.011(4)	.025(4)	.101(4)	.242(4)	.474(4)	.1704(20)
2(120)	.016(4)	.051(4)	.217(4)	.537(4)	.835(4)	.3313(20)
2(200)	.020(4)	.077(4)	.391(4)	.807(4)	.974(4)	.4537(20)
5(20)	.021(4)	.022(4)	.043(4)	.082(4)	.162(4)	.0667(20)
5(60)	.016(4)	.026(4)	.106(4)	.301(4)	.606(4)	.2111(20)
5(120)	.017(4)	.047(4)	.253(4)	.673(4)	.943(4)	.3865(20)
5(200)	.017(4)	.076(4)	.516(4)	.934(4)	.999(4)	.5082(20)
10(20)	.027(4)	.027(4)	.052(4)	.102(4)	.205(4)	.0825(20)
10(60)	.013(4)	.030(4)	.129(4)	.393(4)	.679(4)	.2487(20)
10(120)	.012(4)	.051(4)	.331(4)	.740(4)	.940(4)	.4148(20)
10(200)	.018(4)	.093(4)	.595(4)	.928(4)	.997(4)	.5261(20)
30(20)	.044(4)	.043(4)	.064(4)	.109(4)	.200(4)	.0919(20)
30(60)	.016(4)	.025(4)	.115(4)	.356(4)	.598(4)	.2219(20)
30(120)	.014(4)	.046(4)	.291(4)	.634(4)	.844(4)	.3658(20)
30(200)	.017(4)	.080(4)	.525(4)	.825(4)	.978(4)	.4849(20)

Note: The pattern of δ_i values with one extreme value was (0, ..., 0, δ).

Table 76

Means of Power of H for δ_is with Two Extreme Values by N and k (α = 0.01)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
10(20)	.011(4)	.016(4)	.043(4)	.121(4)	.202(4)	.0785(20)
10(60)	.013(4)	.033(4)	.193(4)	.572(4)	.873(4)	.3366(20)
10(120)	.016(4)	.072(4)	.514(4)	.927(4)	.997(4)	.5053(20)
10(200)	.020(4)	.150(4)	.819(4)	.996(4)	1.000(4)	.5971(20)
30(20)	.011(4)	.015(4)	.038(4)	.111(4)	.263(4)	.0875(20)
30(60)	.012(4)	.029(4)	.186(4)	.542(4)	.788(4)	.3115(20)
30(120)	.014(4)	.065(4)	.497(4)	.852(4)	.985(4)	.4826(20)
30(200)	.018(4)	.142(4)	.744(4)	.983(4)	1.000(4)	.5773(20)

Note: The pattern of δ_i values with two extreme values was $(0, \ldots, 0, \delta, \delta)$.

Table 76.a

Means of Simulated Power of H for δ_is with Two Extreme Values by N and k (α = 0.01)

k*n	δ = 0.10	0.25	0.50	0.75	1.00	Total
10(20)	.025(4)	.034(4)	.063(4)	.156(4)	.243(4)	.1041(20)
10(60)	.015(4)	.037(4)	.206(4)	.595(4)	.899(4)	.3505(20)
10(120)	.019(4)	.078(4)	.528(4)	.936(4)	.998(4)	.5119(20)
10(200)	.022(4)	.161(4)	.819(4)	.997(4)	1.000(4)	.5997(20)
30(20)	.042(4)	.048(4)	.083(4)	.166(4)	.353(4)	.1383(20)
30(60)	.017(4)	.036(4)	.212(4)	.579(4)	.835(4)	.3358(20)
30(120)	.016(4)	.072(4)	.507(4)	.876(4)	.992(4)	.4926(20)
30(200)	.021(4)	.141(4)	.751(4)	.995(4)	1.000(4)	.5796(20)

Note: The pattern of δ_i values with two extreme values was $(0, \ldots, 0, \delta, \delta)$.

Table 77

Means of Power of H for Three Equal Subsets of δ_is by N and k (α = 0.01)

k*n	δ = 0.10	0.20	0.25	0.30	0.40	0.50	Total
5(20)	.012	.018	.024	.032	.055	.093	.0390(24)
5(60)	.016	.042	.069	.110	.241	.430	.1514(24)
5(120)	.024	.096	.175	.291	.589	.840	.3359(24)
5(200)	.036	.193	.358	.558	.877	.985	.5011(24)
10(20)	.013	.022	.032	.045	.092	.175	.0631(24)
10(60)	.019	.065	.120	.208	.474	.757	.2737(24)
10(120)	.031	.176	.345	.559	.898	.991	.5000(24)
10(200)	.052	.382	.663	.876	.996	1.000	.6613(24)
30(20)	.014	.033	.056	.094	.237	.471	.1508(24)
30(60)	.026	.152	.320	.549	.910	.995	.4917(24)
30(120)	.054	.472	.783	.953	1.000	1.000	.7110(24)
30(200)	.114	.830	.982	.999	1.000	1.000	.8208(24)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots, 0, \delta,\ldots, \delta, 2\delta,\ldots, 2\delta)$.

Table 77.a

Means of Simulated Power of H for Three Equal Subsets of δ_is by N and k (α = 0.01)

k*n	δ = 0.10	0.20	0.25	0.30	0.40	0.50	Total
5(20)	.023	.031	.034	.045	.073	.123	.0547(24)
5(60)	.020	.043	.069	.120	.242	.434	.1554(24)
5(120)	.028	.100	.175	.296	.598	.845	.3404(24)
5(200)	.035	.192	.360	.561	.883	.983	.5020(24)
10(20)	.025	.038	.052	.066	.120	.215	.0857(24)
10(60)	.022	.070	.129	.220	.480	.766	.2811(24)
10(120)	.038	.186	.349	.559	.901	.992	.5043(24)
10(200)	.050	.381	.665	.883	.995	1.000	.6624(24)
30(20)	.053	.081	.112	.162	.301	.528	.2060(24)
30(60)	.030	.166	.336	.561	.912	.994	.4998(24)
30(120)	.060	.467	.791	.949	1.000	1.000	.7110(24)
30(200)	.120	.832	.984	.999	1.000	1.000	.8225(24)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots, 0, \delta,\ldots, \delta, 2\delta,\ldots, 2\delta)$.

Table 78

Means of Power of H for Five Equal Subsets of δ_is by N and k (α = 0.01)

k*n	$\frac{1}{2}\delta$ = 0.10	0.20	0.30	0.40	0.50	Total
10(20)	.012	.020	.038	.073	.135	.0556(20)
10(60)	.017	.053	.160	.375	.646	.2502(20)
10(120)	.027	.136	.451	.815	.972	.4800(20)
10(200)	.043	.297	.785	.984	1.000	.6218(20)
30(20)	.013	.026	.062	.147	.304	.1104(20)
30(60)	.021	.096	.362	.752	.958	.4380(20)
30(120)	.038	.304	.836	.994	1.000	.6343(20)
30(200)	.074	.638	.990	1.000	1.000	.7404(20)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots,0,\frac{1}{2}\delta,\ldots,\frac{1}{2}\delta,\delta,\ldots,\delta,1\frac{1}{2}\delta,\ldots,1\frac{1}{2}\delta,2\delta,\ldots,2\delta)$.

Table 78.a

Means of Simulated Power of H for Five Equal Subsets of δ_is by N and k (α = 0.01)

k*n	$\frac{1}{2}\delta$ = 0.10	0.20	0.30	0.40	0.50	Total
10(20)	.023	.037	.060	.104	.169	.0786(20)
10(60)	.021	.057	.165	.386	.660	.2575(20)
10(120)	.030	.144	.458	.817	.976	.4848(20)
10(200)	.042	.295	.785	.986	1.000	.6215(20)
30(20)	.051	.067	.119	.214	.372	.1645(20)
30(60)	.027	.103	.377	.760	.958	.4449(20)
30(120)	.042	.313	.842	.994	1.000	.6383(20)
30(200)	.075	.642	.991	1.000	1.000	.7416(20)

Note: The pattern of three equal subsets of δ_i values was $(0,\ldots,0,\frac{1}{2}\delta,\ldots,\frac{1}{2}\delta,\delta,\ldots,\delta,1\frac{1}{2}\delta,\ldots,1\frac{1}{2}\delta,2\delta,\ldots,2\delta)$.

Table 79

Means of Power of H_+ at $\alpha = 0.10$ for $\mu_\delta = 0$ for the Random-effects Model

σ^2_δ	N = 20k	60k	120k	200k	Total
.00	0.10(16)	0.10(16)	0.10(16)	0.10(16)	0.10(64)
.00-.02	0.12(16)	0.21(20)	0.33(20)	0.45(20)	0.29(76)
.02-.04	0.17(16)	0.39(20)	0.58(20)	0.70(20)	0.47(76)
.04-.06	0.22(16)	0.52(20)	0.62(16)	0.73(16)	0.52(68)
.06-.08	0.27(16)	0.61(20)	0.59(12)	0.71(12)	0.53(60)
.08-.10	0.33(32)	0.55(28)	0.66(24)	0.76(24)	0.56(108)
.15	0.44(16)	0.61(12)	0.75(12)	0.82(12)	0.64(52)
.20	0.51(16)	0.67(12)	0.79(12)	0.85(12)	0.69(52)
.25	0.56(16)	0.71(12)	0.81(12)	0.80(8)	0.70(48)

Table 80

Means of Power of H_+ at $\alpha = 0.025$ for $\mu_\delta = 0$ for the Random-effects Model

σ^2_δ	N = 20k	60k	120k	200k	Total
.00	.025(16)	.025(16)	.025(16)	.025(16)	.025(64)
.00-.02	.033(16)	.080(20)	.169(20)	.288(20)	.148(76)
.02-.04	.054(16)	.213(20)	.425(20)	.579(20)	.332(76)
.04-.06	.081(16)	.351(20)	.478(16)	.615(16)	.379(68)
.06-.08	.114(16)	.459(20)	.433(12)	.587(12)	.387(60)
.08-.10	.169(32)	.394(28)	.526(24)	.660(24)	.413(108)
.15	.260(16)	.452(12)	.646(12)	.746(12)	.506(52)
.20	.340(16)	.535(12)	.706(12)	.786(12)	.572(52)
.25	.406(16)	.598(12)	.741(12)	.719(8)	.590(48)

Table 81

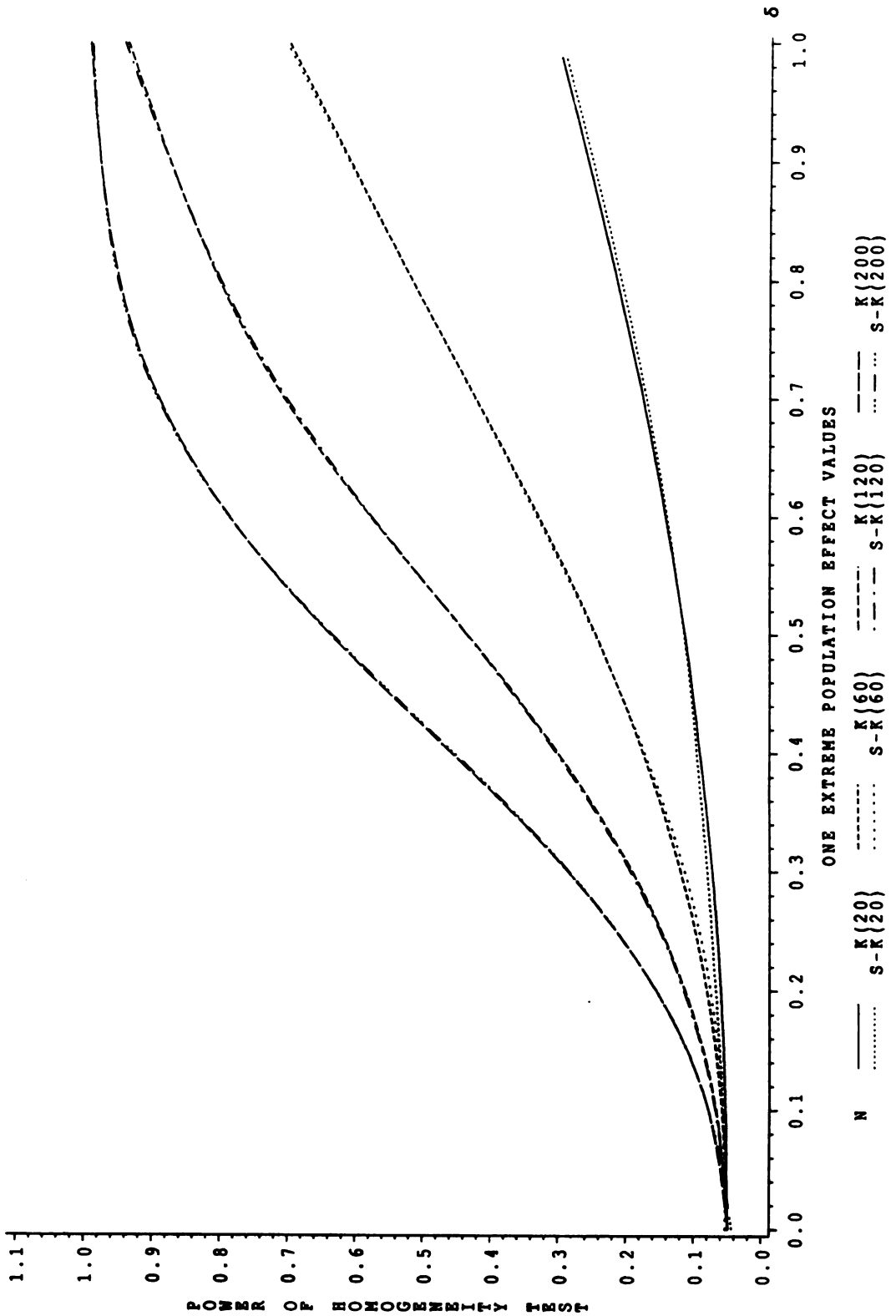
Means of Power of $\underline{H}_+$ at $\alpha = 0.01$ for $\mu_\delta = 0$ for the Random-effects Model

σ^2_δ	$\underline{N} = 20\underline{k}$	$60\underline{k}$	$120\underline{k}$	$200\underline{k}$	Total
.00	.010(16)	.010(16)	.010(16)	.010(16)	.010(64)
.00-.02	.014(16)	.042(20)	.110(20)	.217(20)	.100(76)
.02-.04	.025(16)	.142(20)	.346(20)	.516(20)	.269(76)
.04-.06	.042(16)	.267(20)	.404(16)	.554(16)	.314(68)
.06-.08	.065(16)	.380(20)	.351(12)	.518(12)	.318(60)
.08-.10	.100(32)	.315(28)	.450(24)	.603(24)	.345(108)
.15	.184(16)	.370(12)	.587(12)	.704(12)	.440(52)
.20	.260(16)	.460(12)	.657(12)	.752(12)	.511(52)
.25	.326(16)	.531(12)	.699(12)	.676(8)	.529(48)

APPENDIX D

FIGURE 4.1.1

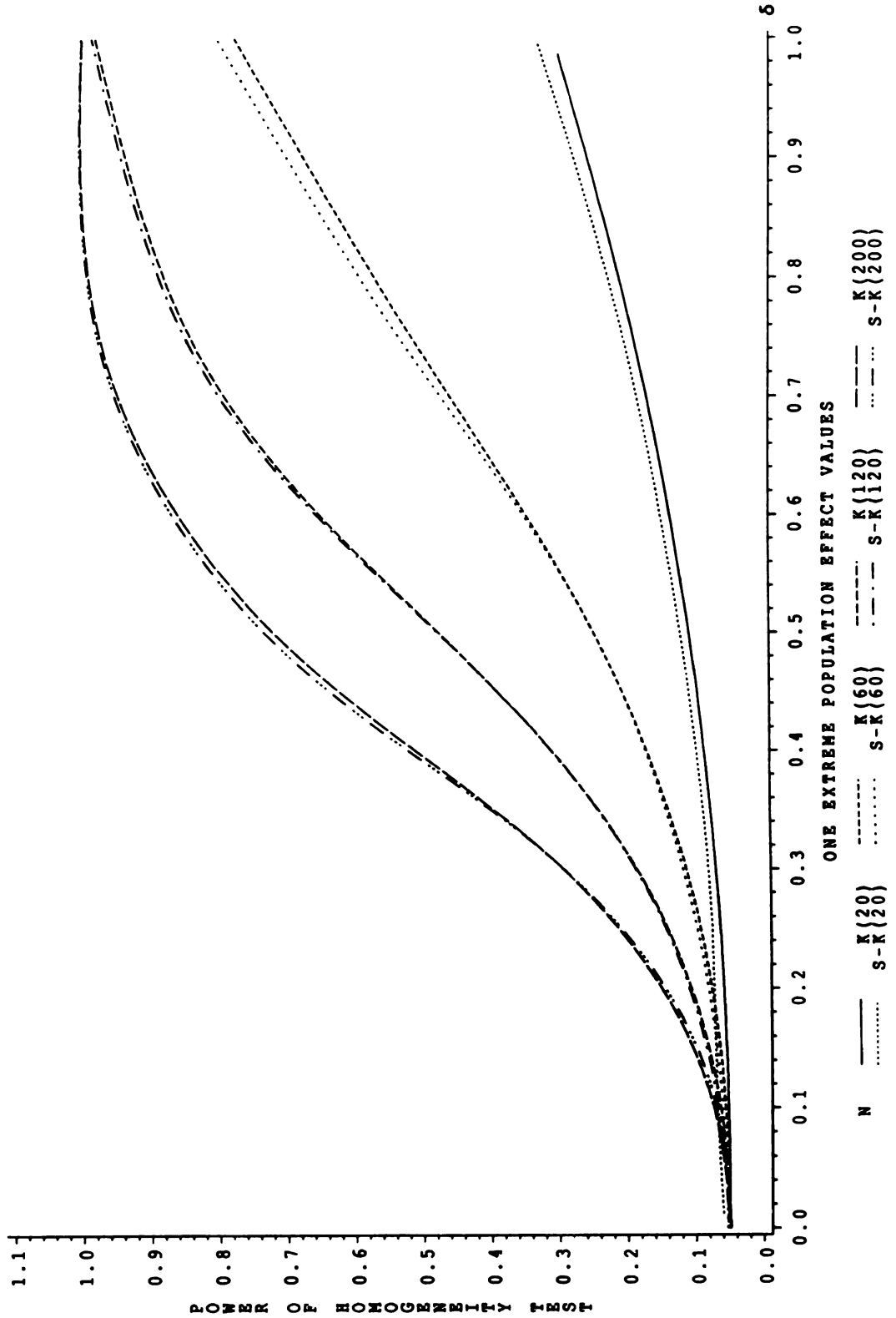
POWER CURVE WITH K = 2 (ALPHA = 0.050)
FIXED-EFFECTS MODELS



Note: "S-" represents the simulated power curves.

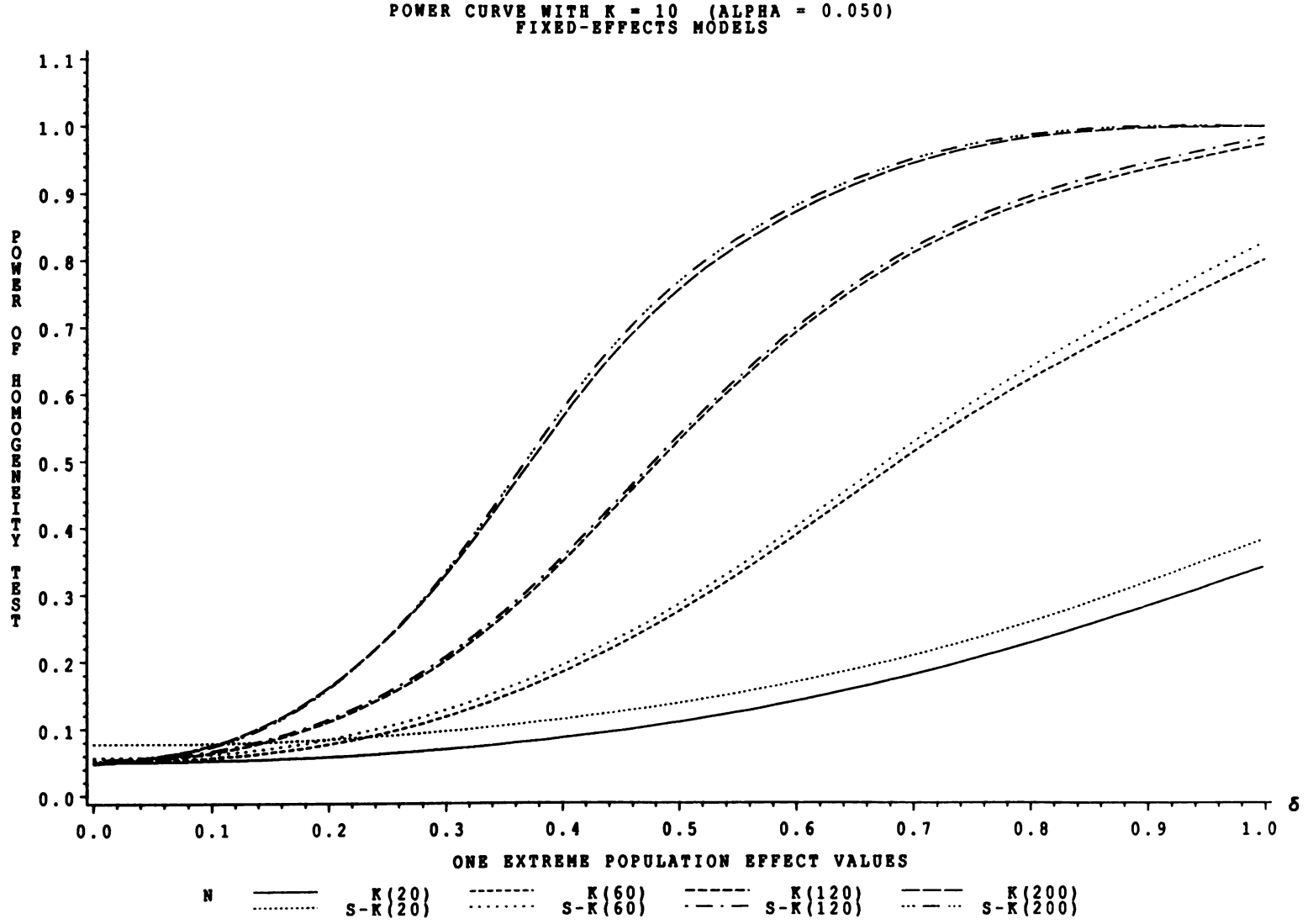
FIGURE 4.1.2

POWER CURVE WITH K = 5 (ALPHA = 0.050)
FIXED-EFFECTS MODELS



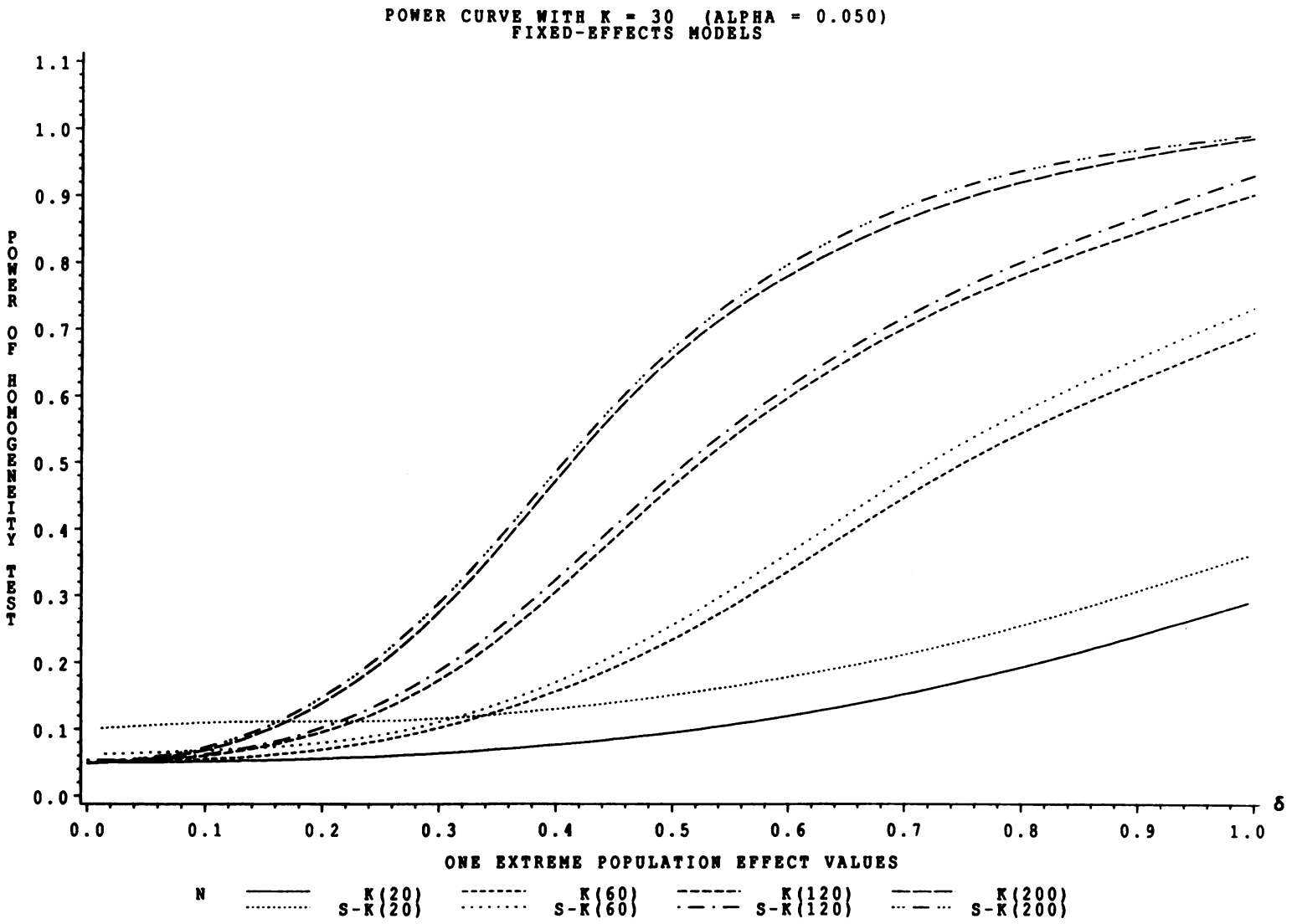
Note: "S-" represents the simulated power curves.

FIGURE 4.1.3



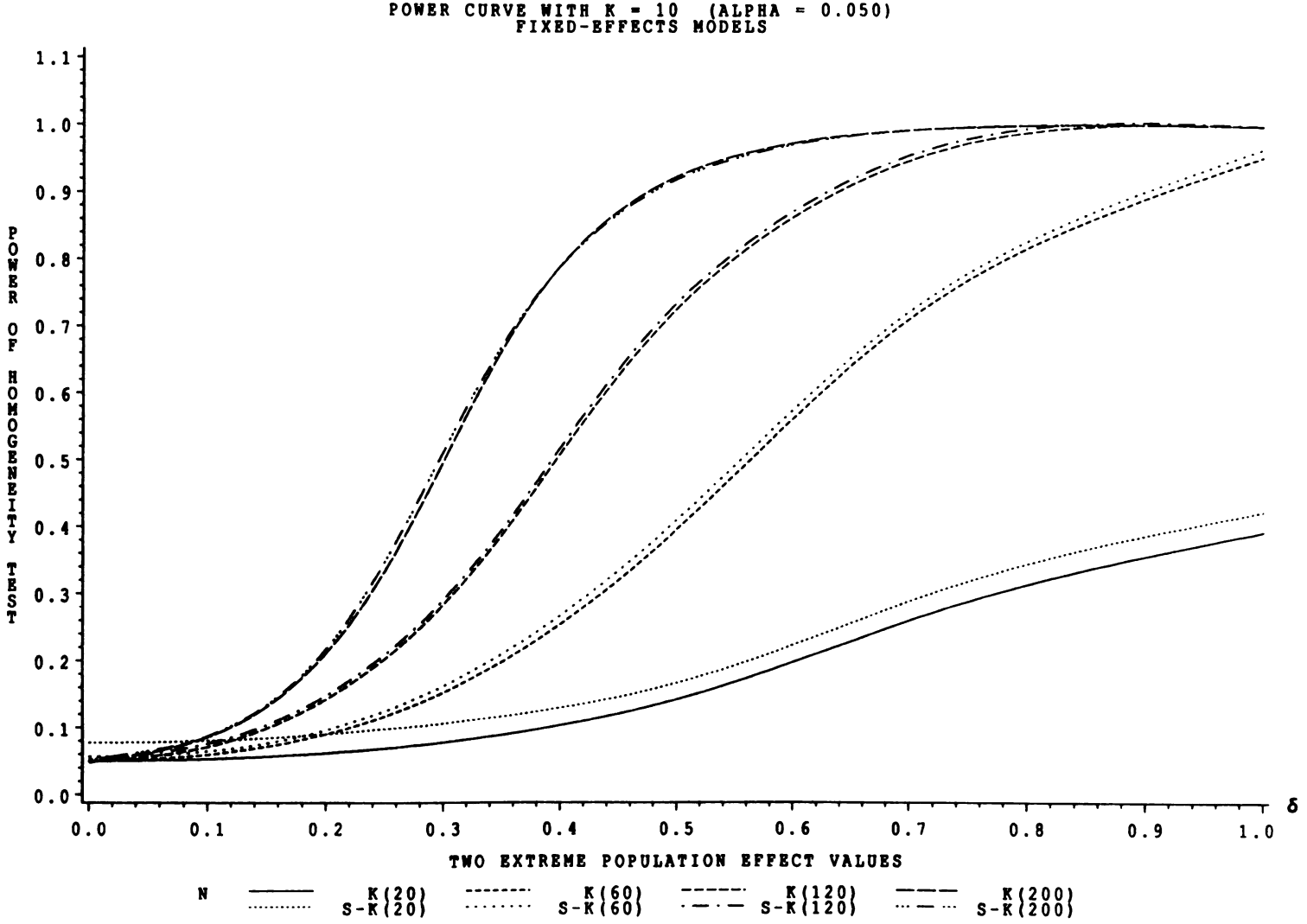
Note: "S-" represents the simulated power curves.

FIGURE 4.1.4



Note: "S-" represents the simulated power curves.

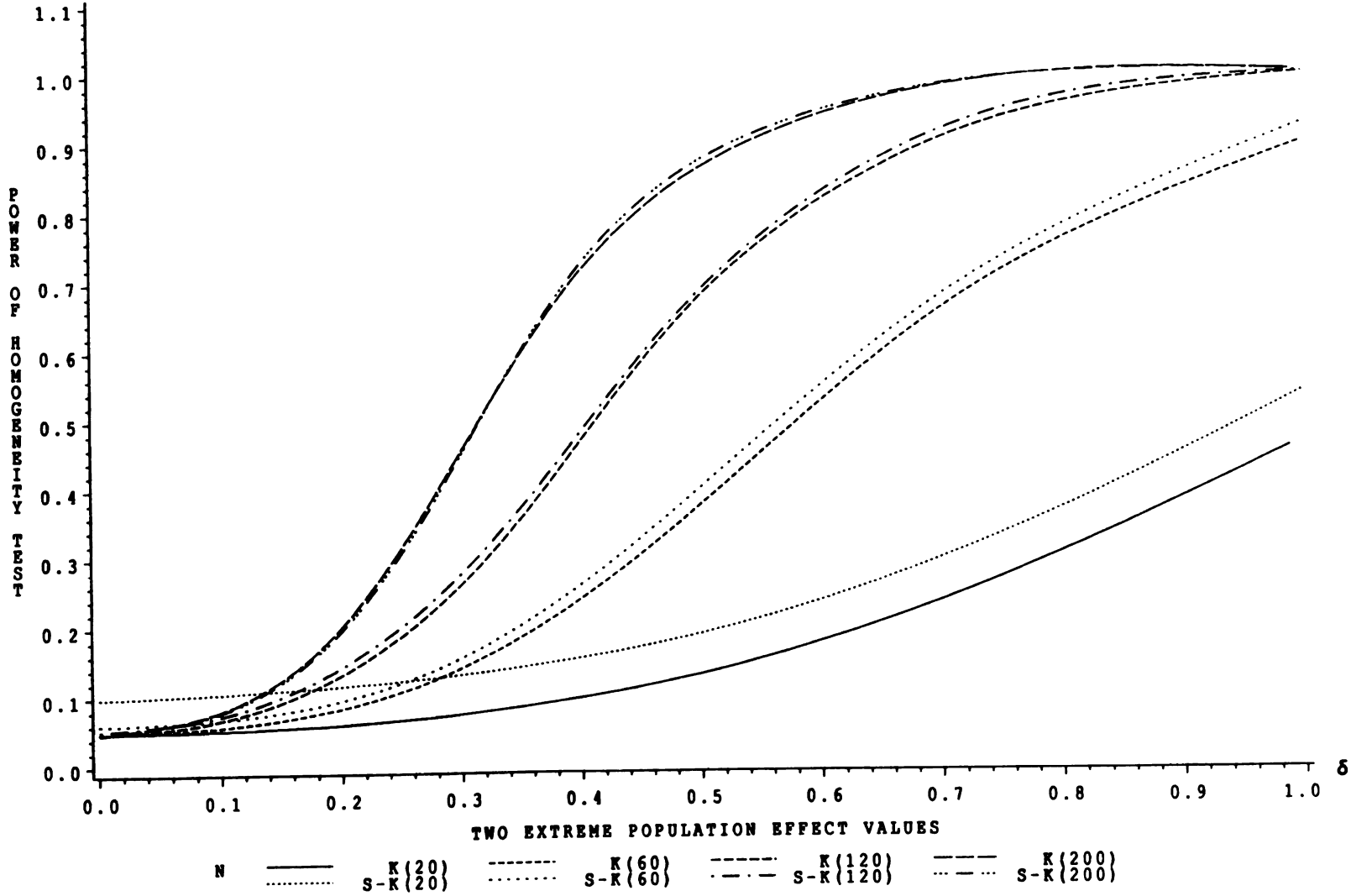
FIGURE 4.2.1



Note: "S-" represents the simulated power curves.

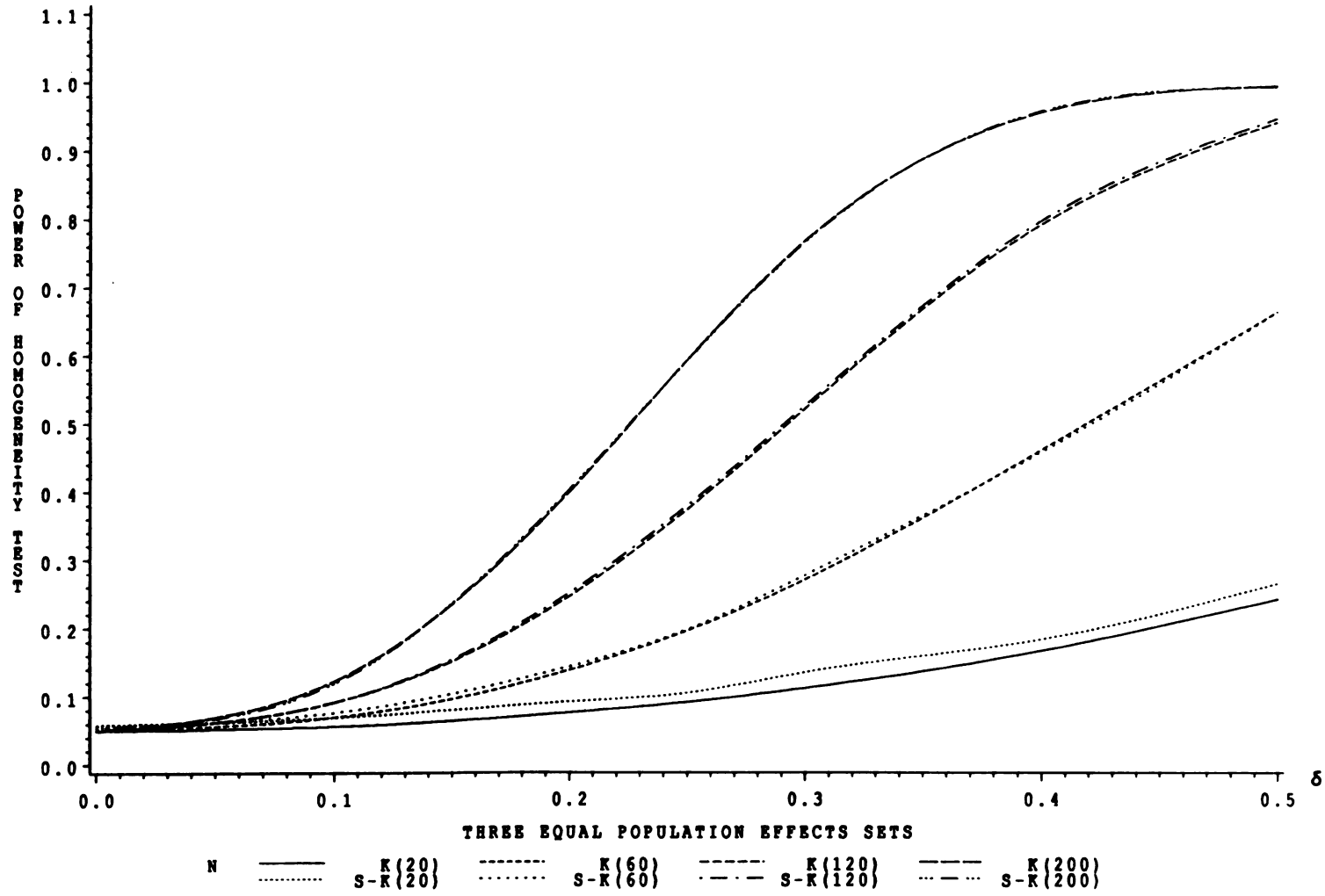
FIGURE 4.2.2

POWER CURVE WITH K = 30 (ALPHA = 0.050)
FIXED-EFFECTS MODELS



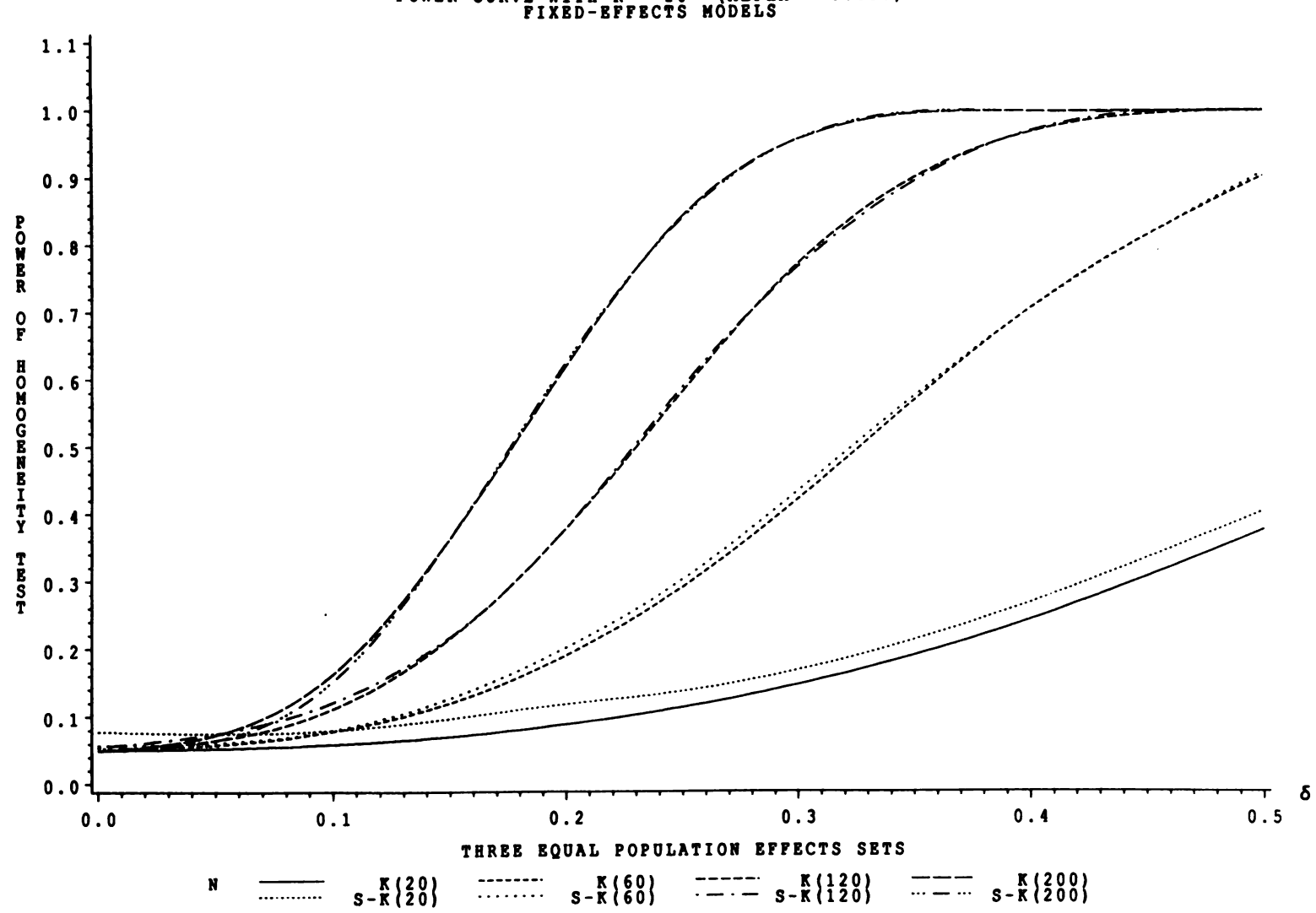
Note: "S-" represents the simulated power curves.

FIGURE 4.3.1



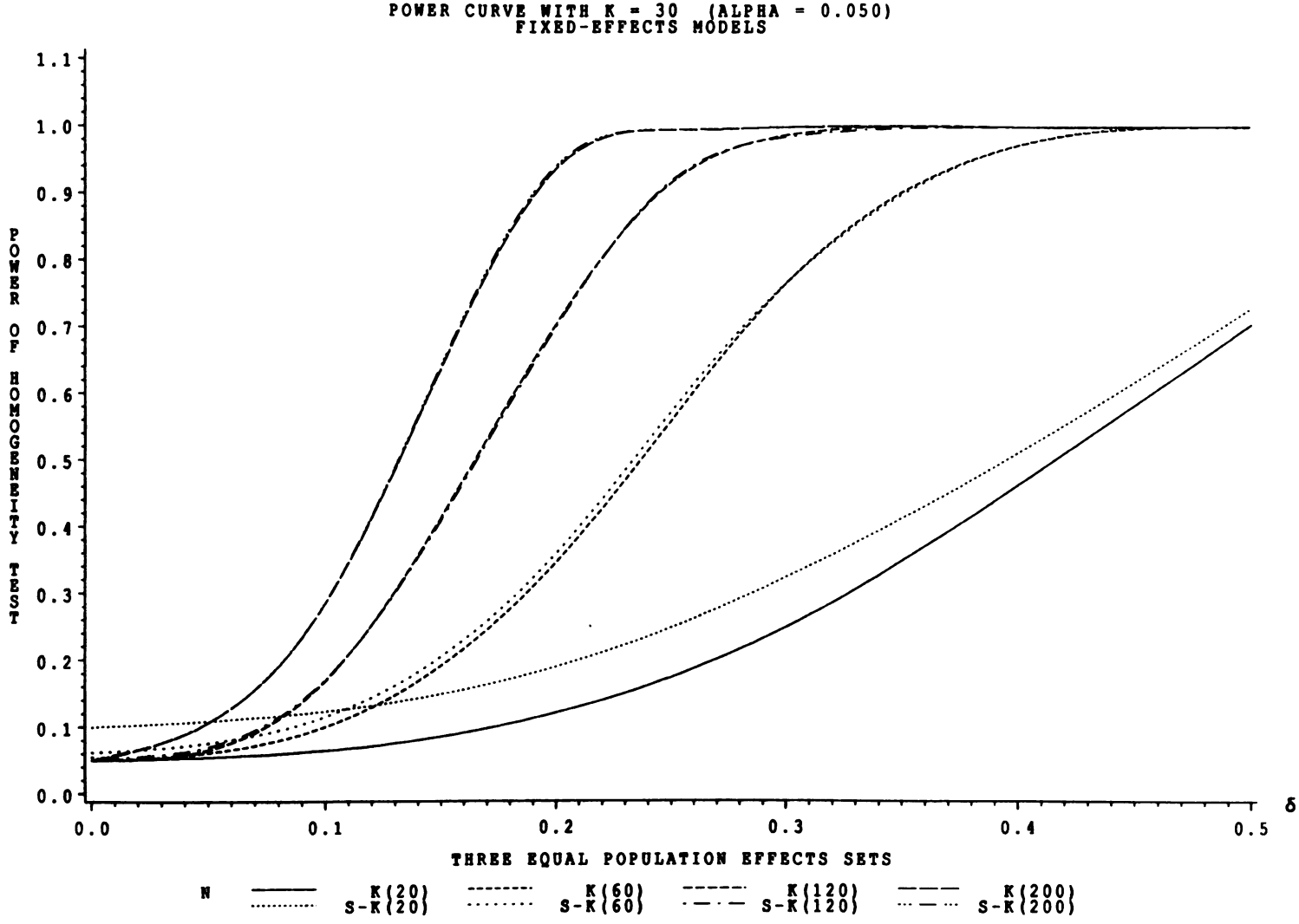
Note: "S-" represents the simulated power curves.

FIGURE 4.3.2



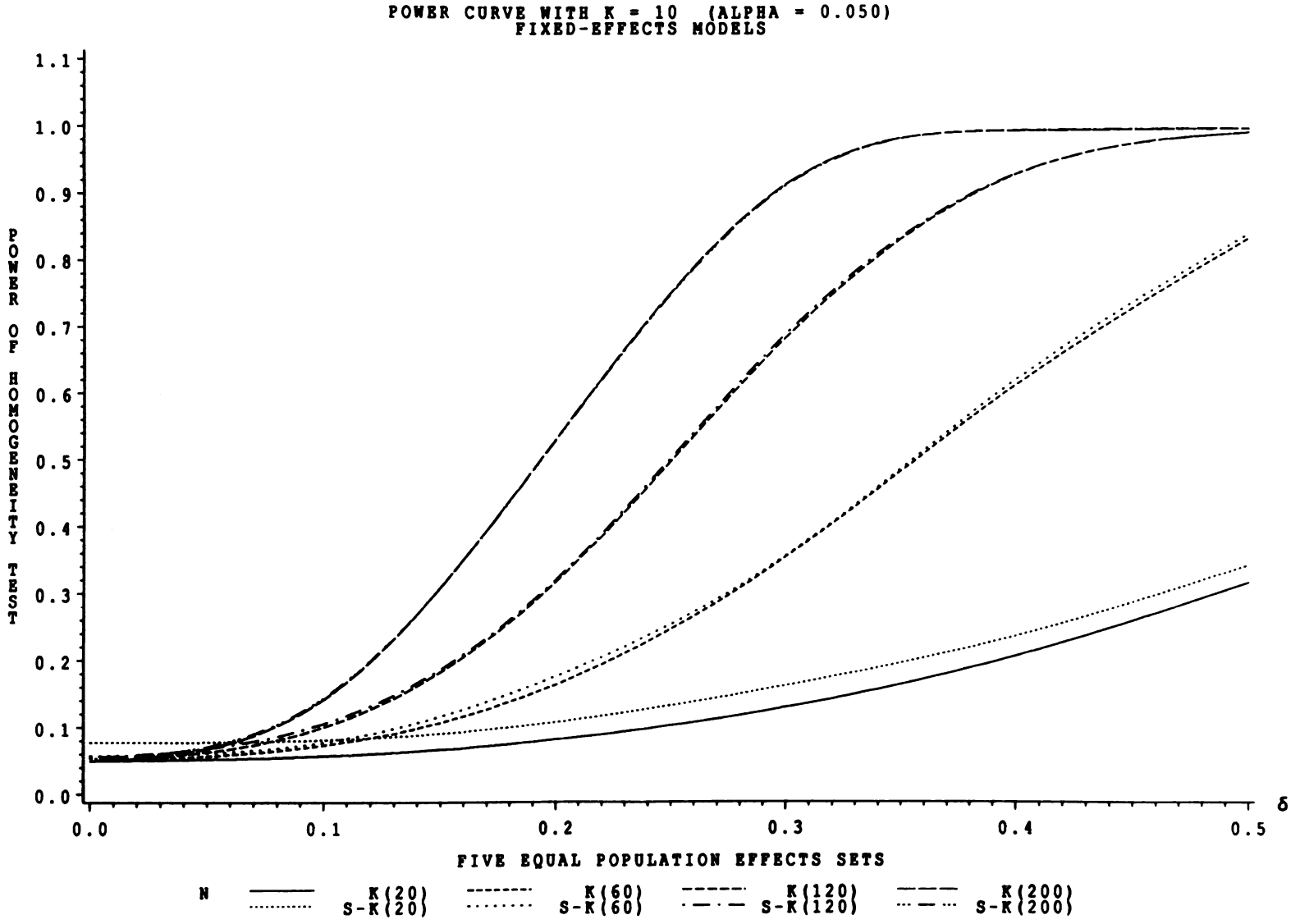
Note: "S-" represents the simulated power curves.

FIGURE 4.3.3



Note: "S-" represents the simulated power curves.

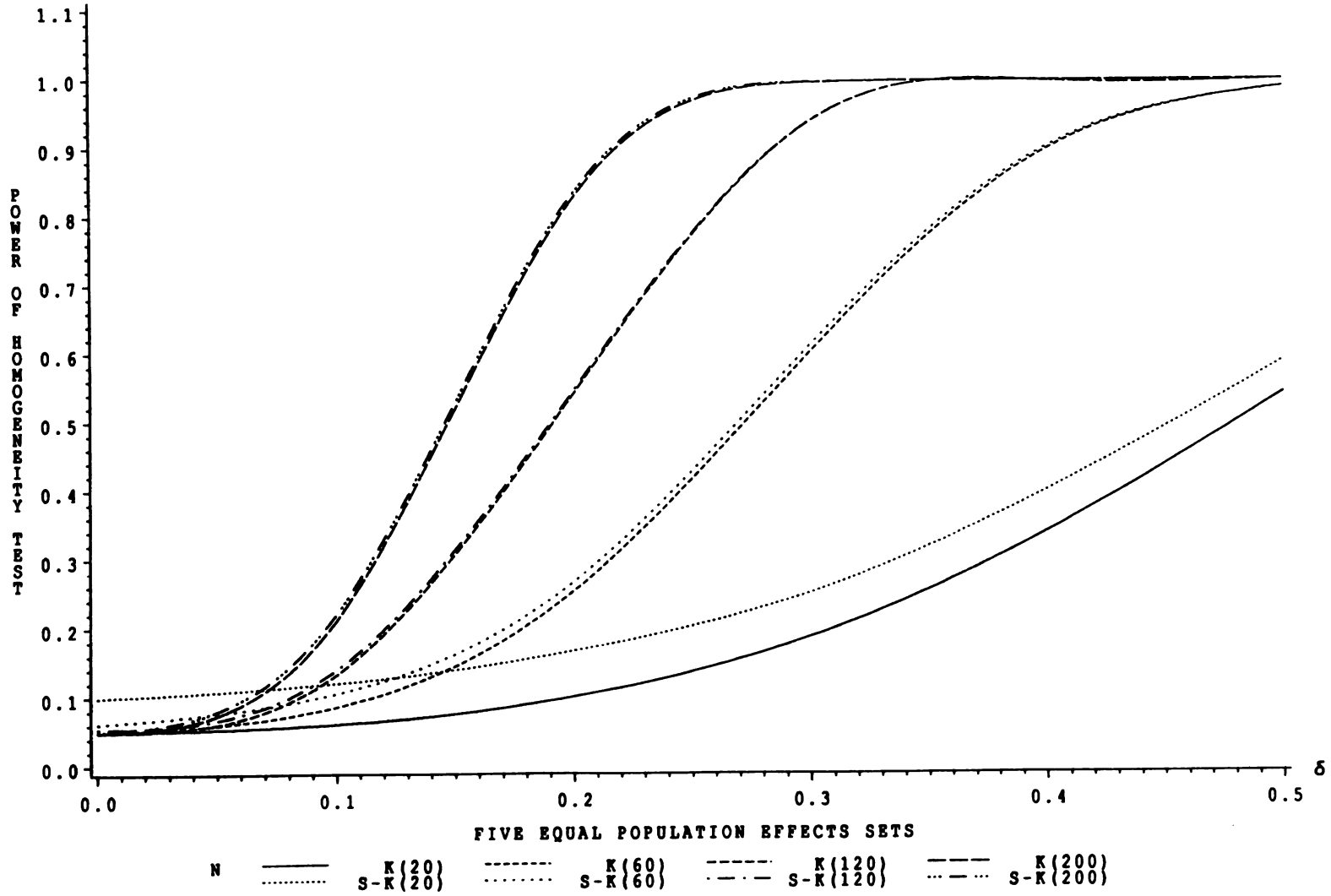
FIGURE 4.4.1



Note: "S-" represents the simulated power curves.

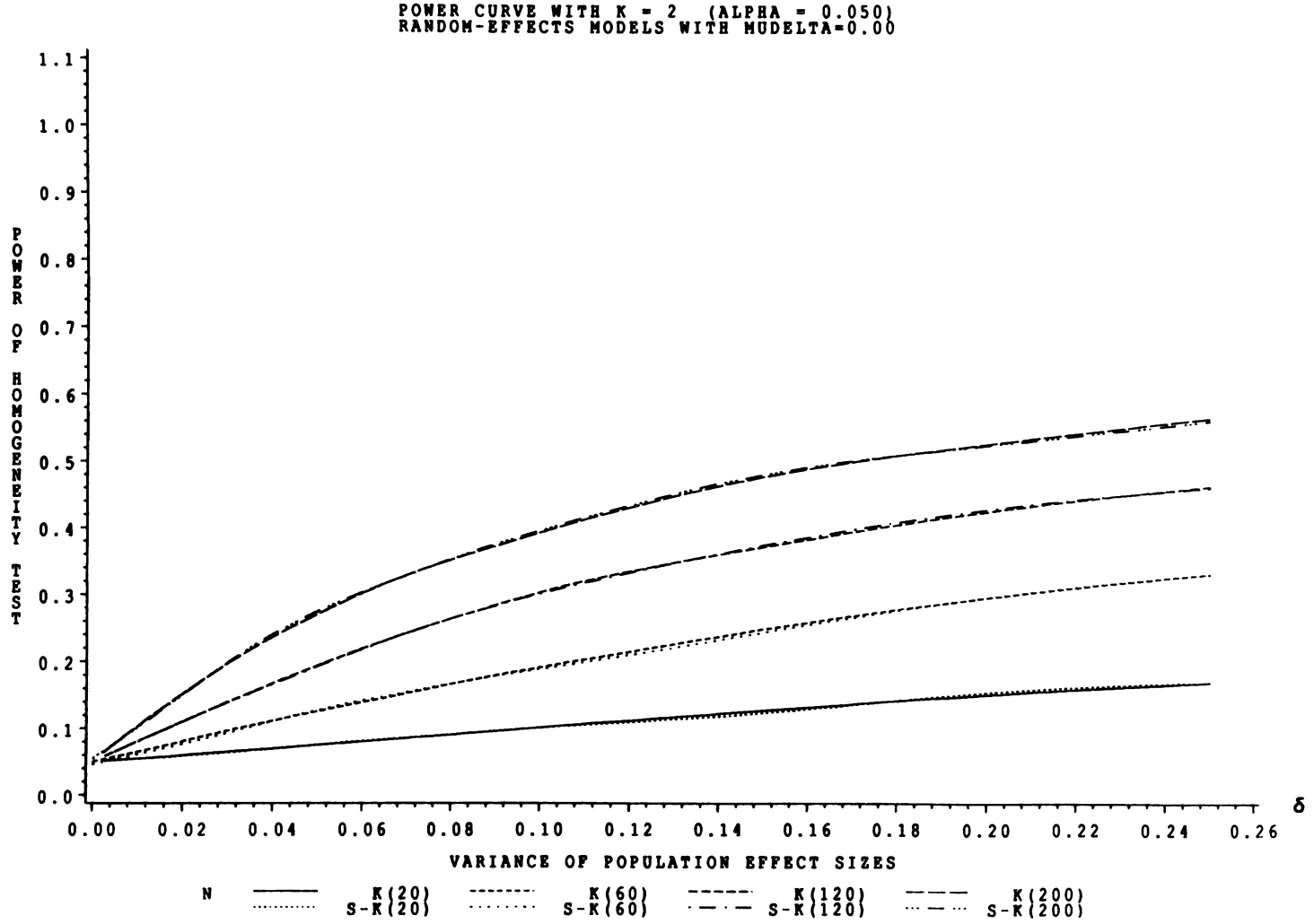
FIGURE 4.4.2

POWER CURVE WITH K = 30 (ALPHA = 0.050)
FIXED-EFFECTS MODELS



Note: "S-" represents the simulated power curves.

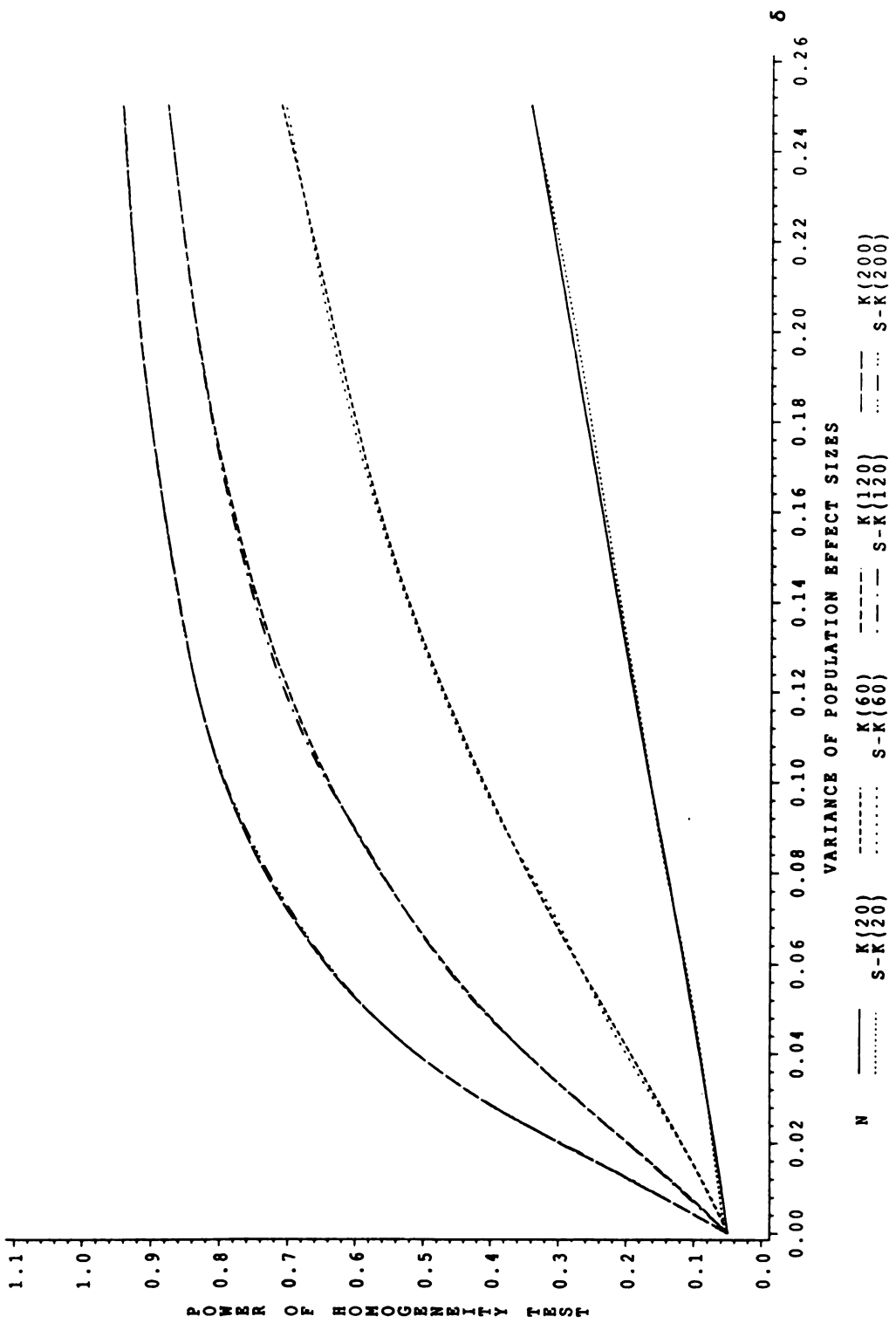
FIGURE 4.5.1



Note: "S-" represents the simulated power curves.

FIGURE 4.5.2

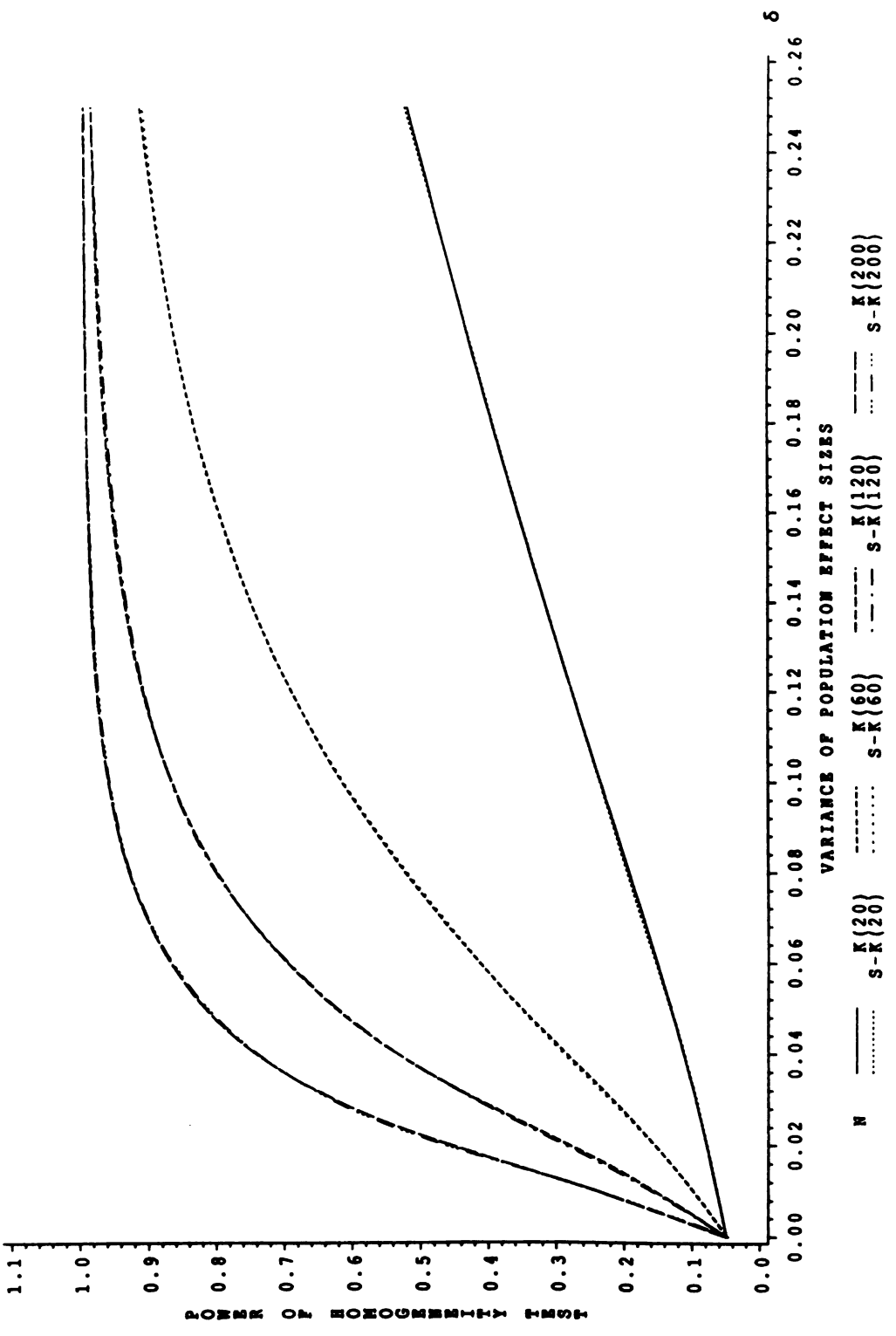
POWER CURVE WITH K = 5 (ALPHA = 0.050)
RANDOM-EFFECTS MODELS WITH MUDELTA=0.00



Note: "S-" represents the simulated power curves.

FIGURE 4.5.3

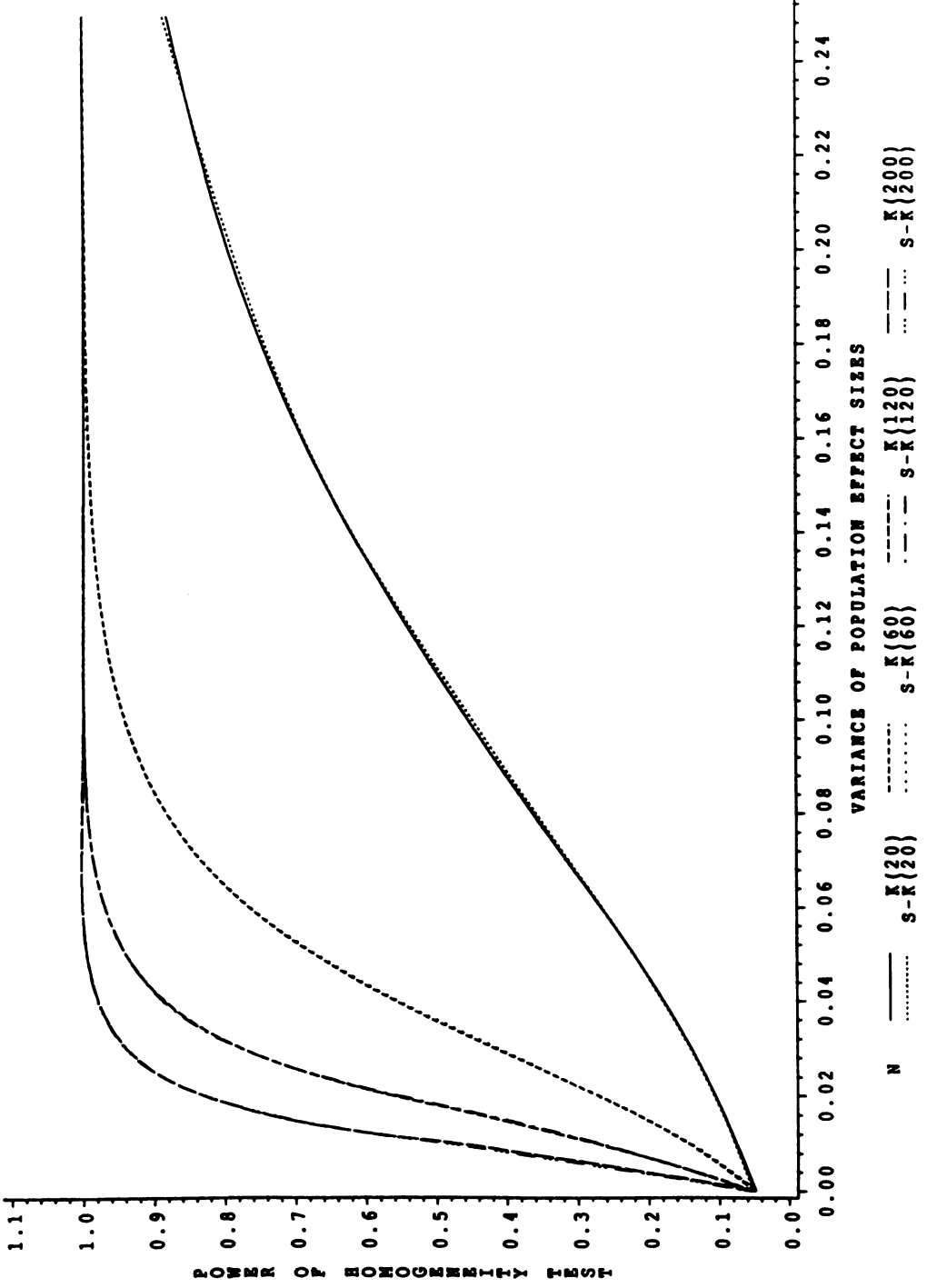
POWER CURVE WITH K = 10 (ALPHA = 0.050)
RANDOM-EFFECTS MODELS WITH NUDELTA=0.00



Note: "S-" represents the simulated power curves.

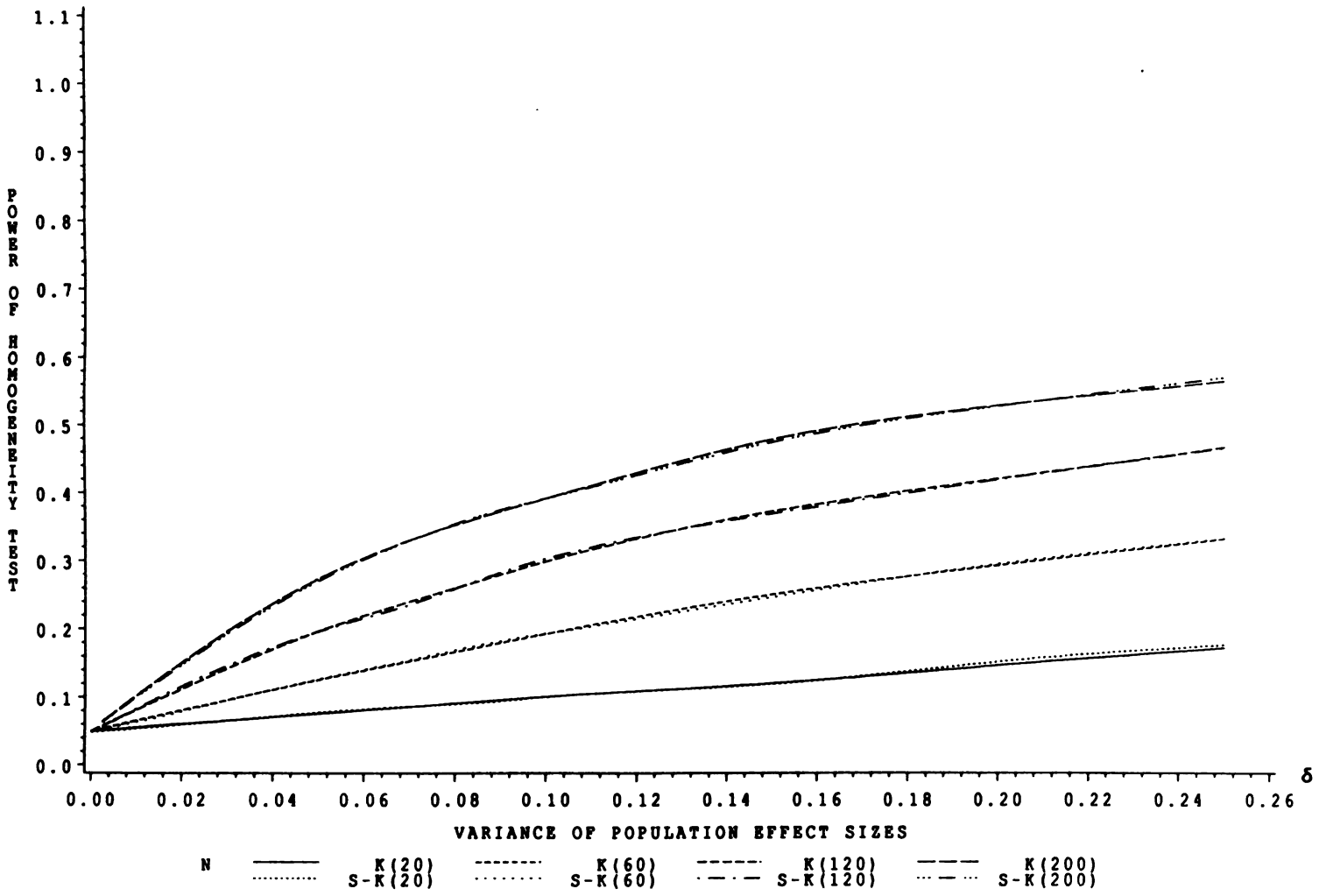
FIGURE 4.5.4

POWER CURVE WITH K = 30 (ALPHA = 0.050)
RANDOM-EFFECTS MODELS WITH NUDELTA=0.00



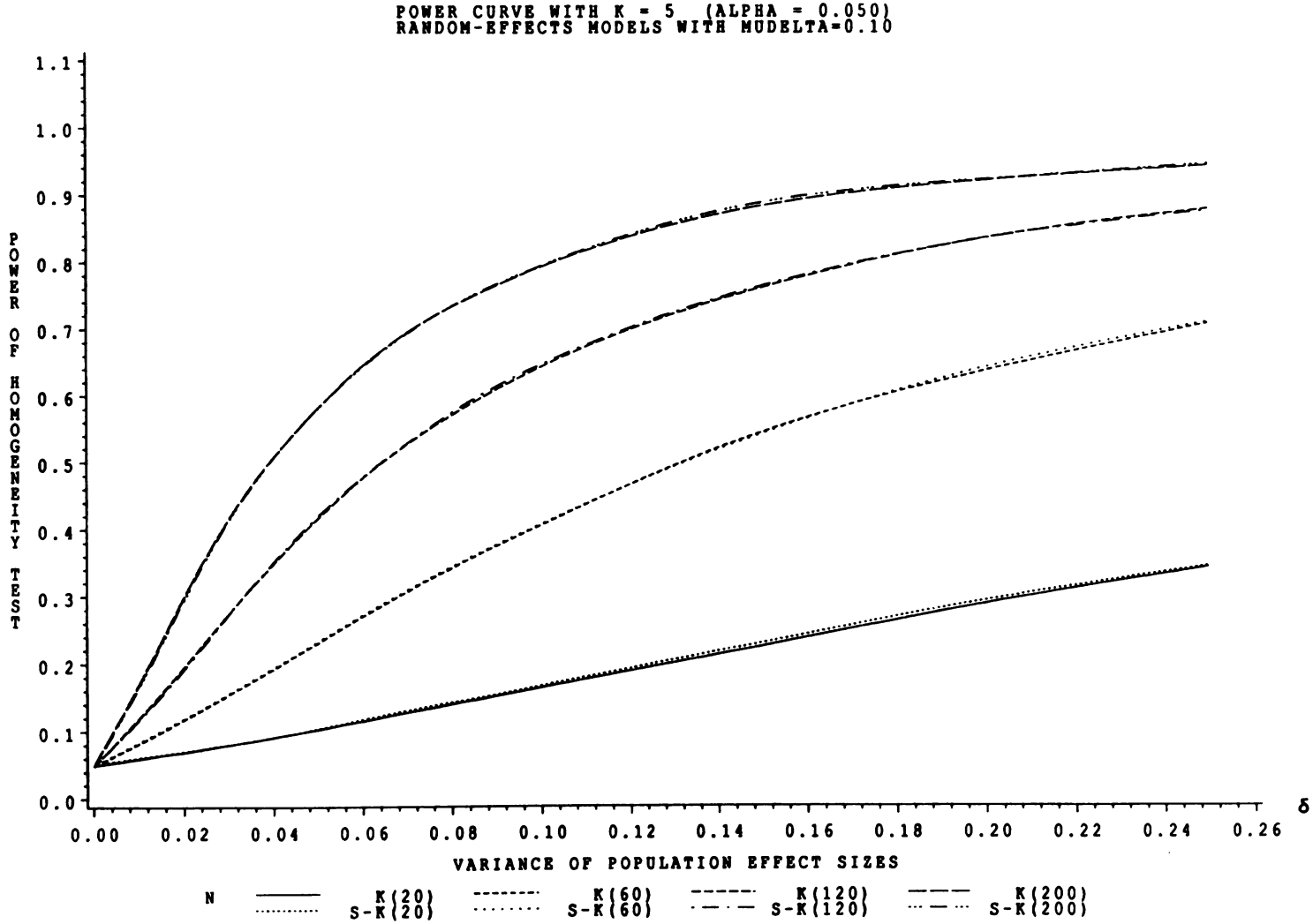
Note: "S-" represents the simulated power curves.

FIGURE 4.6.1



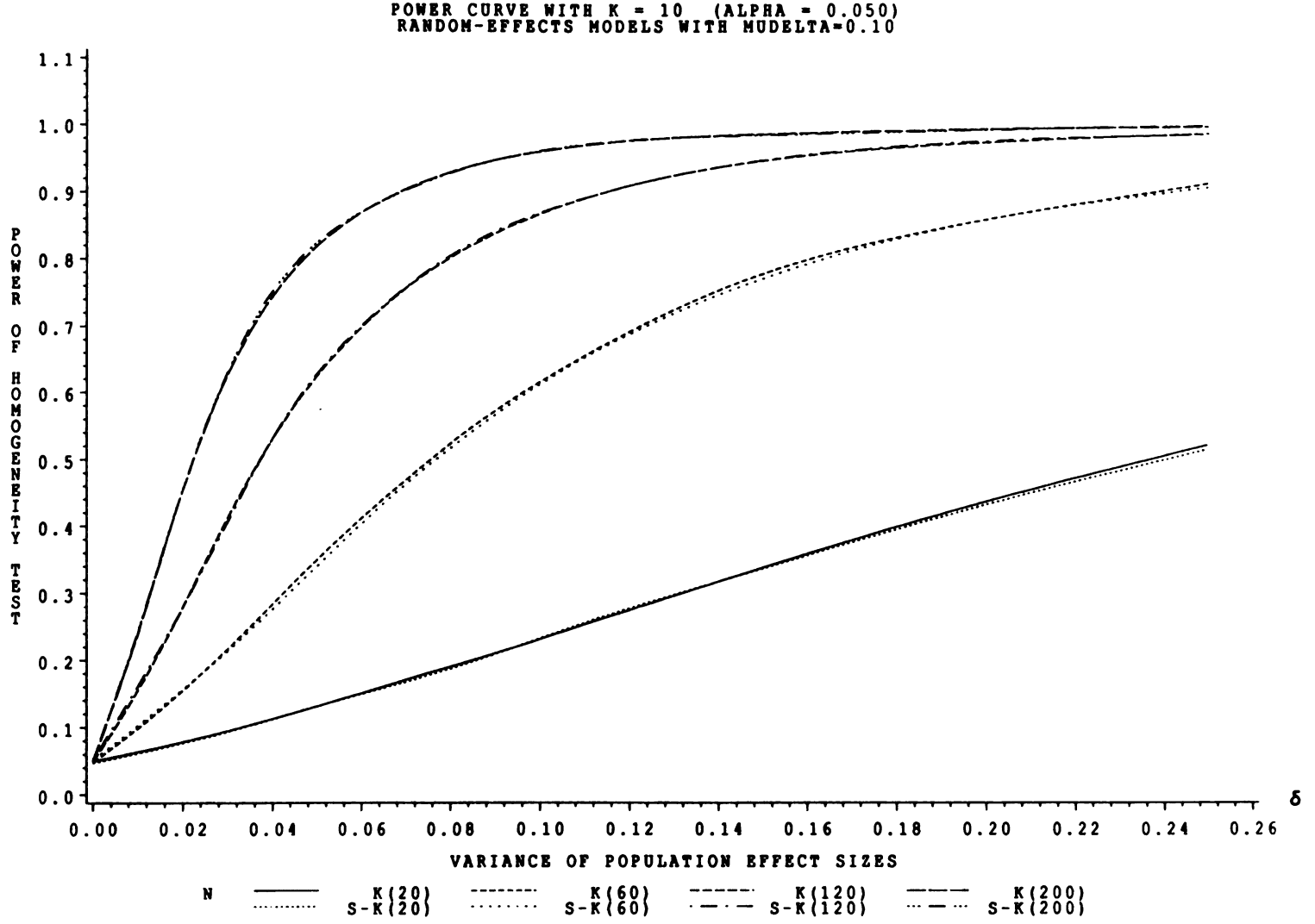
Note: "S-" represents the simulated power curves.

FIGURE 4.6.2



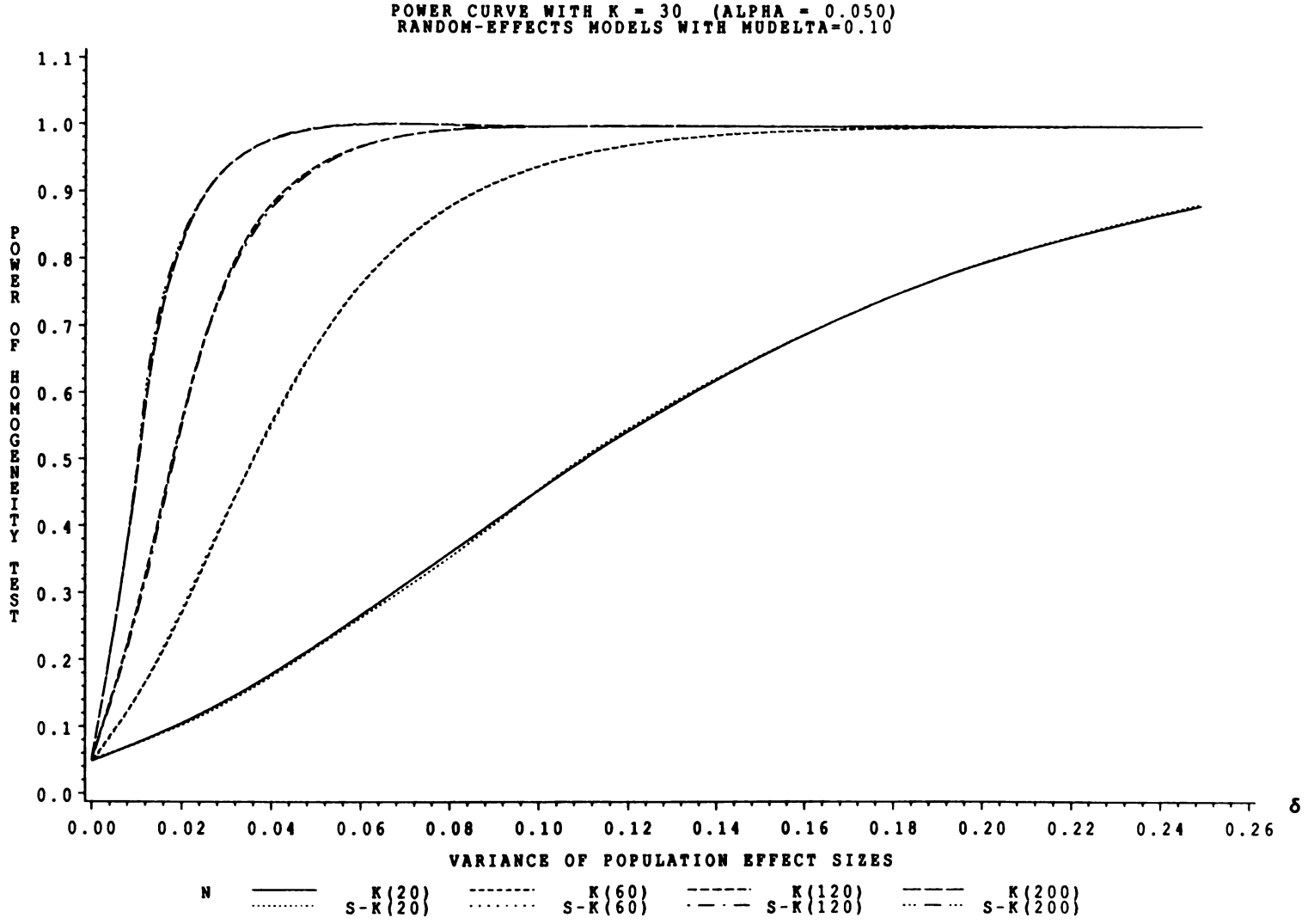
Note: "S-" represents the simulated power curves.

FIGURE 4.6.3



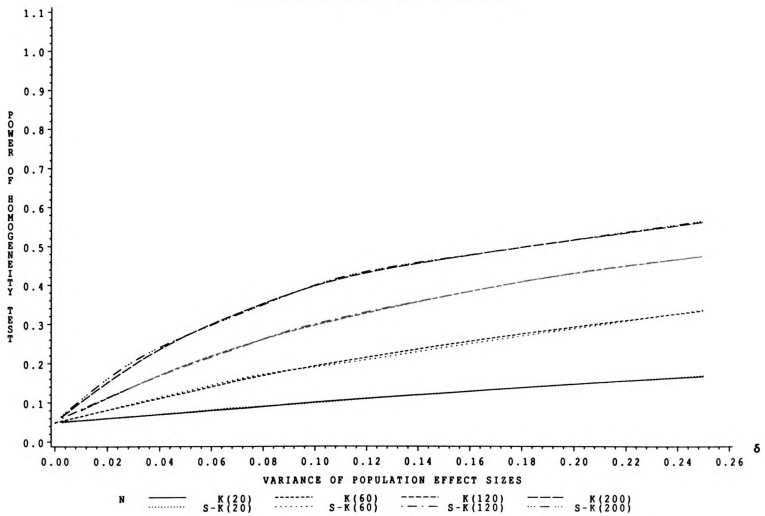
Note: "S-" represents the simulated power curves.

FIGURE 4.6.4



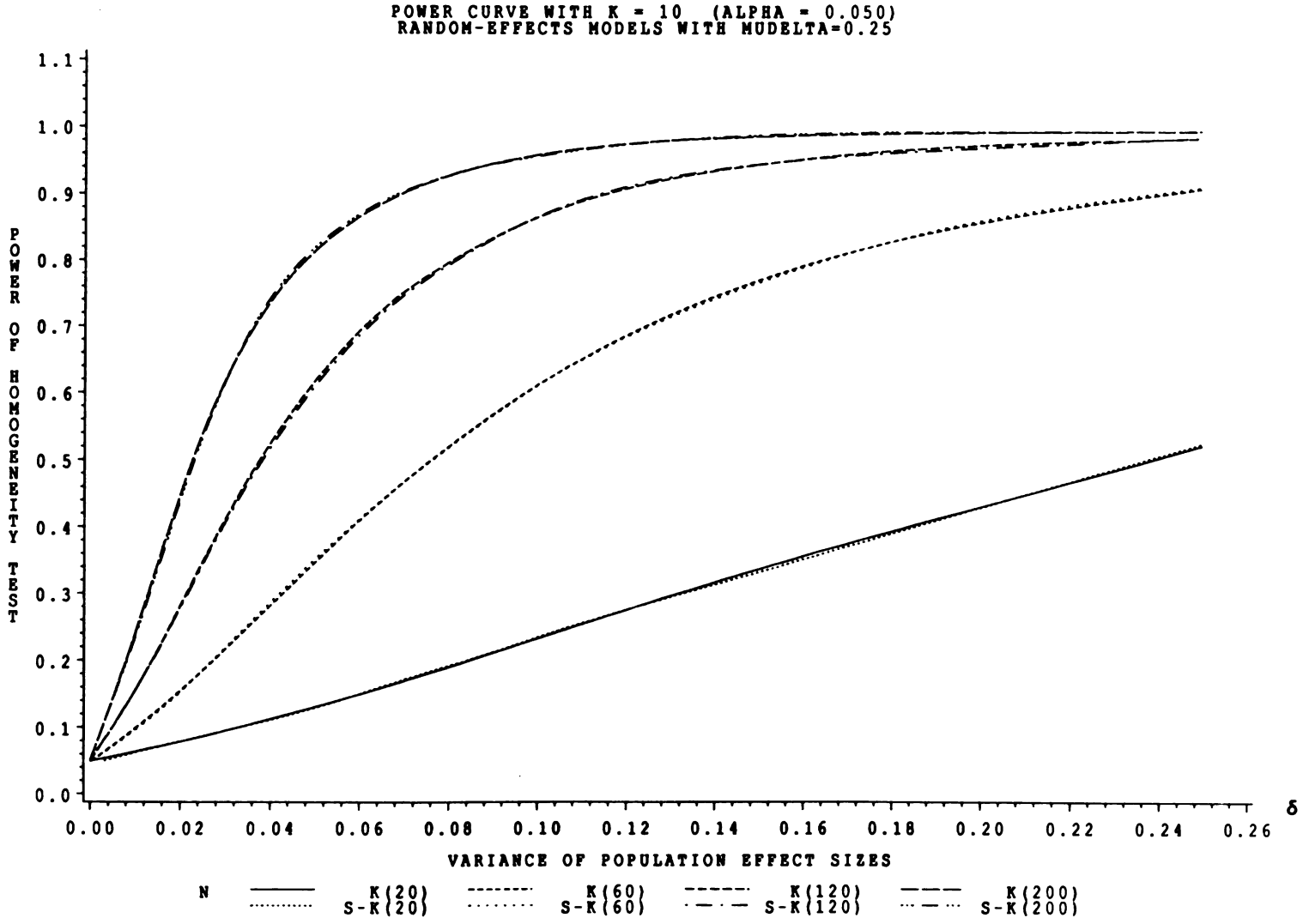
Note: "S-" represents the simulated power curves.

FIGURE 4.7.1



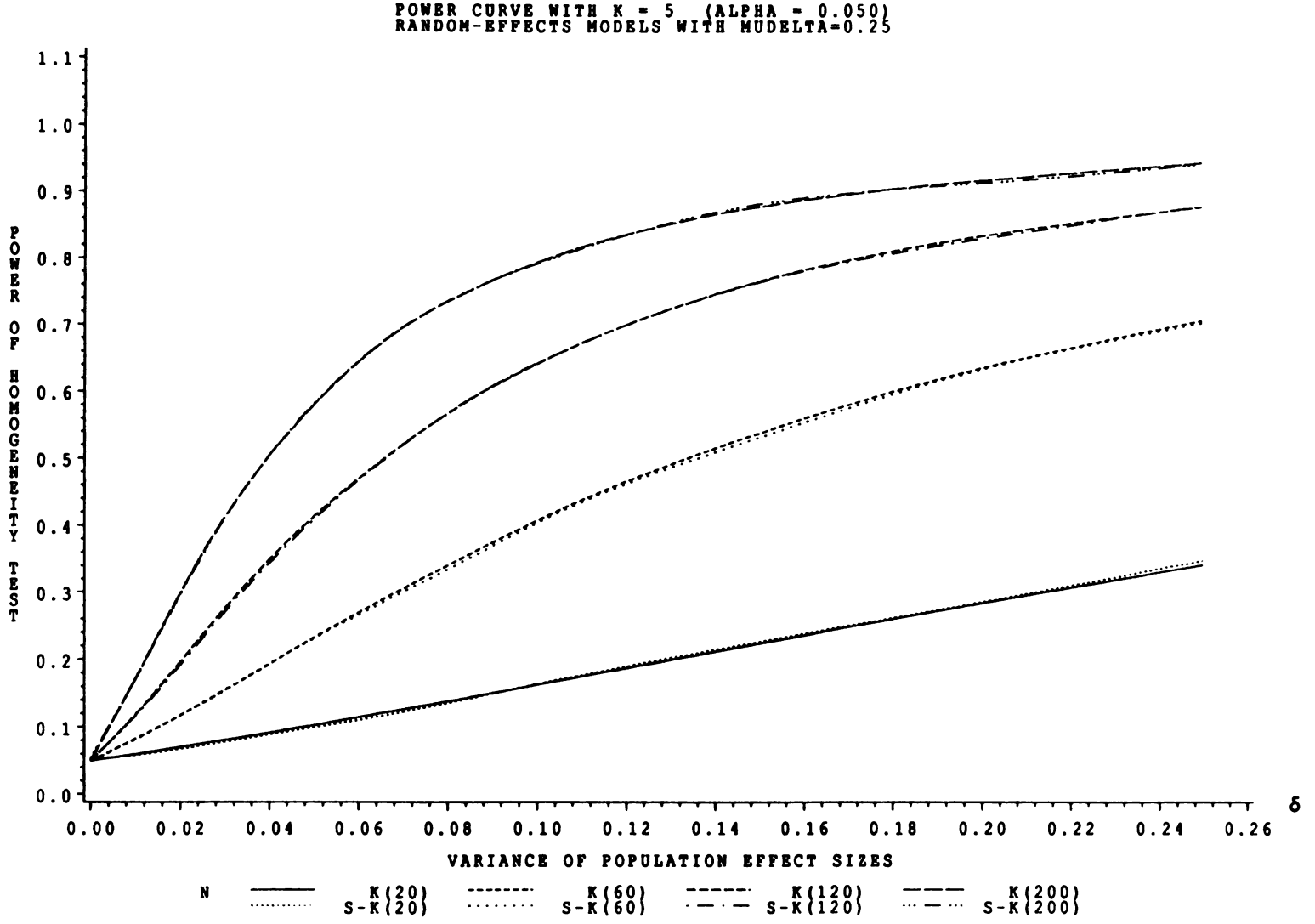
Note: "S-" represents the simulated power curves.

FIGURE 4.7.3



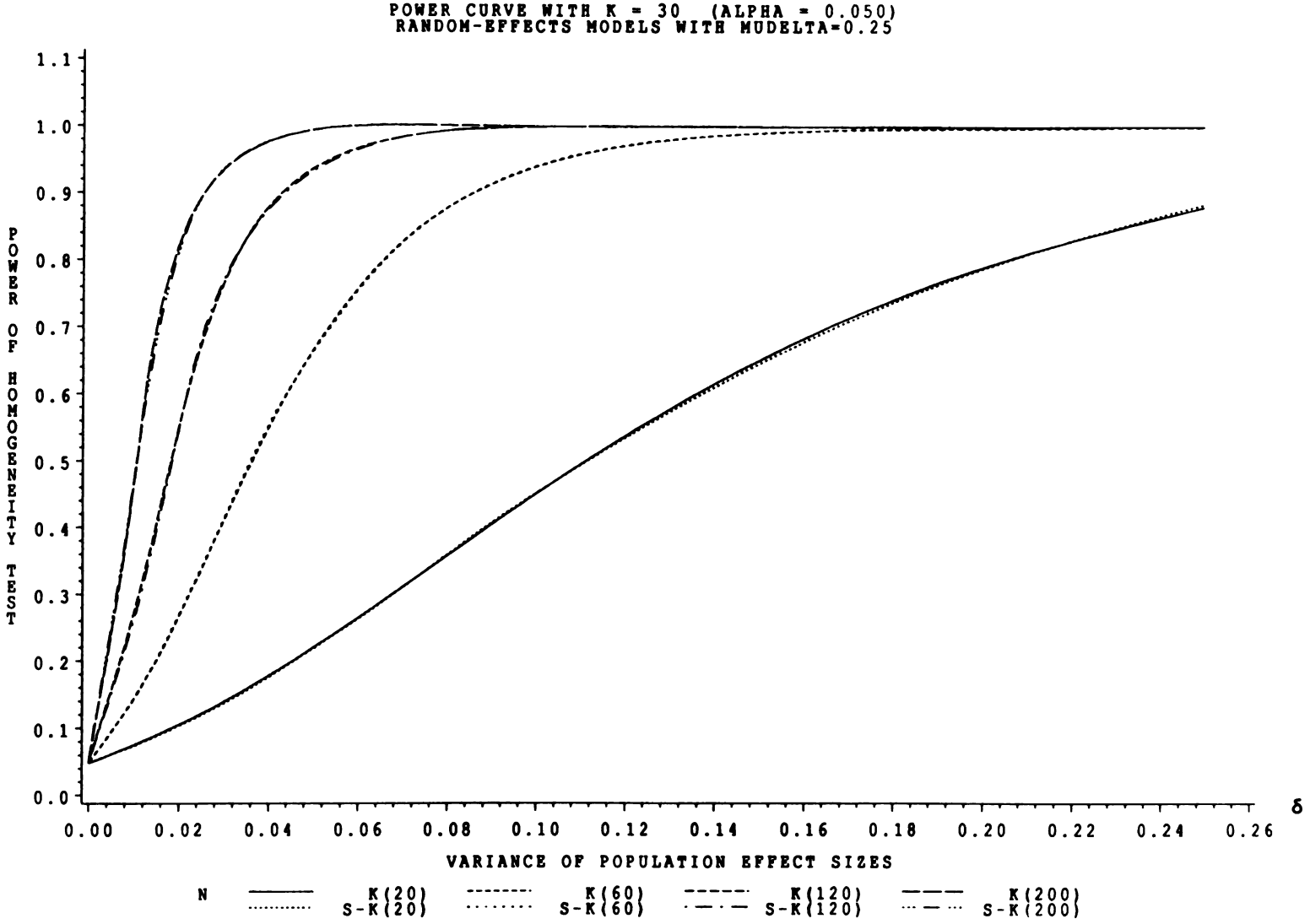
Note: "S-" represents the simulated power curves.

FIGURE 4.7.2



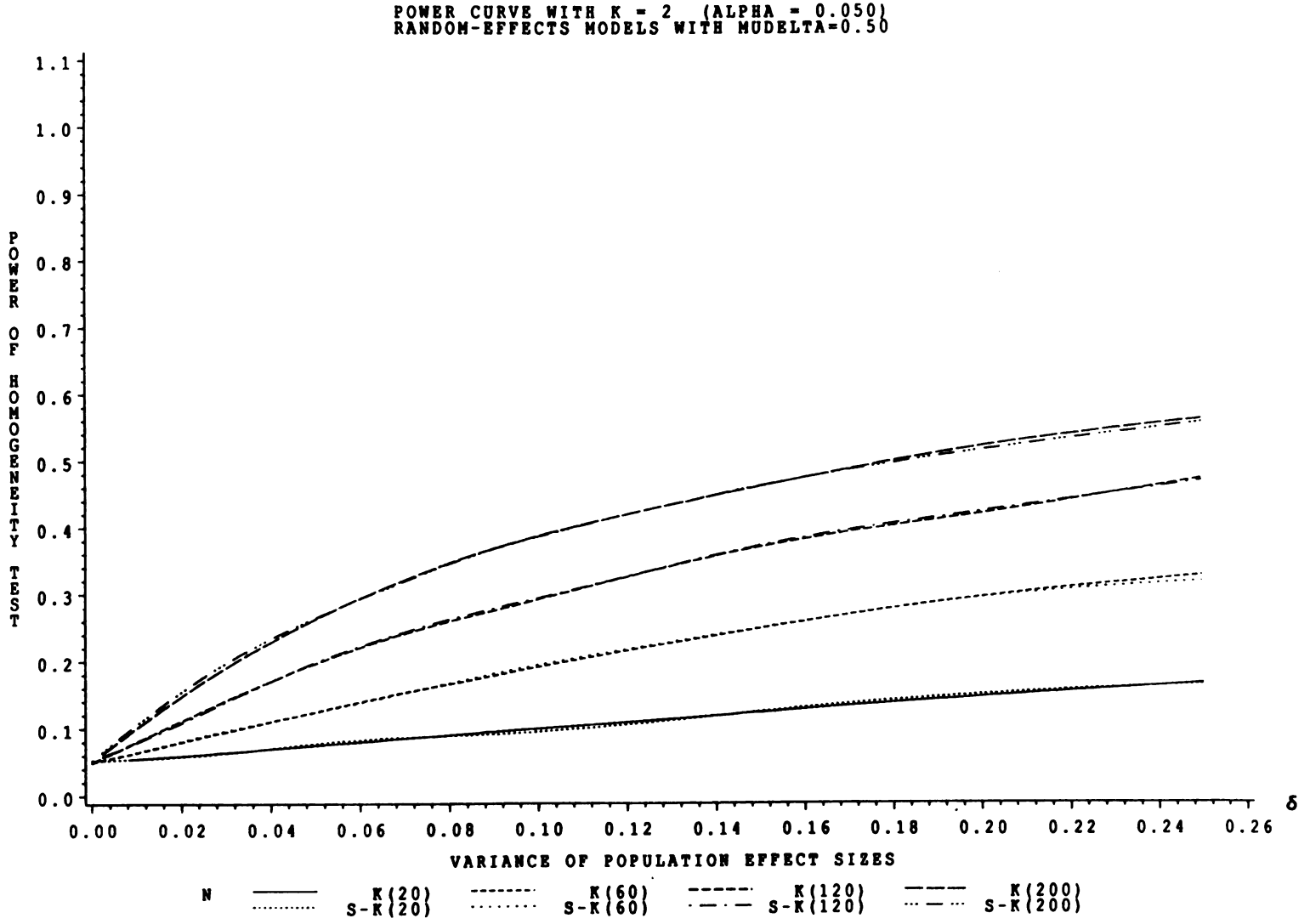
Note: "S-" represents the simulated power curves.

FIGURE 4.7.4



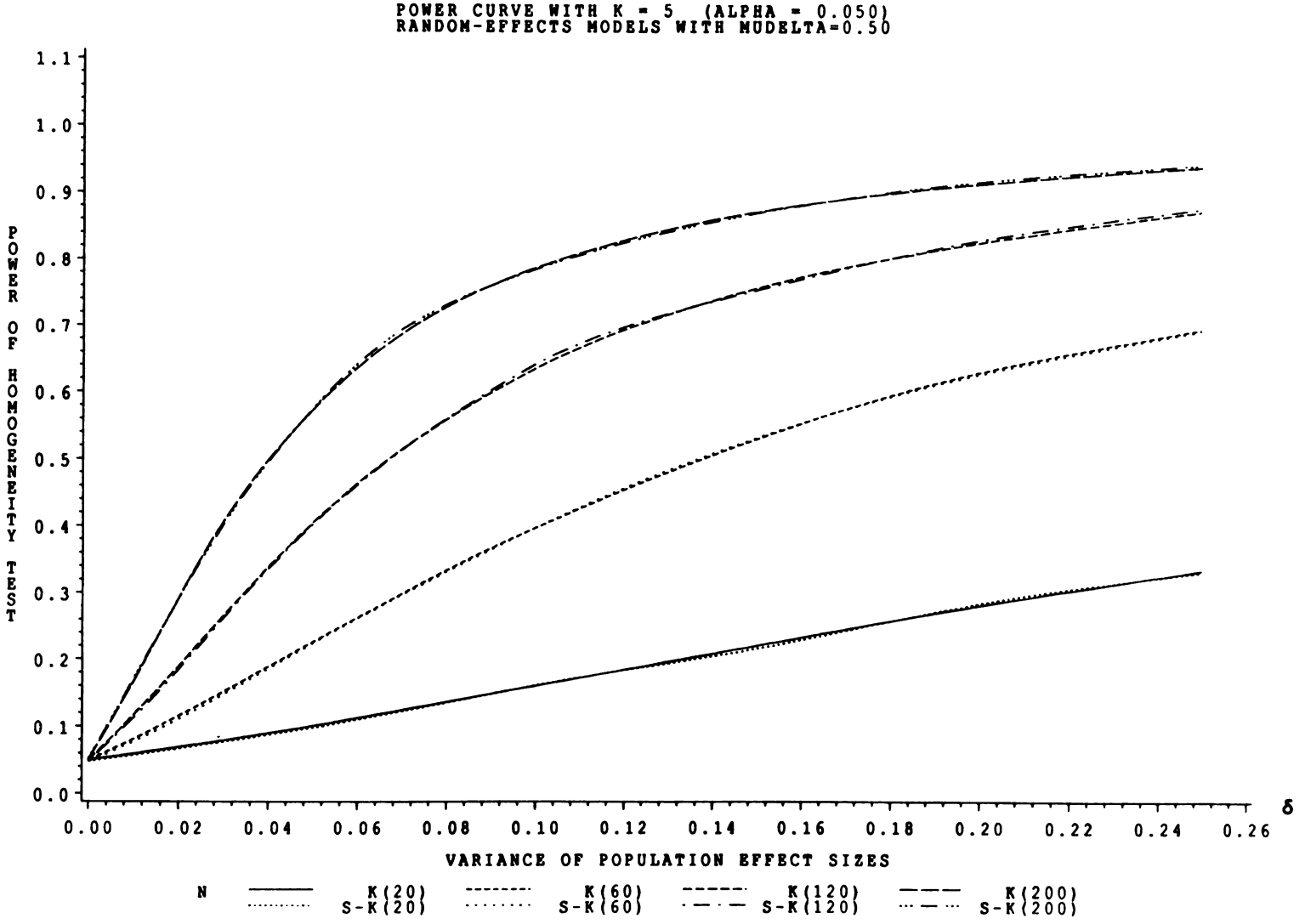
Note: "S-" represents the simulated power curves.

FIGURE 4.8.1



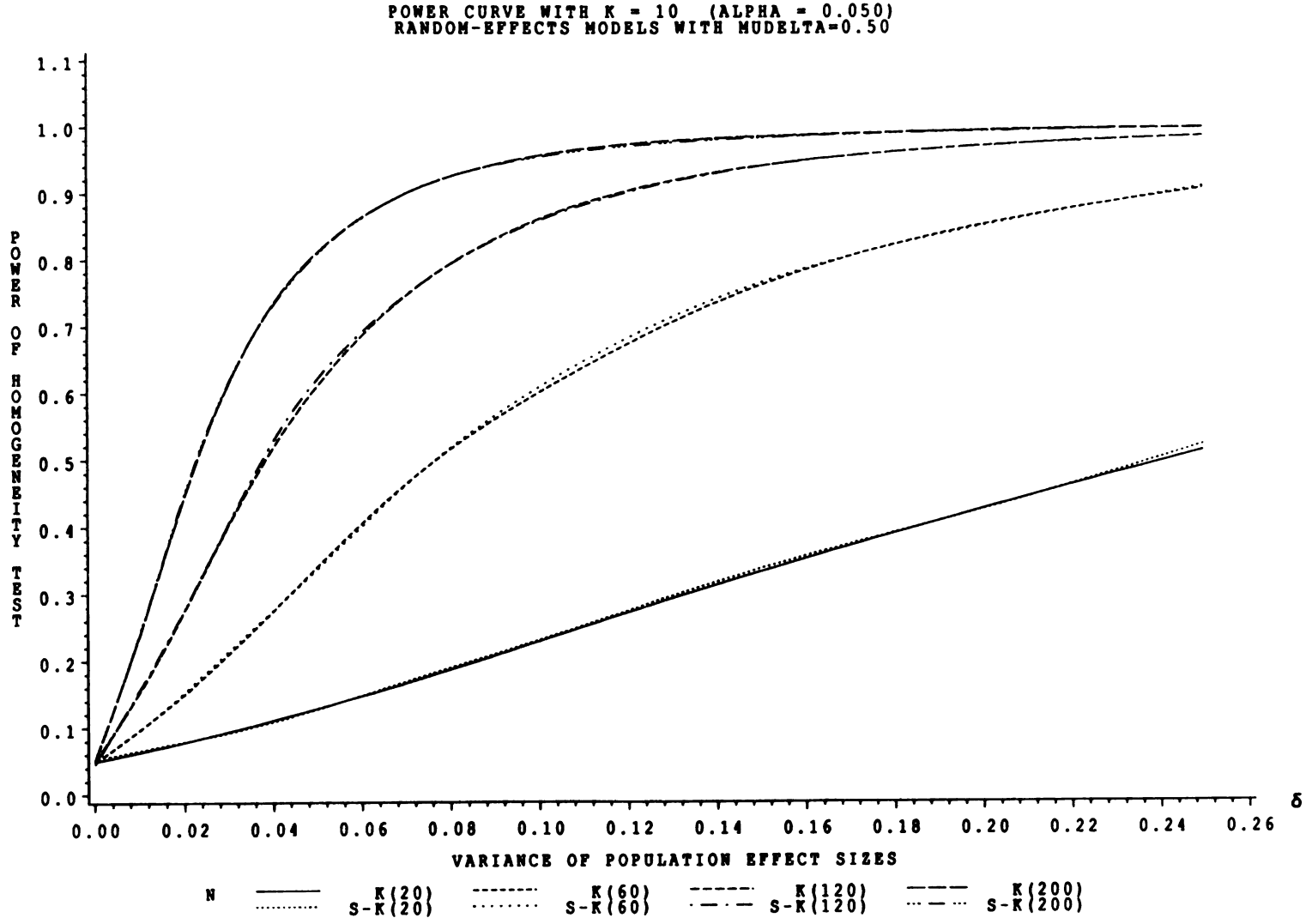
Note: "S-" represents the simulated power curves.

FIGURE 4.8.2



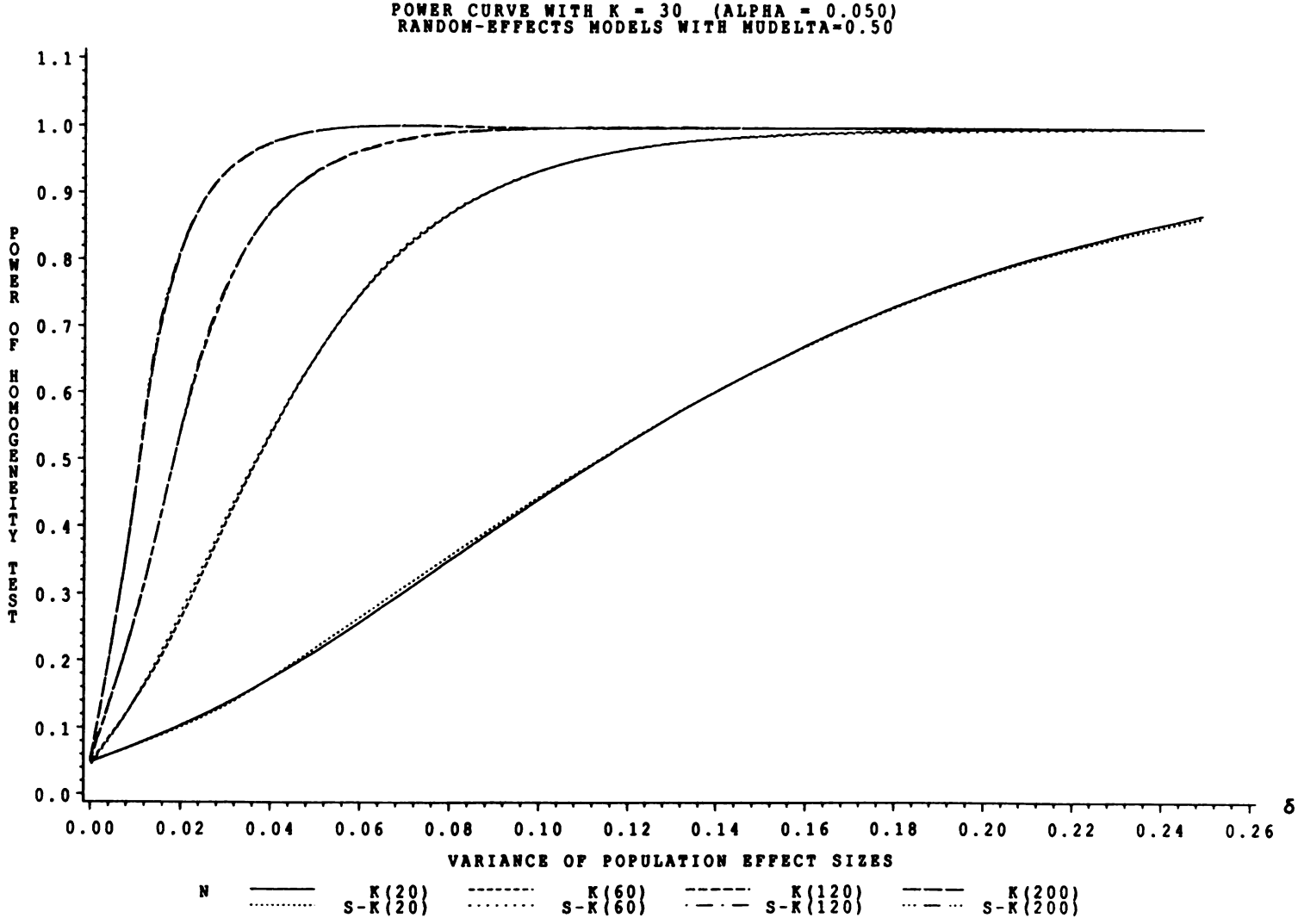
Note: "S-" represents the simulated power curves.

FIGURE 4.8.3



Note: "S-" represents the simulated power curves.

FIGURE 4.8.4



Note: "S-" represents the simulated power curves.

APPENDIX E

SYNTHESIZED STUDIES

Anderson, R. D., Kahl, S. R., Glass, G. V., & Smith, M. L. (1983). Science education: A meta-analysis of major questions. Journal of Research in Science Teaching, 20, 379-385.

Bucknam, R. B., & Brand, S. G. (1983). FBCE really works: A meta-analysis on experience based career education. Educational Leadership, 40:6, 66-71.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 51, 281-309.

Fleming, M. L., & Malone, M. R. (1983). The relationship of student characteristics and student performance in science as viewed by meta-analysis research. Journal of Research in Science Teaching, 20, 481-495.

Horak, V. M. (1981). A meta-analysis of research findings on individualized instruction in mathematics. Journal of Educational Research, 74, 249-253.

Johnson, D. W., Johnson, R. T., & Maruyama, G. (1983). Interdependence and interpersonal attraction among heterogeneous and homogeneous individuals: A theoretical formulation and a meta-analysis of the research. Review of Educational Research, 53, 5-54.

Kavale, K. (1980). Auditory-visual integration and its relationship to reading achievement: A meta-analysis. Perceptual & Motor Skills, 51, 947-955.

Kavale, K. (1981). Functions of the Illinois Test of Psycholinguistic Abilities (ITPA): Are they trainable? Exceptional Children, 47, 496-510.

Kavale, K., & Mattson, P. D. (1983). One jumped off the balance beam: Meta-analysis of perceptual motor training. Journal of Learning Disabilities, 16, 165-173.

Kulik, C. C., Kulik, J. A., & Cohen, P. A. (1979). A meta -analysis of outcome studies of Keller's personalized system of instruction. American Psychologist, 34, 307 -318.

Parker, K. (1983). A meta-analysis of the reliability and validity of the Rorschach. Journal of Personality Assessment, 47, 227-231.

Shapiro, D. A., & Shapiro, D. (1983). Comparative therapy outcome research: Methodological implications of meta -analysis. Journal of Consulting and Clinical Psychology, 51, 42-53.

Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. American Educational Research Journal, 17, 419-433.

Steinkamp, M. W. & Maehr, M. L. (1983). Affect, ability, and science achievement: A quantitative synthesis of correlational research. Review of Educational Research, 53, 369-396.

Steinkamp, M. W. & Maehr, M. L. (1984). Gender differences in motivational orientations toward achievement in school science: A quantitative synthesis. American Educational Research Journal, 21, 39-59.

Sweitzer, G. L., & Anderson, R. D. (1983). A meta-analysis of research on science teacher-education practices associated with inquiry strategy. Journal of Research in Science Teaching, 20, 452-466.

White, K. R. (1982). The relation between socioeconomic status and academic achievements. Psychological Bulletin, 91, 461-481.

Whitley, B. E. (1983). Sex role orientation and self-esteem: A critical meta-analytic review. Journal of Personality and Social Psychology, 44, 765-778.

Willett, J. B., Yamashita, J. J. M., & Anderson, R. D. (1983). A meta-analysis of instructional systems applied in science teaching. Journal of Research in Science Teaching, 20, 405-417.

Willson, V. L. (1983). A meta-analysis of the relationship between science achievement and science attitude: Kindergarten through college. Journal of Research in Science Teaching, 20, 839-850.

Yeany, R. H., & Miller, P. A. (1983). Effects of diagnostic remedial instruction on science learning: A meta -analysis. Journal of Research in Science Teaching, 20, 19-26.

BIBLIOGRAPHY

Alexander, R. A., Scozzaro, M. J., & Borodkin, L. J. (1989). Statistical and empirical examination of the chi-square test for homogeneity of correlations in meta-analysis. Psychological Bulletin, 106, 329-331.

Anscombe, F. J. (1963). Sequential medical trials. Journal of American Statistical Association, 58, 365-383.

Armitage, P. (1960). Sequential medical trials. Oxford: Blackwell Scientific Publications.

Bangert-Drawns, R. L. (1986). Review of developments in meta-analytic method. Psychological Bulletin, 99, 388-399.

Becker, B. J. (1985). Applying tests of combined significance hypotheses and power considerations. (Unpublished Doctoral dissertation, University of Chicago, 1985).

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two meta-analyses. Journal of Research in Science Teaching, 26, 141-169.

Brewer, J. K. (1972). On the power of statistical tests in the American Educational Research Journal, American Educational Research Journal, 9, 391-401.

Chang, L. & Becker, B. J. (1987). A comparison of three integrative review methods: Different methods, different findings? Paper presented at the annual meeting of the American Educational Research Association at San Francisco.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal Psychology, 65, 145-153.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen, J. (1973). Statistical power analysis and research results. American Educational Research Journal, 10, 225-230.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (Rev. ed.). New York: Academic Press.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291-302.

Cronbach, L. V. (1980). Toward reform of program evaluation. San Francisco: Jossey-Base.

Daly, J. A. & Hexamer, A. (1983). Statistical power in research in English education. Research in the Teaching of English, 17, 157-164.

Fabian, V. (1991). On the problem of interactions in the analysis of variance. Journal of American Statistical Association, 86, 362-367.

Fisher, R. A. (1932). Statistical methods for research workers. (4th ed.), London: Oliver and Boyd.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. Educational Research, 5, 3-8.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.

Hedges, L. V. (1982). Fitting categorical models to effect size data. Journal of Educational Statistics, 7, 245 -270.

Hedges, L. V. (1983). A random effect model for effect sizes. Psychological Bulletin, 93, 388-395.

Hedges, L. V. (1986) Estimating effect size from vote counts or box score data. Paper presented at the annual meeting of the American Educational Research Association at Chicago.

Hedges, L. V. & Olkin, I. (1980). Vote-counting methods in research synthesis. Psychological Bulletin, 88, 359-369.

Hedges, L. V. & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando: Academic Press, Inc.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Lewis, R. J. (1990). Sequential clinical trials in emergency medicine. Annals of Emergency Medicine, 19, 1047.

Massey, F. J., Jr. (1956). The Kolmogorov-Smirnov test for goodness of fit. Journal of American Statistical Association, 46, 68-78.

Overall, J. E. (1969). Classical statistical hypothesis testing within the context of Bayesian theory. Psychological Bulletin, 71, 285-292.

Pigott, T. D. (1986). An analogue to analysis of variance for correlations. Paper presented at the annual meeting of the American Educational Research Association at Chicago.

Pillemer, D. B., & Light, R. J. (1980). Synthesizing outcomes: How to use research evidence from many studies. Harvard Educational Review, 50, 176-195.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. Psychological Bulletin, 86, 1165-1168.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Snedecor, G. W. & Cochran, W. G. (1967). Statistical Methods. Iowa: The Iowa State University Press.

Sobel, M. & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. The Annals of Mathematical Statistics, 10, 502-522.

Steinkamp, M. W. & Maehr, M. L. (1983). Affect, ability, and science achievement: A quantitative synthesis of correlational research. Review of Educational Research, 53, 369-396

Steinkamp, M. W., & Maehr, M. L. (1984). Gender differences in motivational orientations toward achievement in school science: A quantitative synthesis. American Educational Research Journal, 21, 39-59.

Tippett, L. H. C. (1931). The methods of statistics (1st ed.). London: Williams and Norgate, Ltd.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

Wald, A. (1952). Sequential Analysis. New York: John Wiley & Sons, Inc.

Whitehead, J. (1983). The Design and Analysis of Sequential Medical Trials. New York: John Wiley & Sons, Inc.

Whitehead, J. (1987). Supplementary analysis at the conclusion of a sequential clinical trial. Biometrics, 42, 461-471.