12-14-2015

Statistical Power in Meta-Analysis

Jin Liu
University of South Carolina - Columbia

STATISTICAL POWER IN META-ANALYSIS

by

Jin Liu

Bachelor of Arts
Chongqing University of Arts and Sciences, 2009

Master of Education
University of South Carolina, 2012

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2015

Accepted by:

Xiaofeng Liu, Major Professor

Christine DiStefano, Committee Member

Robert Johnson, Committee Member

Brian Habing, Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

ACKNOWLEDGEMENTS

I would like to express my thanks to all committee members who continuously supported me in the dissertation work. My dissertation chair, Xiaofeng Liu, always inspired and trusted me during this study. I could not finish my dissertation at this time without his support. Brian Habing helped me with R coding and research design. I could not finish the coding process so quickly without his instruction. I appreciate the encouragement from Christine DiStefano and Robert Johnson. When I feel depressed, they always make me believe in myself. They also offered valuable suggestions for my dissertation.

I am very grateful to my husband, who always supports and loves me. I cannot imagine how I can go through the whole Ph.D. process without him. I also want to thank my friends at the USC who give me encouragement during the process. I also would like to express my gratitude to my parents and sister for everything they have given to me.

ABSTRACT

Statistical power is important in a meta-analysis study, although few studies have examined the performance of simulated power in meta-analysis. The purpose of this study is to inform researchers about statistical power estimation on two sample mean difference test under different situations: (1) the discrepancy between the analytical power and the actual power and (2) the influence of unequal sample size and unbalanced design on the power. Results indicated that there are noticeable discrepancies between the estimated power and actual power under certain conditions. In general, unbalanced design decreases the statistical power in the meta-analysis. Recommendations are provided for researchers who are interested in power of meta-analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

STATISTICAL POWER

A research study usually starts with the development of a significant research question. For example, a school psychologist may want to know whether a certain intervention can help manage children with behavioral problems. A program evaluation specialist may be interested in knowing whether technology usage (e.g., iPad/iPod) in classroom instruction helps improve students' engagement levels. An educational researcher might want to examine the difference in math abilities between boys and girls. After the research question has been developed, researchers develop an appropriate study design to initiate the research study.

Common research designs include experimental studies, survey research, focus group research, and case studies. Most of the time, the entire population cannot be observed due to limited time and resources. A sampling scheme is used to obtain a sample representative of the population. After data are collected from the sample, the researcher will conduct data analysis on the sample to generalize to a larger population and thus, shed new light on the research question (Bhattacherjee, 2012).

There are two branches of research methodology in social science research: qualitative and quantitative methods. Qualitative methodology involves the examination of the data obtained from interviews, observations, and focus group studies. For instance, a qualitative researcher can use diverse coding strategies to categorize the observations, identify themes, and reflect on the research question. With these analysis demands, a

qualitative inquiry does not utilize a large amount of empirical data. On the other hand, quantitative methodology uses numeric data and statistical analysis to summarize information and draw inferences. A mixed method design is to integrate qualitative and quantitative data into one study (Johnson & Onwuegbuzie, 2004). Depending on the nature of the quantitative analysis, the empirical analysis can use descriptive statistics and inferential statistics. The descriptive statistics can be used to summarize the information from the collected data (e.g., frequency distribution, dispersion, and correlation). Nevertheless, the descriptive statistics have limitations because they only provide information based on the sample data and do not allow inferences about the target population. A researcher invariably wants to generalize the results of a study to a much larger population. To draw an inference about the population, a researcher often resorts to inferential statistics and hypothesis testing.

A hypothesis test has a null hypothesis and an alternative hypothesis about the population parameter, which is tied to the research question. A certain statistic (e.g., Z test or t-test) can then be applied to analyze the data and make a decision about the hypotheses. The null and alternative hypotheses are generated with reference to the population of interest. Typically, a researcher hopes to generalize the findings to the population through examining the sample data. That is why it is important to obtain a sample that is representative of the target population. Admittedly, convenience sampling is commonly used in social science studies, which limits the generalizability.

A two-group comparison study provides an illustration of the hypothesis testing (e.g., gender difference in self-efficacy). The null hypothesis (H_0) is defined as the lack of treatment effect or group difference on the continuous outcome. The alternative

hypothesis (H_a) is the direct opposite of the null hypothesis, or the complement of the null hypothesis. It suggests the existence of a treatment effect or some difference between the two genders.

Researchers need to make a decision in the hypothesis testing and, ultimately, provide a yes or no answer to the research question after examining the test statistic. This decision is based on probability theory. It involves examining the likelihood of observing the test statistics by assuming that the null hypothesis is true. For example, in a two-group comparison study with a known population variance, the computed test statistic (i.e., Z) should fall in a certain area of the standard normal distribution. If the Z statistic is not far away from the center of the probability distribution, the test result is deemed as expected under the assumption of the null hypothesis. This does not constitute strong evidence against the null hypothesis. Therefore, the null hypothesis cannot be rejected in this instance.

The null hypothesis can be rejected when the Z statistic deviates from the center of the standard normal distribution. Typically, researchers wish to reject the null hypothesis in order to confirm the treatment effect (a significant mean difference between the two groups). A threshold value (denoted as alpha, α) is used to decide whether the null hypothesis should be rejected or not. The alpha is the maximum probability that a true null hypothesis can be rejected. It is also called the significance level and is traditionally set to .05. This value provides a benchmark for rejecting the null hypothesis and achieving statistical significance. The benchmark is used to calibrate the statistical significance in obtaining a deviant statistic. The probability of obtaining the Z statistic at least deviant from its most expected values under the null is defined as the p-value. If the

p-value is less than .05, it suggests that the test statistic is discrepant enough from the expected value to reject the null hypothesis of no treatment effect.

The rejection of the null hypothesis can occur when the null hypothesis is either true or false, which implies different consequences. The consequences can be illustrated in Table1.1. There are four possible scenarios, regardless of which decision is made. First, there are two possibilities in the population: the lack of a mean difference and the existence of a mean difference. Second, the decision can be either rejecting the null hypothesis or retaining the null hypothesis on the basis of the sample data.

If the null hypothesis is true (the lack of a mean difference), the correct decision is to retain the null hypothesis. If the null hypothesis is not true (existence of a mean difference), the correct decision is to reject the null hypothesis. Although researchers always want to make the correct decisions, they need to acknowledge the possibility of making incorrect ones. There are two types of error because of the existence of the two competing hypotheses and the two possible decisions. Type I error is the probability of rejecting the null hypothesis when it is actually true. Type I error is limited by the significance level (α) or the maximum probability of rejecting a true null hypothesis. The probability of not rejecting a false null hypothesis is Type II error (denoted as beta, β). A false null hypothesis means that the alternative hypothesis is true: there is a treatment effect existing in the population. If the researcher fails to reject the null hypothesis, he or she commits a Type II error. The probability of not making a Type II error when the null hypothesis is false refers to statistical power. As statistical power is inversely related to this error, it can be expressed as 1- β. In other words, statistical power is conceptually defined as the probability of rejecting a false null hypothesis. Researchers often like to

increase statistical power to raise the chances of confirming the possible group differences in hypothesis testing.

In short, Type I error and statistical power are important properties of a hypothesis test. Researchers always aim to control both types of errors. Type I error is traditionally controlled by the significance level. All the hypothesis tests must have a predetermined significance level to limit the Type I error rate. However, research studies vary widely in statistical power. Several factors can influence statistical power. In the following, the determinants of statistical power are described and discussed.

Determinants of Statistical Power

Statistical power is related to the following factors: sample size, population effect size, and the Type I error rate (Borenstein, Hedges, Higgins, and Rothstein, 2010; Cohen, 1988; 1992; Ellis, 2010; Lipsey and Hurley, 2009; Liu, 2013). All of them are discussed in the following paragraphs.

Sample size

Sample size affects the sampling error in a study, and it is one of the most important determinants of statistical power. Unlike other determinants, sample size can be controlled by researchers. A goal of a research study is to find an appropriate sample size to attain the desired statistical power. Increasing sample size is a straightforward way to increase statistical power while other parameters are held constant. However, researchers may not be able to obtain a large sample size, due to high costs and the limited pool of participants (Lipsey & Hurley, 2009). For instance, the attrition of participants in a longitudinal study may influence the final sample size in data analysis and, in turn, the statistical power. Besides the total sample size, other sample size related

factors may influence the statistical power as well. For instance, the unbalanced design between two groups contributes to the loss of statistical power (Hsu, 1994).

Effect size

Another factor is the treatment effect (i.e., group difference) in a study. A large effect size contributes to high power; a small effect size returns low power. The effect size describes the magnitude of the treatment effect (e.g., population mean differences). It is the degree to which the null hypothesis is false (Cohen, 1988, p10). Other things being equal, effect size is positively related to statistical power. For the comparison of the two group mean difference on a continuous outcome (e.g., gender difference in the behavioral and emotional problems), the simple effect size is the mean difference of the outcome between the two groups. Dividing the simple effect size by the common standard deviation yields the standardized effect size, *ES*. The standardized effect size does not depend on the original measurement scale and can be compared across studies.

$$ES = \frac{\mu_1 - \mu_2}{\sigma}$$

The standardized effect size expresses the mean difference in the unit of a common standard deviation. Positive values indicate higher outcomes in first group and negative values indicate higher outcomes in the second group. Cohen (1988) defined .2, .5, and .8 as small, medium, and large effects, respectively, in the behavioral sciences. Standardized effect size is widely used in power analysis. This parameter is influenced by both the mean difference and the variance. For instance, even though the mean differences between groups are large, the standardized effect can be decreased by a large standard deviation. Researchers can target a subpopulation with similar characteristics so that the variation of the outcome (standard deviation) is controlled. Researchers cannot

obtain the effect size from the population directly and they need to estimate that from previously done studies. A minimum detectable effect size may be assumed for the effect size in power analysis (Liu, 2013). Alternatively, researchers can select plausible values based on substantive or clinical importance, and existing data (Borenstein, Hedges, Higgins & Rothstein, 2010).

Significance Level

The significance level defines the risk of committing Type I error (i.e., $\alpha = .05$). It is less likely to reject a true null hypothesis with a lower value, regardless of the used sample size or the actual effect size. Type I error has a negative relationship with Type II error. In other words, a smaller Type I error corresponds to a higher Type II error or a lower statistical power, while other things being equal. Since these two errors are both important, researchers need to pay attention not only to Type I error and but also to Type II error (i.e., statistical power). Lipsey and Hurley (2009) discussed how to set the balance in controlling the two types of error. Researchers need to weigh the relative seriousness of the two errors. The most common approach is to set alpha and beta equal. Borenstein, Hedges, Higgins, and Rothstein (2010) also suggested that researchers should adjust the two errors as appropriate for a given study instead of simply applying the .05 and .8 guideline.

The three determinants are all related to statistical power and all need to be simultaneously considered in power analysis. Researchers typically conduct statistical power analysis to find the necessary sample sizes with respect to the minimum detectable effect size to achieve the desired power (Ellis, 2010). Admittedly, other factors such as

the type of statistical test, the reliability of the outcome measure, and the quality of the study design all influence the statistical power.

Prospective and Retrospective Power

There are two main kinds of power analyses – prospective and retrospective power analysis. Prospective power analysis is a part of research planning and is completed prior to the implementation of a study. It is mostly used to estimate the required sample size with reference to the other parameters in hypothesis testing. For instance, when researchers intend to conduct a replicate study, they need to search the past research to identify the potential population effect sizes. For example, if the Type I error and power are set .05 and .8, the necessary sample needed for error control is determined in power analysis for the research planning. As mentioned above, the two error values can vary if the researchers can justify their standards. Researchers may also check if the past research attained the ideal power, using the sample size from the previous study.

After a study is completed, a retrospective power analysis can be conducted. For example, if researchers cannot reject the null hypothesis but believe in a treatment effect, they may consider low statistical power as a possible explanation for failure to confirm the treatment effect. However, some scholars are cautious about the retrospective power analysis. They suggest that power should not be based on the effect size obtained from the sample due to its possibly large sampling error. The effect size estimate from the sample cannot guarantee a good estimation of the statistical power. The post-hoc power analysis should assume the population effect sizes from previous studies of a similar nature (Thomas, 1997).

History of Statistical Power

Statistical power has not received its due attention in social science research until several decades ago (Borenstein, Hedges, Higgins, and Rothstein, 2010; Cohen, 1988; 1992; Ellis, 2010; Liu, 2013). Researchers appear not to be concerned with statistical power, as the research studies often lack sufficient statistical power. Fisher, who is credited for the creation of the significance test, did not think it was possible to calculate the statistical power. About a century ago, Neyman and Pearson recognized not only the error of rejecting a true null hypothesis (α) but also the error of not rejecting a false null hypothesis (β). Despite the controversial beginning of statistical power, it has been a popular topic in textbook and research articles (Cohen, 1988; Kraemer, Yesavage and Brooks, 1998; Lindsay, 1993; Liu, 2013; Murphy and Myors, 2004; Rossi, 1990). There is no uniform guideline for the desired statistical power across different studies.

The importance of power analysis derives from the fact that investigators always want to reject the null hypothesis, which confirms the existence of a treatment effect (Cohen, 1992). Despite its importance, the current practice of power analysis leaves much to be desired. Statistical power sometimes was not done properly or completely ignored. Some researchers saw little use of conducting power analysis (Mone et al., 1996). Onwuebuzie and Leech (2004) found that statistical power was ranked thirty-fourth out of the thirty-nine topics discussed by the methodology instructors. Such an oversight may be due to the lack of references on statistical power (Nickerson, 2000). In the new century, most people who have experience in statistics should have learned the concept of power, but they may still have difficulty in performing a proper power analysis. The practical difficulty (e.g., unknown population effect size) may have

explained the lack of statistical power analysis in some new fields. Also, there is no strict requirement of including power estimates in a published study unless the scale of the prospective study is large and the funding agency demands a power analysis. In fact, the surveys of statistical power in the social sciences have all indicated insufficient power in published research studies (Cook and Hatala, 2014; Ellis, 2010; Mone et al. 1996). As power is essential to the goals of a research study, power analysis should be required for all the social science studies.

Computation of Statistical Power

Although power analysis is accessible in most of the statistical software (e.g., R and SAS), it is still necessary to understand the formula of power calculation. This will help researchers understand how the parameters in power analysis influence each other so that they can conduct a proper power analysis. In addition, the simulation of statistical power will be introduced because the simulation method can be used to check the accuracy of the power estimates based on approximation. The basic simulation code will be provided to illustrate the steps in the simulation of statistical power here. As power formulas vary from one statistical test to another one, the formula for statistical power in common statistics tests will be discussed in the following. A more comprehensive review of power analysis in different tests can be found in the book *Statistical Power Analysis for the Social and Behavioral Sciences* (Liu, 2013).

Power in a Z test

In a two-group comparison study with a known population variance, the null hypothesis is $\mu_1 - \mu_2 = 0$ and the alternative hypothesis is $\mu_1 - \mu_2 \neq 0$ (two sample

mean difference). The assumption of not knowing the directionality of a test is made, and two-sided test is considered.

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The test statistics Z follows a standard normal distribution when the null hypothesis is true. After the test statistics is obtained, the Type I error rate – probability of rejecting the null when it is true – can be calculated.

$$P = P\left[Z > Z_{1-\frac{\alpha}{2}}\right] + P\left[Z < Z_{\frac{\alpha}{2}}\right],$$

where the $Z_{1-\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$ are the critical values on the two sides for the predetermined significance level (e.g., .05).

Statistical power is calculated under the assumption that the alternative hypothesis is true or there is a mean difference between the two comparison conditions. When the alternative hypothesis is true, the test statistic no longer follows the central distribution (mean of the distribution is 0). Instead, it follows a non-central distribution with a shifted mean related to the population effect size, which is one of the determinants of the statistical power. The non-central distribution can be viewed as shifting the standard normal distribution to the left or right with a different mean but the same standard deviation. To simplify the illustration, the two sample sizes are set to be equal ($n_1 = n_2 = n$) and a common standard deviation is assumed to the same between the two comparison populations ($\sigma_1 = \sigma_2 = \sigma$).

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma}\sqrt{\frac{n}{2}}$$

The non-centrality parameter (λ) can be obtained by substituting the sample estimates with the population parameters in the formula for the Z-test.

$$\lambda = \frac{\mu_1 - \mu_2}{\sigma}\sqrt{\frac{n}{2}}$$

The non-central Z' under the alternative hypothesis is determined by the non-centrality parameter lamda.

Statistical power in a two-sided Z test can be expressed as:

$$1 - \beta = P[Z' > Z_{1-\frac{\alpha}{2}}] + P[Z' < Z_{\frac{\alpha}{2}}]$$

This power value is related to the cumulative probability of rejecting the null hypothesis when it is false. It can be easily calculated with the help of statistical software. The distance between the two distributions is related to the non-centrality parameter and affects the statistical power. Other things being equal, the larger the non-centrality parameter is, the higher the statistical power will be for the significance test. Larger sample sizes and population effect sizes lead to higher λ and higher power. If a less stringent rejection criterion is used, it requires a less deviant Z' to exceed the critical value, which increases the statistical power.

Power in a t test

In practice, the population standard deviation is rarely known, so the t-test is more widely used with the sample estimates of the population standard deviation. As in the Z test, researchers can assume a pooled sample standard deviation for both groups ($\widehat{\sigma_1} = \widehat{\sigma_2} = \hat{\sigma}$).

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

The pooled sample standard deviation is

$$\hat{\sigma} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}},$$

where s_1^2 and s_2^2 are the sample variances of the two groups.

The test statistic follows a *t* distribution, which is influenced by degrees of freedom. The degrees of freedom are equal to $(n_1 + n_2 - 2)$ in a two-sample independent *T* test. The *p*-value of obtaining a t statistics is calculated in a similar way as that in the Z test, except that a *t* distribution has a degree of freedom. It should be noted that a *t* distribution is close to the standard normal distribution for large degrees of freedom. When the degrees of freedom are large, the sample size is large and the sample estimate of the variance is very close to its population counterpart. Similar to the power in a Z-test, the power function for a two-sided *t*-test is

$$1 - \beta = P[T' > t_{1-\frac{\alpha}{2},N-2}] + P[T' < t_{\frac{\alpha}{2},N-2}]$$

Power in an *F* test

In practice, researchers may have more than two groups in comparing mean differences (e.g., ethnicity). An ANOVA analysis can be done to compare those multiple groups, and an *F* test can be used here. Unlike Z-test or t-test, the *F* statistic is used to check if there are any significant differences among the multiple groups. The *F* statistic is the ratio of the average between-group variance and the average within-group variance.

$$F = \frac{Between - group\ Sum\ of\ squares\ /v_1}{Within - group\ Sum\ of\ sqares\ /v_2}$$

The *F* ratio compares the between-groups variation (treatment effects or group differences) with the within-groups variation (random variation among individuals). There are two degrees of freedoms: one in the nominator and the other in the denominator (v_1 is the number of groups – 1 and v_2 is total sample size – the number of groups). Larger *F* values lead to smaller p-values, which mean that the treatment effect appears prominent against the error variation due to individual variation. If the p-value of the *F* statistic is less than the significance level, the null hypothesis will be rejected and the treatment effect is detected.

The statistical power of the *F* test is the probability of obtaining an *F* statistic exceeding the critical value that is used to reject/retain the null hypothesis. The *F* statistic follows central *F* distribution when there are no group differences or treatment effects. Unlike the Z and T distribution, an *F* distribution is not symmetric and is skewed to the right side. When the alternative hypothesis is true, the *F* statistics follows a non-central distribution or *F'* for short. The non-central *F* has the same degrees of freedom as the central *F* statistic and a centrality parameter:

$$\lambda = \sum \frac{n_j \alpha_j^2}{\sigma^2},$$

where σ^2 is the population variance, α_j is the difference between the population mean of the group j and the population grand mean, and n_j is the number of people in each group. The process of obtaining the parameter is beyond the scope of this review but can be found in the related text (e.g., Liu, 2013). The power function for the *F* test can be expressed in terms of the cumulative distribution of the non-central *F*,

$$1 - \beta = P[F'(v_1, v_2, \lambda) \geq F_0] = 1 - P[(v_1, v_2, \lambda) < F_0].$$

The *F* test is used only to examine whether there are any mean differences among all the groups. It does not show how groups might differ between themselves. Researchers usually follow up an *F* test with simultaneous means comparisons (post-hoc tests) to locate any mean differences among the groups. The power of the post-hoc tests is based on the t-test. The t-test statistic can be written as:

$$T = \frac{\overline{Y_j} - \overline{Y_{j'}}}{\hat{\sigma}\sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}},$$

where $\overline{Y_j}$ and $\overline{Y_{j'}}$ are the group means, $n_{j'}$ and nj are the group sizes, and $\hat{\sigma}$ is the root mean square error or the square root of the mean squares for error.

The non-centrality parameter λ is:

$$\lambda = \sqrt{\frac{n_j \mathrm{n_{j'}}}{\mathrm{n_{j'}} + n_j}} \frac{u_j - \mathrm{u_{j'}}}{\hat{\sigma}}$$

The statistical power for the two-sided T test is

$$1 - \beta = (T'(N - J, \lambda) > t_{1-\frac{\alpha}{2m}, N-J} + P(T'(N - J, \lambda) < t_{\frac{\alpha}{2m}, N-J},$$

where *N* is the total sample size, *J* is the number of groups, and m is the number of comparison tests. The computation is similar to the power analysis in a regular T test, except the Bonferroni adjustment is applied to control the family-wise Type I error. However, the Bonferroni adjustment may make researchers hard to reject the null hypothesis. Some other procedures, such as Turkey's HSD test and Dunnett's test, should be considered.

Simulation of Statistical Power

Power functions become rather complicated when more advanced statistical tests are used (Liu, 2013). Sometimes, simulation can be used to avoid computational complexity in power analysis. Simulation studies are widely used in empirical research. The simulation studies involve generating data from computer programs to study the performance of the statistical estimates under different conditions (Hutchinson & Bandalos, 1997). For example, most people can use simulation to learn about Type I error when the model assumptions are not met. This can be done in simulation by generating data under different model assumptions.

The idea of simulation uses the same logic of hypothesis testing. If there is no real effect, researchers hope to retain the null hypothesis most of the time and control the Type I error. If there is a real existing effect, researchers hope to reject the null hypothesis as much as possible. Simulation can be used to check the performance of actual Type I error and power by repeating the same statistical procedures many times under regular model assumptions or under different model assumptions.

Simulations can be conducted with the help of computer software (e.g., R). For instance, in a two sample t test, the effect size can be simulated a certain number of times (e.g., Simultime=1000) by assuming a certain mean (e.g., PopulationEffect=0.2) and standard deviation (e.g., SD=.1). In addition, the sample size is supplied in each repetition (e.g., Samplesize=100). These numbers can vary in practice according to the research settings. An example of such code from R is shown below.

```
PopulationEffect<-0.2
SD<-1
Simultime<-10
Samplesize<-100
```

Repeating the same process many times can help researcher calculate the rejection rate of a test or simulated power. In each repetition, a *p*-value is retained to make a statistical decision (reject or retain the null hypothesis).

```
pv<-rep(NA,Simultime)
for (i in 1: Simultime)
{print (i)
 SimuValues<-rnorm(Samplesize, PopulationEffect, SD)
 pv[i]<-t.test(SimuValues,alternative= "two.sided",
mu=0)$p.value
     }
mean(pv<.05)
```

Finally, 1000 p-values are stored in the output. The same strategy applies to simulated statistical power. The proportion of the rejected null hypotheses among all the simulated tests is the simulated statistical power when the simulated tests assume a non-zero treatment effect. The power under the above condition is around .5. Repeating the process more times can improve the stability of the results. The computing time of the simulation should be considered because time cost is important in a study.

The complete code of the two sample t test is given in the appendix, and it can be adapted to simulate the power in the Z and F tests. Researchers can use computer simulation to compare the discrepancy between the simulated power and the power based on formulas. The simulation can provide an easy and direct way to cross check the power based on the analytical formulas with the observed power obtained from the simulated studies. In particular, simulation can be utilized to check the accuracy of statistical power in meta-analysis, which is based on the approximate formulas in the literature. The research questions will be stated clearly after the introduction of meta-analysis in Chapter 2.

Table 1.1 *Decision Making in a Hypothesis Test*

Decisions from the hypothesis testing of the sample data	Truth (Population)	
	H_0 True (No effect & difference)	H_0 False (Real effect & differences)
Retain H_0	Correct Decision	Type II error (β)
Reject H_0	Type I error (α)	Correct Decision (Power)

CHAPTER 2

META-ANALYSIS

Meta-analysis is a quantitative review method, which synthesizes the results of several studies on the same topic. In a meta-analysis, a researcher combines the effect size estimates from a set of small studies to get a common effect size estimate (i.e., the direction and magnitude of the treatment effect). Thus, meta-analysis has the potential to overcome the shortcomings of a single primary study because a small primary study can be limited in sample size, estimate precision, and generalizability (Ellis, 2010; Hedges & Pigott, 2001).

Limitations of Primary Studies and Narrative Review

In quantitative research, numerous studies use primary samples collected on a small scale. Due to the time and resource constraints, there are always some limitations of those primary studies of small size, that is, the lack of generalizability of the findings and the low level of statistical power.

Generalizability or external validity refers to the extent, to which the study results can be generalized to a broader setting (Trochim, 2000). The generalizability of a study may be limited in a small primary study because of the specific sampling strategies involved. It is well-known that a primary study can use either a probability sample (e.g., simple random sampling) or a non-probability sample (e.g., convenience sampling). If it is a probability sample, the researchers can generalize the conclusion of a study to a larger population, from which the samples were randomly selected. If a study uses a

convenience sample, the study results, however, might not generalize to a larger population. Despite this flaw, convenience sampling or other non-probability sampling strategies are still used in practice. Even when a probability sample is used, a primary study has other challenging issues. For example, the problem of small sample size can be exacerbated if there is anticipated participant drop-out like that in longitudinal studies (Hogan, Roy & Korkontzelou, 2004). A small sample size can lower the statistical power in testing the treatment effect.

As discussed in Chapter 1, power analysis is often overlooked in social science research (Bezau & Graves, 2001), and the surveys of power in different social science fields indicate that many published articles have low statistical power (Ellis , 2010). Given the fact that the published research has more significant results than the unpublished research, the actual power of unpublished studies could have been even lower in practice. As noted by Cafri, Kromrey and Brannick (2010), the power in social science research generally did not have sufficient power to detect small and medium effect sizes in the populations. In addition, researchers may not be able to collect data to reach the ideal statistical power, due to the time and logistic constraints. In fact, low statistical power is often the explanation of inconclusive conclusions among small primary studies.

A qualitative review can be used to inquire about inconclusive study results. Researchers can review the literature to "circumscribe the boundaries of existing knowledge and to identify potential avenues for further inquiry (Ellis, 2010, p. 90)." However, the narrative summaries of past research cannot overcome the shortcomings of small primary studies, especially the issue of low statistical power. First, researchers have

practical difficulties, such as no access to certain resources, including all studies of the target topic. They cannot obtain a satisfactory level of generalizability. Even though researchers can assume that they include all studies, they will not be able to address the concern over low statistical power. Meta-analysis allows researchers to overcome the limitations of a qualitative review. They can use meta-analysis to combine the effect size estimates and reconcile the inconsistent findings across a large number of small studies (Hunter & Schmidt (2004).

Advantages of Meta-analysis

Meta-analysis is better suited to addressing the limitations of small primary studies or qualitative reviews. Scholars first noted the importance of developing strategies of meta-analysis almost forty years ago. Glass (1976) first introduced the method:

> Most of us were trained to analyze complex relationships among variables in the primary analysis of research data. But at the higher level, where invariance, non-uniformity and uncertainty are no less evident, we too often substitute literary exposition for quantitative rigor. The proper integration of research requires the same statistical methods that applied in primary data analysis. (p. 6)

Reference books have been written on the subject of meta-analysis. In *Statistical Methods for Meta-analysis*, Hedges, and Olkin (1985) suggested that meta-analysis could address the two issues that could not be solved in conventional studies: (1) the impossibility of testing the inconsistency across studies and (2) the impossibility of conducting a test for the average effect size of studies.

Hunter and Schmidt (2004, p. 16) stated that meta-analysis can help improve the limited generalizability of primary studies and summarize research literatures to form a

cumulative knowledge base. For example, researchers can use meta-analysis to broaden the applicability of the findings. Also, meta-analysis can suggest directions for new research.

Borenstein, Hedges, Higgins and Rothstein (2009) noted that the goal of a synthesis is to understand the results of any study in the context of all other studies. There are two fundamental changes in meta-analysis: (1) we work directly with effect size in each study instead of p-value; (2) we include all of the effects in a single statistical synthesis. This is critically important for the goal of computing a summary effect size, while any narrative reviews cannot provide any means to synthesize such data.

Ellis (2010) discussed the advantages of a meta-analysis over a narrative review. He listed several benefits of using a meta-analysis:

(1) Bring a high level of discipline to the review process. It is a more objective process.

(2) Cumulating data (effect size) instead of conclusions (p-value).

(3) Provide definitive answers to questions regarding the nature of an effect even in the presence of conflicting findings

(4) Work as a tool for theory development and a guide for future research.

Meta-analysis can also increase statistical power in testing a treatment effect. Borenstein, Hedges, Higgins and Rothstein (2009) and Liu (2013) demonstrated why meta-analysis could increase statistical power when compared with a single study. The reason for the increased power can be simply explained by the fact that the combined data from small studies increase the overall sample size in a meta-analysis. The increased

total sample size in a meta-analysis helps increase the non-centrality parameter, which in turn improves statistical power in testing the treatment effect.

Effect Size in Meta-analysis

A key concept in meta-analysis is the effect size. Researchers decide to reject or retain the null hypothesis, based on the p-value of a test statistics (statistical significance), while they use effect size to measure the magnitude of an effect, sometimes referred to as the practical significance of a test. As suggested by Cohen (1990, page 1310), "the primary product of a research inquiry is one or more measures of effect size, not p values." Effect size is not only important in the primary studies but also critical in meta-analysis as scholars combine the effect size from studies to get an average estimate of the treatment effects across studies. Ellis (2010) included a good summary of different kinds of effect sizes. There are two major families of effect size: d (e.g., odds ratio, Cohen's d; differences between groups) and r (e.g., Pearson correlation, Cohen's f; measure of association). The current study focuses on two group differences in continuous outcomes.

Two kinds of conceptual models can be employed in meta-analysis. They are formulated, according to the property of the effect sizes in individual studies. A fixed-effects model treats the population effect sizes from individual studies as the same. In other words, there is a common population effect size across studies in the fixed-effects model. By contrast, a random-effects model treats the population effect sizes from individual studies as a random sample of all possible effect sizes with an underlying distribution (e.g., normal distribution). In a fixed-effects model, the only reason the effect size varies is the random error. In a random-effects model, the effect size can be influenced by random error and the effects of different studies. While discussing model

selection, Hedges and Vevea (1998) stated that fixed-effects models are designed to make inferences about a population exactly like the studies sampled, but random-effects models are designed to draw inferences about a population that may be not exactly identical. The fixed-effects models were used frequently in practice, and the random-effects models have seen increased use over time (Cafri, Kromrey & Brannick, 2010). Until now, there are no absolute guidelines for model selection. However, the model selection does affect how the effect size indexes are combined in the meta-analysis.

In meta-analysis the effect sizes are always combined to compute the average effect size and its variance, which form a statistic test (i.e., Z-test). One can then use the test statistic to make a decision about retaining or rejecting the null hypothesis about the average effect size. In the following, the fixed-effects and random-effects meta-analysis are described in details for easy reference.

Analytical Procedures

The first step of a meta-analysis is to define the research questions and the study design. Researchers need to perform a comprehensive literature review to include and summarize the studies for the meta-analysis. Well-formulated research questions and thorough literature review contribute to the high quality of a meta-analysis study. Once the information from primary studies has been processed, researchers need to identify a common measure to all studies and combine the effect sizes from individual studies (Normand, 1999).

To investigate the estimation and power function of meta-analysis, we first review the basic analytical procedures in fixed-effects and random-effects meta-analyses (Borenstein, 2009; Hedges, Borenstein, Hedges, Higgins & Rothstein, 2010). The

analytical procedures and the power calculations are explained for both the fixed-effects meta-analysis and the random-effects meta-analysis later in this chapter.

The two-group mean difference is used as the effect size index for several reasons. First, it is a widely used effect size index in practice. For instance, many researchers are interested in the gender differences on certain continuous outcomes, such as achievement levels and behavioral problems. Similar studies can be found online easily. If we search the "meta-analysis" and "gender difference" through the ERIC and Education Resource database, there are 389 results. A certain amount of them used continuous outcomes. Other similar two group tests with continuous outcomes can be searched online too. Hattie (2009) reviewed over 800 meta-analysis related to achievement using the effect size index d, which also indicates that the popularity of this index. Secondly, few simulation studies have been conducted to analyze the performance of this effect size index.

Cohen's d is used as the effect size index of each study to investigate the mean differences across groups. Cohen's d is used frequently when there is a continuous outcome for two groups of subjects, such as treatment and control groups in the experimental design. For example, female and male students naturally form two comparison groups. This kind of analysis is widely used in applied research in social science research. The formula to calculate Cohen's d (e.g., Ellis, 2010; Liu, 2013) is:

$$\text{Cohen}'\text{d} = \frac{\overline{\text{X}_1} - \overline{\text{X}_2}}{s_p}.$$

where $\overline{\text{X}_1}$ and $\overline{\text{X}_2}$ are the sample means for two groups, and s_p is the pooled standard deviation of two groups.

It is noted that the assumption of pooled standard deviation is not always met in practice especially when the sample size between two groups are not balanced. In addition, d tends to overestimate the population variance. The bias can be removed by Hedge's g, which weights the standard deviation by its sample size (Hedges, 1981). It can be converted from d using the following correction factor (Borenstein, Hedges, Higgins, and Rothstein, 2010):

$$J = 1 - \frac{3}{4df - 1}$$

Where the degree of freedom is the overall sample size – 2.

$$\text{Hedge's g } = J * \text{ d}.$$

One of the major meta-analysis methods were developed by Hedges and his colleagues (Hedges & Olkin, 1985; Hedges & Vevea, 1998). The analytical power formulas were developed by Hedges and Pigott (2001). The fixed and random-effects model were discussed separately.

Fixed-effects Meta-analysis

The common effect size estimate for the *i*th individual study is equal to the standardized mean difference between the treatment condition and control condition (Cohen, 1988)

$$\text{d}_i = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p}.$$

In this formula $\bar{Y}_1$ and $\bar{Y}_2$ are the means for two groups, and s_p is the pooled standard deviation in a two independent sample t- test. The effect size estimate d_i corresponds to a population effect size of θ_i.

I denote t_i, $\bar{n}_1$, and $\bar{n}_2$ as the reported t statistics, the treatment group size, and the control group size of the *i*th study in a meta-analysis,

$$t_i = \frac{\bar{Y}_e - \bar{Y}_c}{s_p\sqrt{\frac{1}{\bar{n}_{1i}} + \frac{1}{\bar{n}_{2i}}}}.$$

Thus, d_i can be expressed in terms of the *t* statistic,

$$d_i = t_i\sqrt{\frac{1}{n_{1i}} + \frac{1}{n_{2i}}}.$$

The effect size d_i is assumed to have an underlying T distribution with mean of θ_i and variance of v_i. According to Hedges and Olkin (1985), the variance term is known to be

$$v_i = \frac{n_{1i} + n_{2i}}{n_{1i}n_{2i}} + \frac{d_i^2}{2(n_{1i} + n_{2i})}.$$

The corrected variance of Hedge's g is

$$v_{gi} = J^2 * v_i.$$

The null hypothesis for the population effect size for each individual study is $\theta_1 = \theta_2 \ldots = \theta_i = \cdots = \theta = 0$. The fixed-effects model becomes

$$d_i = \theta + e_i,$$

where e_i has a mean of zero and variance of v_i. The common effect size can be estimated by pooling the estimates from individual studies, where the effect size estimates from those studies are weighted by the sampling variances of individual studies. An effect size estimate from a study with a larger sample size will receive more weight because the estimate is more precise with a smaller sampling variance. The weight w_i is the

reciprocal of the variance term v_i ($w_i = 1/v_i$). The estimate of common effect size is a weighted average.

$$\hat{\theta} = \bar{\mathrm{d}} = \sum_{i=1}^{I} w_i d_i \,/ \sum_{i=1}^{I} w_i$$

The variance of the weighted average v or Var($\bar{\mathrm{d}}$) is simply the reciprocal of the sum of weights.

$$v = 1/\sum_{i=1}^{I} w_i$$

An approximate Z-test can be used to test the null hypothesis that the common effect size θ is zero, using the weighted average estimate.

$$Z = \frac{\bar{\mathrm{d}} - 0}{\sqrt{v}}$$

The *p*-value in a two-sided test is the probability of obtaining a z statistic at least deviant from the center of the standard normal distribution as the computed one. A small *p*-value less than or equal to five percent will result in the rejection of the null hypothesis, which is followed by pronouncement of a non-zero common effect size. A confidence interval can be computed to accompany the significance test for the common effect size. The 95% confidence interval for the common effect size is estimated as:

$$\bar{d} \pm 1.96 * \sqrt{v}.$$

When the alternative hypothesis is true, the common effect size is equal to a non-zero constant θ_a. The Z test follows a non-central normal distribution Z' with a non-centrality parameter λ:

$$\lambda = \frac{\theta_a}{\sqrt{v_{\theta_a}}}.$$

The current procedure assumes a common variance of all studies ($\bar{v}_i$) to simplify the v for power computation because it can greatly simplify the variance formula. If variances of all the studies are thought be approximately equal, that is, $v_1 = v_2 \ldots = v_i = \cdots = v_I$. It is noted that this an ideal assumption, because the variance of all studies are not identical. The variance v can be simplified to

$$v_{\theta_a} = \frac{\bar{v}_i}{I},$$

where $\bar{v}_i$ is the average of overall variance for all studies. $\bar{v}_i$ can be computed by using the average sample sizes for n_{ei} and n_{ci} and the estimated $d_i = \theta_a$. The variance thus computed is an approximation of the actual variance (Hedges & Pigott, 2001),

$$\lambda = \frac{\theta_a}{\sqrt{v_{\theta_a}}} \approx \frac{\theta_a}{\sqrt{\frac{\bar{v}_i}{I}}} = \frac{\sqrt{I}\theta_a}{\sqrt{\bar{v}_i}} .$$

In order to simplify the calculation, the treatment group and control group size are assumed to be equal ($\bar{n}_{1i} = \bar{n}_{2i} = n$):

$$\bar{v}_i \approx \frac{\bar{n}_{1i} + \bar{n}_{2i}}{\bar{n}_{1i}\bar{n}_{2i}} + \frac{\theta_a^2}{2(\bar{n}_{1i} + \bar{n}_{2i})} .$$

The non-centrality parameter in the meta-analysis can be changed to

$$\lambda = \frac{\sqrt{I}\theta_a}{\sqrt{\frac{2}{n} + \frac{\theta_a^2}{4n}}},$$

where θ_a is the standardized mean difference common to all individual studies. The term $\theta_a^2/4n$ is very small, especially when the population effect size (θ_a) is small and the sample size for each group (n) is large. Dropping the negligible term in λ yields

$$\lambda \approx \sqrt{I}\theta_a\sqrt{\frac{n}{2}}.$$

The power function for the two sided test is, therefore,

$$1 - \beta \approx P[|Z'(\lambda)| \geq Z_0]$$

$$= 1 - \Phi(Z_0 - \lambda) + \Phi(-Z_0 - \lambda).$$

Random-effects Meta-analysis

In the random-effects model the effect size estimates from individual studies have an underlying distribution. The effect size estimate d_i follows a normal distribution with mean of θ_i and variance of v_i, that is,

$$d_i = \theta_i + e_i.$$

The parameter θ_i has an underlying distribution with a mean θ and a variance of τ. It is assumed that the population effect sizes from individual studies follows a normal distribution. Unlike the fixed-effects model, the random-effects model suggests that the effect sizes bounce around the grand average effect size θ. Thus, d_i becomes

$$d_i = \theta + \alpha_i + e_i.$$

The random effect α_i is due to different individual studies with its variance τ. The random effect e_i is the sampling error of d_i with its variance of v_i.

The random-effects model can be reformulated so that the same procedure can be applied the fixed-effects model. The random-effects α_i and e_i can be combined into a single error term e_i^*. Thus d_i becomes

$$d_i = \theta + e_i^*,$$

where $e_i^* = \alpha_i + e_i$ and $v_i^* = Var(e_i^*) = v_i + \tau$. Now the random-effects model can be treated as a special case of the fixed-effects model with a more complex variance v_i^*. An approach that is similar to that used for the fixed-effects can be followed. The weight w_i^* in the random-effects model is the reciprocal of the variance term v_i^* ($w_i^* = 1/v_i^*$).

The weighted mean of the random-effects model can be computed:

$$\bar{\mathrm{d}} = \frac{\sum_{i=1}^{I} w_i^* d_i}{\sum_{i=1}^{I} w_i^*}.$$

The variance of $\bar{\mathrm{d}}$ is:

$$v^* = \frac{1}{\sum_{i=1}^{I} w_i^*},$$

where v_i is the same way estimated before. Hedge's g correction is used for the random-effects model is similar as the fixed-effects model. The variance τ can be estimated, according to Hedges and Vevea (1998):

$$\tau = \frac{Q - (k-1)}{c},$$

where $Q = \sum_{i=1}^{k} w_i \, (d_i - \bar{\mathrm{d}})^2$ and $c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} {w_i}^2}{\sum_{i=1}^{k} w_i}$.

An approximate Z test can be used to test the null hypothesis$(\theta = 0)$, based on the weighted average estimate:

$$Z = \frac{\bar{\mathrm{d}} - 0}{\sqrt{v^*}}.$$

A small *p*-value less than or equal to five percent will result in the rejection of the null hypothesis, which is followed by declaration of a non-zero common effect size. A confidence interval can be computed to accompany the significance test for the common effect size. The 95% confidence interval for summary effect is estimated as

$$\bar{d} \pm 1.96 * \sqrt{v^*}.$$

Under the alternative hypothesis, the common (grand average) effect size is equal to a non-zero constant θ_a. The Z test follows a non-central normal distribution Z' with a non-centrality parameter λ,

$$\lambda = \frac{\theta_a}{\sqrt{v^*{}_{\theta_a}}}.$$

Some conjectures are needed to approximate $v^*{}_{\theta_a}$. One assumes that the sample sizes are equal among individual studies, following Hedges and Pigott (2001). So one obtains $v_1^* = v_2^* \ldots = v_i^* \ldots = v_I^*$. The variance v^* can be computed as:

$$v^*{}_{\theta_a} = \frac{\bar{v}_i^*}{I}.$$

Then the non-centrality parameter can be rewritten as

$$\lambda = \frac{\sqrt{I}\theta_a}{\sqrt{\bar{v}_i^*}},$$

where the overall variance for all studies is equal to

$$\bar{v}_i^* = \bar{v}_i + \tau \approx \frac{\bar{n}_{1i}+\bar{n}_{2i}}{\bar{n}_{1i}\bar{n}_{2i}} + \frac{\theta_a^2}{2(\bar{n}_{1i}+\bar{n}_{2i})} + \tau .$$

The λ in the random-effects model is usually smaller, compared with the non-centrality parameter in fixed-effects model. The ratio of $\bar{v}_i$ (within-study variance) and τ (between-study variance) can be denoted by $p = \tau/\bar{v}_i$. Thus the non-centrality parameter can be expressed in this way,

$$\lambda = \frac{\sqrt{I}\theta_a}{\sqrt{\bar{v}_i + \tau}} = \frac{\sqrt{I}\theta_a}{\sqrt{\bar{v}_i(1+p)}}.$$

After setting up the p ratio, lamda can be calculated in the same way as in the fixed-effects model. Although the random-effects model makes it easy to generalize the research findings to a broader context than the fixed-effects model, the fixed-effects meta-analysis tends to have higher power than the random-effects meta-analysis. The power function for a two-sided test is the same as the fixed-effects model:

$$1 - \beta \approx P[|Z'(\lambda)| \geq Z_0]$$

$$= 1 - \Phi(Z_0 - \lambda) + \Phi(-Z_0 - \lambda).$$

It should be noted that Hunter and Schmidt (2000, 2004) advocate a random-effects model, based on the belief that a fixed-effects model is often inappropriate for real-world data and can limit the generalizability of the findings in a meta-analysis study. However, they apply a slightly different analytical procedure in the meta-analysis. They still use the Z statistic to test the significance of combined effects size, but they weigh the effect sizes by the sample sizes of the individual studies instead of the variances of the studies. In other words, the larger the sample size, the more weight the study will receive in the combined effect size estimate.

$$\bar{\mathrm{d}} = \frac{\sum_{i=1}^{I} w_i^* d_i}{\sum_{i=1}^{I} w_i^*},$$

where $w_i^* = n_i$, and d_i of each study was calculated as the above methods. In addition, there is a different formula to calculate the combined variance,

$$v^* = \frac{\sum w_i^* \left[d_i - \bar{\mathrm{d}}\right]^2}{\sum w_i^*}.$$

Cited by Ellis (2010, p.150), the variance term should be corrected by dividing v^* by the number of studies in a meta-analysis. To calculate the test statistics, a similar procedure is followed to calculate the value of the Z statistic.

Comparisons of different models and methods have been reviewed in the literature. For example, Field (2001) investigated the random-effects meta-analysis in combining correlation coefficients, and he found that Type I error for both strategies was not controlled for small number of studies (<15) in the heterogeneous case (population effect size is not fixed). The fixed-effects model caused biased results if the real data contained varied population effect sizes across studies (Field, 2003). Homogeneity test

(Q statistics) could be used for model selection, but Broenstein (2009, p. 84) suggested that the decision should be based on the understanding of whether or not all studies shared a common effect size rather than on the outcome of a statistical test.

In this chapter I first conducted meta-analysis, using two real datasets obtained from an online database. Second, I used the real data sets to estimate the parameters for power analysis. Using the parameter estimates, I showed how to compute the statistical power and discuss some issues surrounding low statistical power in meta-analysis. Third, I described how to simulate the statistical power in meta-analysis.

Challenges in Meta-analysis

Although meta-analysis is an objective process of synthesizing studies, there are subjective decisions to make in the process. The results can be biased if the following issues are not handled appropriately: (1) exclude relevant research; (2) include bad results; (3) use the inappropriate statistical models and methods; and (4) complete analysis with insufficient statistical power (Ellis, 2010).

Ellis (2010) recommended that the first step in meta-analysis is to select "good" studies, based on the well-defined research topic. Excluding relevant research (e.g., publication bias) or including low quality studies may lead to biased results. However, it may be difficult to include studies that are not published. The quality of a study is sometimes difficult to judge if the information on sampling and implementation are not available.

Researchers have expressed their concerns over power in meta-analysis. Cafri, Kromrey, and Brannick (2010) asserted that "power analysis is more important in meta-analysis because such studies summarize similar research and influence more on theory

and practice." Field (2001) investigated different meta-analytical models for correlation coefficient studies, and he examined the procedures that produced the most accurate and powerful results under different conditions. Cohen and Becker (2003) demonstrated how meta-analysis could increase statistical power. In their study, three indices were examined: standardized mean difference, Pearson's r, and odds ratio. Statistical power could be increased by reducing the standard error of the weighted effect size. However, the number of studies would not always increase statistical power and the between-study variance should be considered under the random-effects model. Stern, Gavaghan and Egger (2000) found that power was limited in meta-analyses, based on a small number of individual studies. In this case, results should be interpreted with great care. Thus, statistical power in meta-analysis has great implications for the study result, which is the main focus of the current study.

Meta-analysis Application

The low statistical power in a meta-analysis may be due to the small number of studies or the low minimum detectable population effect size. To illustrate this issue, I will use real meta-analysis data to estimate the parameter values and assess the statistical power in the context.

The first dataset came from studies on gender differences in mental rotation and cognitive abilities (Voyer, 2011). The previous research had well documented that men were better at mental rotation and cognitive abilities, as compared with women. Six studies were included in the meta-analysis to examine the gender differences in mental rotation tasks with long time limits. Table 2.1 showed the summary information of each study including the sample size of each group and the standardized mean difference of

each group. The fixed-effects method and the two random-effects methods were used as introduced before in this chapter. Table 2.2 displayed the summary results of the three models. Three analyses indicated the same conclusion that men were better at the tasks than women. The Q statistics (4.23, p=.52) did not show that the heterogeneity among groups was statistically significant. The between-study variance was zero because Q statistics is smaller than the number of studies. With a large combined effect size and a small amount of heterogeneity, it was easy to find statistical significance even with a small number of individual studies. It did not matter whether the fixed-effects or random-effects model was used. Because the between study variance was zero, the test statistics value is the same for both random-effects and fixed-effects models. The final results were basically the same across models. In other words, when the heterogeneity among groups was small, there were no big differences among models.

The second dataset came from studies of children's self-conscious emotions (Else-Quest, Higgins, Allison & Morton, 2012). The research on gender stereotypes of emotion suggested that men experienced more pride than women. Table 2.3 included the summary information of each study including the sample size of each group and the standardized mean difference of each group.

A fixed-effects model and two random-effects models were employed in the analysis. Table 2.4 displayed the summary information of the three models. The two random-effects models produced the same conclusion that there were no gender differences, which was different from the conclusion from the fixed-effects model. Even though the fixed-effects model result indicated statistical significance, the model may not be appropriate for this dataset due to the high value of Q statistic (250, $p < .05$) Thus, a

random-effects model was warranted. However, the consensus in the literature is that men were more prideful. In other words, the gender difference did exist, and the effect size was not zero. If researchers believed there was a real difference between the gender effects, they might attribute the lack of statistical significance in the meta-analysis to low statistical power.

Most of the meta-analyses have enough statistical power but there is no guarantee that a meta-analysis will not lead to Type II error, just as is sometimes the case with individual studies (Ellis, 2010). Thus, the power investigation in meta-analysis is still worthwhile.

Research Questions

It is important to investigate the factors that influence statistical power in meta-analysis. The current study seeks to extend the previous research to gauge the performance of statistical power in meta-analysis (two-group differences on continuous outcomes) under various conditions such as the number of studies, the sample sizes of individual studies, and the between-study variances. In particular, the current study will compare the estimated power with the simulated power, which is designed to be the actual power in meta-analysis. The comparison study will yield a better understanding of power in real meta-analyses. For instance, if there is a discrepancy between the estimated power and the simulated power, researchers may consider adjusting the estimation values to compensate for the differences between the estimated power and the real power in planning a meta-analysis.

Areas in power analysis for meta-analysis to be investigated are included in the following:

1. Many meta-analyses use standardized mean difference to combine study results, but few simulation studies have focused on investigating the real statistical power based on this index.
2. The extant literature provides the approximate formulas for computing the statistical power, although its accuracy has never been thoroughly vetted. Hedges and Pigott (2001) showed the power functions through approximation, which is, averaging the variance across studies as the combined variance estimate. The discrepancy between estimated power and actual power is of importance.
3. Sample sizes influence the power estimates. In meta-analysis, the sample sizes of small studies vary from one study to another. Sample size difference between two groups also influences the statistical power. Some studies included in a meta-analysis do not have balanced designs with equal sample sizes between the two groups. The sample ratio between the two groups may influence the statistical power. Thus, the degree to which unequal sample sizes affect the statistical power in meta-analysis will be investigated. The other factors, such as number of studies and population effect size, will be included for consideration as well. Mainly, these factors are under the researchers' control and are the main focus of the study.

The current study is intended to simulate statistical power in meta-analyses under more realistic conditions. Therefore, I pose two broad research questions:

1. Is there any discrepancy between the approximate power and the simulated power in fixed-effects and random-effects meta-analyses?
2. How do unequal sample sizes across studies and unbalanced designs within a

study of studies affect statistical power in the fixed-effects meta-analysis and the random-effects meta-analysis?

Recommendations will be given for practical researchers who are interested in power of meta-analysis.

Table 2.1 *Effect Sizes and Sample Sizes of Studies for Mental Rotation Tasks*

Study No.	Female	Male	D
1	53	32	0.79
2	117	97	0.67
3	153	106	0.85
4	63	43	0.68
5	431	312	0.95
6	29	48	0.74

Table 2.2 *Summary Results of Meta-analysis across Methods*

Methods	Average Effect Size	95% CI	Z statistics (p value)
Fixed-effects Model	.85	[.74,.96]	15.47 (p<.01)
Random-effects Model -1	.85	[.74,.96]	15.47 (p<.01)
Random-effects Model -2	.85	[.76,.94]	18.63(p<.01)

Note: Random-effects Model -1 – Hedge and Colleagues;
Random-effects Model -2 - Hunter & Schmidt.

Table 2.3 *Effect Sizes and Sample Sizes of Studies for Pride*

Study No.	Male	Female	D
1	515	308	0.44
2	22	139	0.44
3	30	142	-0.03
4	39	130	0.24
5	38	85	0.3
6	20	73	-0.13
7	61	29	-0.08
8	809	1802	0.3
9	99	285	-0.04
10	97	192	-0.34
11	26	72	0.44
12	44	35	0.05
13	814	1513	0.36
14	129	219	0.44
15	300	699	0.45
16	616	1432	-0.43
17	148	190	-0.13

Table 2.4 *Summary Results of Meta-analysis across Methods*

Methods	Average Effect Size	95% CI	Z statistics (p value)
Fixed-effects Model	.16	[.12,.20]	7.72 (p<.01)
Random-effects Model -1	.14	[-.04,.31]	1.49 (p=.137)
Random-effects Model -2	.15	[-.001,.31]	1.94(p=.052)

Note: Random-effects Model -1 – Hedge and Colleagues;
Random-effects Model -2 - Hunter & Schmidt.

CHAPTER 3

ANALYSIS AND SIMULATION OF STATISTICAL POWER IN META-ANALYSIS

Meta-analysis Practice

Statistical power in meta-analysis can be computed, according to the formulas provided by Hedege and Pigott (2001). The following parameters influence the statistical power in a meta-analysis, and some of them are also relevant to a single primary study.

1. Sample size. Unlike primary studies, there are more varying conditions in meta-analysis. The sample size may vary from one study to another. The two groups in the same study may have unequal sample sizes. Balanced designs of individual studies also influence the estimation results.
2. Population effect size. As in primary studies, the population effect size is positively related to statistical power. Standardized effect size is commonly used in meta-analysis to unify the measurement scale across studies.
3. Number of small studies in a meta-analysis. Other things being equal, more studies are included in the meta-analysis, the higher the statistical power will be.
4. Analytical model. The fixed-effects model usually yields higher statistical power than the random-effects model. The former model does not consider the between-study variance.

The power analysis for a meta-analysis parallels the issues of power analysis for a primary study (Borenstein, Hedges, Higgins, & Rothstein, 2010). The procedures for computing statistical power in meta-analysis have been described by Hedges and Pigott

(2001), and they found statistical power is not always high in meta-analysis. Cafri, Kromrey, and Brannick (2010) suggested that researchers should pay close attention to power at the planning phase of a meta-analysis. Liu (2013) offered an explanation of how to compute power in meta-analysis, using statistical software (e.g., SAS and R).

Computation of Power in Meta-analysis

Statistical power was calculated based on the analytical procedures in Chapter 2. Different parameters were used in the computation of statistical power. The average sample size for each group, the number of individual studies, and the population effect size were varied to check the effects on power. It was of interest to learn how many individual studies, how many samples per study, and how large of a population effect size were needed to achieve the desired statistical power (e.g., 80%). The R codes were developed to compute power under different situations (see Appendix A). The average sample size n varied between 30 and 100, the number of studies I varied between 5 and 80, and the effect size ES ranged between 0.1 and 0.8. For the random-effects model, the between-group variance was varied to represent small, medium, and large amounts of heterogeneity across studies (Hedges & Pigott, 2001). The between-study variance Tau square was set to .33, .67 and 1.0 times the within-study variance. These values were used for results illustration. After discussing the simulation procedures in the next part, the parameters being selected in the simulation and power computation will be discussed in details.

A table and a figure (Table 3.1 and Figure 3.1) were generated to show the relationships between the model parameters and the sample sizes necessary to achieve the desired power.

Power Curves of different conditions were drawn in Figure 3.1. First, power increases with sample size and number of studies increases. A close examination reveal the differences in power between the fixed-effects and random-effects models through different line types. The previous research has frequently suggested that the fixed-effects model has higher statistical power than the random-effects model. This is confirmed in the graph. The fixed-effects model generally tends to have higher statistical power than the random-effects model. This is more pronounced for the random-effects model with elevated heterogeneity among individual studies. The population effect size was fixed as .1 for easy graph reading. The relationship between population effect size and power was illustrated in Table 3.1.

The desired power is set to 0.80 for discussion. If there is a high population effect size (0.8) in the study, the lower limit computation setting (n = 30, I = 5) is enough to reach .8 statistical power in all models. If there is a medium population effect size (around 0.5) in the study, a few more studies (n=10) are needed to reach .8 in all models. The power can be increased by increasing the average sample size of each study as well. For a small population effect size (0.1 or 0.2), a larger average sample size for each group and a large number of studies are needed to achieve the desired power. Around 100 subjects per study and 80 studies are required to obtain the desired power .8 when the population effect size is 0.1. Around 80 subjects per study and 20 studies are required to obtain the desired power .8 when the population effect size is 0.2. It is suggested that the statistical power is low when the population effect size is small for small sample sizes. In other words, researchers might make an incorrect decision even when there is a real

difference between female and male in hubristic pride (Chapter 2 example). The result highlights the concern over low statistical power in meta-analysis.

Consistent with results in Figure 3.1, differences between fixed and random effects models were identified in Table 3.1. However, the difference in statistical power between the fixed-effects and random-effects models diminishes as the population effect size became large. For the same population effect size, large sample sizes also reduce the difference in power between the fixed-effects and random-effects models. In other words, larger parameter values help equalize the fixed-effects and random-effects models in statistical power. Nevertheless, researchers should be aware of the power differences in the two models when they select a model for the planned meta-analysis.

These conclusions, such as the parameters that needs to reach power of .8, are tenable before the accuracy of statistical power was investigated.

Simulation of Statistical Power in Meta-analysis

Computer simulation can be used to further the understanding of statistical power in real meta-analysis. It can also be utilized to address the concerns over the approximate power. The current methodology simply assumes that each study has the same variance when estimating the population variance in the power formula (page 28 to page 29). The approximation was used to simply the power calculation process and it is seldom true in practice (Hedges & Pigott, 2001). Simulated power is more accurate compared with the analytical power. Comparing the estimated power and simulated power can help researchers check the accuracy of the computation findings at the beginning of this chapter. Also, the discrepancies between the estimated power and simulated power can help researchers identify the potential bias in the power formulas.

In the current study, I simulated the power under various conditions and then compared the simulated power with the approximate power based on the analytical formulas provided in the literature. The study was conducted using R. The following simulation conditions were defined based on similar studies (e.g., Field, 2003) and the pilot study:

(1) Average sample size: The average sample size varied in different meta-analysis studies. In the current study, the sample size ranged from 30 to 100 (i.e., 30, 40, 50, 60, 80, and 100). The average sample size in the real meta-analysis is usually large but this study is intended to check the influence of small sample size. Thus, the sample size larger than 100 was not considered. In addition, large sample size normally yields high/ideal statistical power even when other parameter values are low. In practice, the sample sizes among individual studies are unequal. Therefore, a truncated binomial distribution was used to generate integer positive numbers to meet the requirement of sample size. By varying the maximum value in the distribution, the variation of sample size was varied. The sample size of each study was varied, based on different ratios (e.g., group1:group2 = 1:2). The study started with the simple situation, in which the sample sizes across studies were the same, and the sample sizes between the two groups in each study were the same. Then, the study examined the varying sample sizes between studies and within studies.

(2) The following population effect sizes were used: no effect (0), a small effect (.1, .2, and .3), a moderate effect (.5), and a large effect (.8). These effect sizes were selected, based on Cohen's guidelines (1988). Although Cohen suggested .2 as a small effect, .1, .2, .3 were selected as the small population effect sizes. I chose to study more

on small effect sizes because they occur more often in practice. For instance, Hattie (2009) synthesized over 800 meta-analysis related to achievement. The overall distribution of all the effect sizes indicated that many of the effect sizes were small, i.e., under .4 (72 out of 138 studies). Therefore, the population effect size in the lower range will be studied more carefully.

(3) The number of studies: the number of studies ranged from 5 to 80 (i.e., 5, 10, 20, 50, 80). These numbers were chosen based on the real meta-analysis datasets. For instance, studies of children's self-conscious emotions (Else-Quest, Higgins, Allison, and Morton, 2012) had different number of studies in different emotion aspects ranging from 17 to 307. Different study numbers were used to cover most of the practical situations, and the number of studies higher than 80 was normally with satisfactory statistical power and was not included in this study.

(4) Number of Simulations: The meta-analysis was repeated 10,000 times to obtain a stable simulation result. This is 10 times as many as the minimum recommended (Mooney, 1997).

(5) Type I error rate was set to .05 in the current study.

(6) The fixed-effects model and two random-effects model were considered separately. Random-effects model -1 used the methods developed by Hedges and colleagues. Whereas, the random-effects model-2 used the method developed by Hunter and Schmidt.

The total simulated scenarios were based on four varying factors: 6 population effect sizes (0, .1, .2 .3, .5, .8), 6 average sample sizes (30, 40, 50, 60, 80, 100), 5 number of studies (5, 10, 20, 50, 80), 3 models (fixed-effects model, random-effects model 1, and

random-effects model 2). There were 36 combinations of the average sample size and number of studies. For each combination of the average sample size and number of studies, 10000 Monte Carlo trials were used. All these conditions were clearly defined as initial conditions accordingly. The average sample size and number of studies were defined into two vectors, which included all the selected conditions. Then the number of simulation time and alpha were fixed in the current study. The population effect size was defined as a single value in each simulation condition.

```
#sample size
possible.ns <- c(30,40,50,60,80,100)
#Number of studies
I.ns <- c(5,10,20,50,80)
# Set Type I error rate as .05(fixed)
alpha <- 0.05
# number of simulation iterations(fixed)
sims <- 10000
#Population effect size (set as 0,.1,.2,.3,.5,.8,)
PES <-0
```

The simulation code was developed based on the R code for meta-analysis shown in Chapter 2 (see Appendix A). To run the simulation efficiently, the parameter values were set in the loops (average sample size and number of studies). To limit the output matrix to two dimensions, the population effect size was varied in different simulation runs. The third loop was created to repeat meta-analysis (see the abbreviated R code below).

```
#loop for different average sample size
for (j in 1:n){
N<- possible.ns[j]
#loop for different number of studies
for (k in 1:s){
I<- I.ns[k]
#Simulation loop
for (i in 1:sims){
}
}
}
```

In each simulated run, a meta-analysis was conducted. The same formula was applied as referring to the meta-analysis application examples in Chapter 2. The only difference was the generation of the effect size in primary studies. The *t* distribution was used to generate the effect size in each meta-analysis. To address the bias in Cohen's d, Hedge's g was used in power simulation to provide more accurate effect size estimates. A Z statistic was calculated after each repetition of the simulated meta-analysis (see the R code below). The fixed-effects model with equal sample size between and within studies was used for explanation.

```
# In each simulation, perform the meta-analysis
# Sample size across studies equal in this condition
Nvary<-rep(N,I)
# Simulate the effect size using t distribution
# Sample size between two groups in each study are equal
d0 <- rt(I,Nvary-2)*2*sqrt(1/Nvary)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PES
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-(4/Nvary)*(1+0.125*ES*ES)
Varianceg<-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
SumWd<-sum(Weight*ES)
WeightedD<- SumWd/SumWeight
SEM<-sqrt(1/SumWeight)
Zstat<- WeightedD/SEM
```

In each meta-analysis, a *p*-value was saved. A statistical decision was then made according to the alpha level (.05). The frequency of rejecting the null hypothesis was saved in a 6x5 matrix for different average sample sizes and numbers of studies. They are the simulated statistical power across different conditions. When the population effect size was zero, the simulated statistical power was equal to the actual Type I error rate.

```
p.value[i]<- 2*pnorm(-abs(Zstat))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
prob[j,k] <- mean(significant.experiments)
```

The complete R code for simulating the statistical power was included in the Appendix A across different conditions. The analytical power of each condition was saved for purpose of comparison and results illustration. The similar loops were defined except that no simulation loop was defined for the analytical power. Using the formula, analytical power across different sample size and number of studies can be calculated and saved in a 6*5 matrix.

```
FixPowfunction<-function(possible.ns, I.ns, PES)
    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
   power <- array(rep(NA,n*s),dim=c(n,s))
   #looping at different sample size
   for (j in 1:n){
   N <- possible.ns[j]
   #looping at different number of studies
   for (k in 1:s){
   I<- I.ns[k]
   Vtotal<-(4/N)*(1+0.125*PES*PES)
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
}
}
   return(powerround)
    }
FixPowerFunction<-FixPowfunction(possible.ns,I.ns, PES)
FixPowerFunction
```

The fixed-effects model and two random-effects models were considered separately. Power tables and power curves under different conditions were created to illuminate the

results. Detail information of results organization was given at the beginning of Chapter 4.

The following expectations were made:

(1) Discrepancies between analytical power and simulated power exist because the analytical power is based on the approximation formulas. Discrepancies under or around .05 are assumed to be acceptable.

(2) Unbalanced design decreases the statistical power in meta-analysis as it does in the primary studies.

(3) There is no systematic bias in estimated power, as the discrepancies can show underestimation and overestimation.

Table 3.1 *Results of Power for the Fixed-effects Model and Random-effects Model*

n, I, ES	Fixed-effects model	Random-effects model (small heterogeneity)	Random-effects model (medium heterogeneity)	Random-effects model (large heterogeneity)
30, 5, 0.1	.094	.083	.076	.072
50, 5, 0.1	.124	.105	.094	.086
40, 10, 0.1	.170	.140	.121	.109
80, 20, 0.1	.516	.410	.340	.293
80, 50, 0.1	.885	.782	.686	.608
100, 80, 0.1	.994	.972	.933	.885
30, 5, 0.2	.231	.185	.157	.139
50, 5, 0.2	.351	.278	.231	.200
40, 10, 0.2	.514	.409	.339	.292
80, 20, 0.2	.979	.933	.870	.805
30, 5, 0.3	.447	.354	.293	.253
40, 10, 0.3	.847	.735	.636	.559
40, 20, 0.3	.988	.955	.904	.847
30, 5, 0.5	.854	.743	.646	.568
30, 10, 0.5	.989	.959	.910	.854
30, 5, 0.8	.997	.983	.954	.915

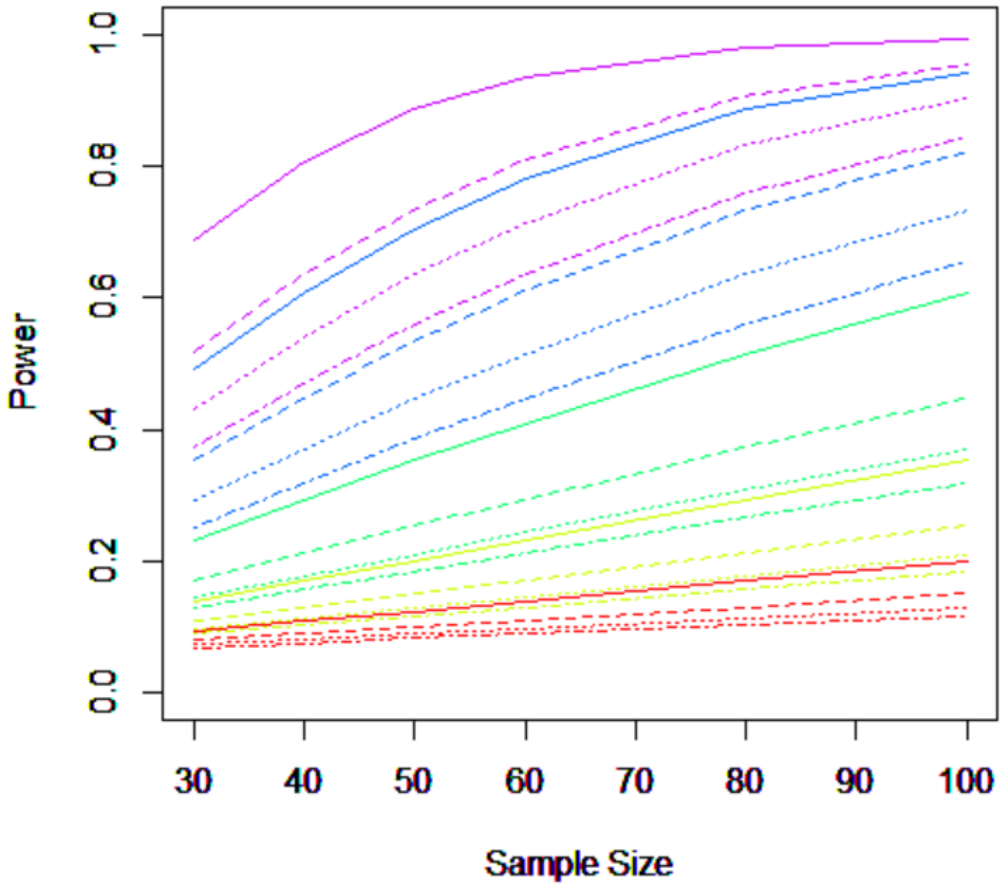

Figure 3.1 Power Curves under Different Parameter Values

Note: Solid lines Fixed-effects model; Dashed lines Random-effects model – low heterogeneity; Dotted lines Random-effects model – medium heterogeneity; Dotdash lines Random-effects model – high heterogeneity; purple, blue, green, yellow, and red lines: large to small number of studies.

CHAPTER 4

RESULTS

Two broad research questions were addressed in this chapter. First, the four different designs regarding the sample size were defined as follows: equal sample size and balanced design, unequal sample size and balanced design, equal sample size and unbalanced design, and unequal sample size and unbalanced design. The equal sample size referred to the same number of subjects of each individual study included in the meta-analysis, and the design balance indicated whether the sample sizes between the two groups in an individual study were equal or not. The unequal sample size across studies only influenced the simulated power, because analytical power only used average sample size across studies. To investigate the research questions, the simulated power and analytical power across selected conditions were generated for four designs. Power differences were considered for both research questions. The power curves of selected conditions were provided to show the results graphically.

Type I Error Control

The actual Type I error rate was checked through power simulation before investigating the research questions. Three models (i.e., fixed-effects model, random-effects model by Hedges and Colleagues, and random-effects model by Hunter and Schmidt). In the null case (population effect size =0), the probability of rejecting the null hypothesis represented the actual Type I error rates. The four designs were checked accordingly. This check was necessary because the Type I error can affect Type II error

and, in turn, the statistical power. The results of the equal sample size and balanced design were shown in Table 4.1. Using a nominal alpha of .05, it was clear that Type I errors were under control and limited to the purported five percent for the fixed-effects model and the random-effects model by Hedges and colleagues (Table 4.1). In other words, the two models produced error rates at around .05. However, the Type I error rate of the random-effects by Hunter and Schmidt was not controlled properly especially for small number of studies in a meta-analysis. The other three designs indicated similar conclusions and the exact Type I error values were not shown in the results table due to the page limit (The R codes were included in the Appendix A). This suggested that the model by Hunter and Schmidt should not be used in power simulation especially for small number of studies. The power values generating using this method should be interpreted cautiously. Since the influence of statistical power on small number of studies was an important concern of the current study, this model was removed from the following analysis.

It was known that the fixed-effects model yielded higher statistical power than the random-effects model (Table 3.1 and Figure 3.1). However, when the population effect size varied, the random-effects model should be used to meet to model assumption even lower power was received. Otherwise, the actual Type I error rate was inflated. This was especially so for a large sample size. Table 4.2 included the statistical power simulated in a fixed-effects model but with varied population effect sizes. The power was not accurately estimated if the fixed-effects model was used under such conditions, because the simulated power was based on the assigned Type I error rate (.05). Thus, model selection was important in statistical power for meta-analysis. In the following simulation

process, the population effect size across studies was fixed in the fixed-effects model. The population effect size of the random-effects model was assumed to follow a normal distribution with a mean of the average population effect size and a standard deviation of .1 to meet the random-effect model assumption. It was noted that different standard deviations were considered but the results were similar, so one setting was shown in the results. In addition, the simulated and analytical power used the same between-study variance to guarantee the comparability of simulated and analytical power.

Table 4.3 to Table 4.10 showed the simulated power values and analytical power values across different population effect size, average sample size, and number of studies for both models under four designs.

First, the results of balanced design and equal sample size across studies were shown in Table 4.3 (the fixed-effects model) and Table 4.4 (the random-effects model). Although this is rarely true in practice, the condition was included as a basis of the following analysis.

Next, unequal sample size across studies and balanced design were considered. It is well-known that equal sample size across studies is an ideal condition. Usually sample size across studies are not equal. The truncated binomial distribution was used to generate the varied sample size across studies (http://www.vosesoftware.com). This guaranteed that the generated sample sizes were positive integer numbers with the specified mean and standard deviation. The maximum sample size was varied, so was the standard deviation of the distribution in the binomial distribution. The maximum sample size was changed by multiplying the average sample size by certain numbers (e.g., average sample size * 3). A larger maximum sample size was related to a larger variation of all the

sample sizes. Different maximum sample sizes were tried in the pilot run, but similar results were obtained. Thus, only one condition was listed here (maximum sample size = average sample size * 3). The sample sizes for the two groups in an individual study were assumed to be equal in this condition (i.e., balanced design). The results of the fixed-effects model was shown in Table 4.5 and the results of the random-effects model was shown in Table 4.6.

In practice, individual studies included in a meta-analysis rarely have perfect design balance. As shown in Table 2.1 and 2.3, most studies did not have exactly the same sample size between the two groups. Thus, the average sample size ratio between the two groups of all studies was set to different values. As different sample size ratios produced similar discrepancy, only one sample size ratio was shown in the results (sample size ratio: 1:2). In practice, the sample size ratio of 1:2 should be enough unbalanced for practical meta-analysis dataset. Equal sample size across studies were assumed. Thus, the simulated power and analytical power of this design were displayed in Table 4.7 (the fixed-effects model) and Table 4.8 (the random-effects model).

Finally, unequal sample size across studies and unbalanced design within studies were examined. This was the most practical design. The results of the fixed and random effects model were shown in Table 4.9 and Table 4.10.

Discrepancy in Power Estimation

First, the discrepancies of different conditions were checked. Overall, the simulated and analytical power were close to each other (<=.05) under almost all the conditions and all designs from Table 4.3 to Table 4.10. There were no systematic discrepancies between simulated power and analytical power. In other words, the

analytical power was overestimated or underestimated in different conditions. Statistical power was generally understandably higher when the population effect size, average sample size, and number of studies were larger. In addition, when the population effect size was at .8, the simulated power and analytical power estimates were close to 1 without any discrepancy no matter what sample size, number of studies, or designs we had. In other words, the influence of other parameters became inconsequential under such conditions. However, this was not true for the average sample size or number of studies. The largest average sample size (100) itself cannot remove the discrepancy when the number of studies and population effect size were small. This is the same for the largest number of studies.

Next, the different patterns in the fixed and random-effects models were discussed. The discrepancies between the simulated power and analytical power for the fixed-effects model were generally minimal under four designs. All the discrepancies were around or less than .01. Power curves of the fixed-effects model for the equal sample size and balanced design were shown in Figure 4.1. It was shown that the analytical power (solid lines) were close to simulated power (dashed lines) and it was hard to tell the difference between two groups of lines from the graph. Only one population effect size (i.e., 0.1) was used on the graph to show the largest discrepancies. Other population effect size has smaller discrepancies and the power curves under larger population effect were too close to each other to read from a graph. Only one graph for the fixed-effects model was drawn due to the similar conclusions across four designs. The discrepancies were larger in the random-effects models compared with the fixe-effects model under certain conditions when other parameters were held constant. There were

noticeable power discrepancies in the random-effects model under a few conditions. This was especially so for the unequal sample size and unbalanced design. For instance, when the population effect size was .3, the number of studies was 5, the average sample size was 60 in the design of unequal sample size and unbalanced design, the discrepancy between simulated power and analytical power was .059 (Table 4.10). When the discrepancies were large, the analytical power estimates were more likely to underestimate the real power. It was also more likely to occur when at least one of the parameters was not large enough. Interestingly, when the population effect size was not large enough, the discrepancies increased with the higher population effect size by fixing the other two parameters under certain conditions. For instance, in Table 4.10, the power discrepancy was .006 when the population effect size was .1 with the average sample size of 30 and number of studies of 5. The power discrepancies were .021, .043, and .051 when the population effect size were .2, .3, and .5. The discrepancy disappeared when the population effect size was .8.

Finally, the discrepancies of random-effects model were different for different designs. The discrepancies were larger for the unequal sample size and unbalanced design. The average discrepancy of all selected conditions for the first three designs were around .005, and the average discrepancy of all selections for the fourth design (i.e., unequal sample size and unbalanced design) was around .02. Figure 4.2 to Figure 4.5 showed the power curves of the four designs. One population effect size (i.e., .1) was used to show the pattern. Figure 4.2 to Figure 4.4 showed similar discrepancy patterns across different conditions. However, Figure 4.5 showed that unequal sample size and unbalanced design had larger power discrepancies compared with the other three designs.

It was noticed that all the discussions above were based on the power estimates with variations. When the parameters were large enough, power estimates were close to 1. No discussion of power discrepancies were needed.

Overall, the approximate analytical power was close to the real simulated power with acceptable discrepancies when the average sample size, population effect size, and number of studies were varied. Some of the conditions in the random effects models had noticeable power discrepancies as shown in Table 4.4, Table 4.6, Table 4.8, and Table 4.10.

Influence of Unequal Sample Size and Unbalanced Design on Statistical Power

Next, influences of unequal sample sizes across studies and unbalanced design on statistical power were examined. Although the simulated power and analytical power were close to each other, the simulated power across different conditions was used for the analysis because it was construed as the actual power. Each condition was examined separately and then combined together for the final investigation. The unequal sample size and

Population effect size can improve statistical power as seen from the power tables. Large population effect size (0.8) was not a big concern since it yielded perfect power estimates under various conditions. Different population effect sizes (0.1, 0.2, 0.3, and 0.5) were discussed. Power difference under compared conditions were used to investigate the influence. Population effect size of .1 with different average sample size and number of studies were used to generate power curves.

First, the influence of the unequal sample size were checked. The difference of the fixed-effects model can be checked by comparing the power estimates from Table 4.3

and Table 4.5. The difference of the random-effects model can be checked by comparing Table 4.4 and Table 4.6. The power differences were calculated and shown in Table 4.11. The unequal sample size did not have a systematic influence on the statistical power. Power values from two conditions were close to each other. When the population effect size was .1, the power curves of the fixed and random-effects models were drawn separately (Figure 4.6 and Figure 4.7). It was also hard to see the trend when all the power values were close to one side in the graphs, so power curves under other population effect sizes were not shown. Similarly as the results from the Table 4.11, the curves indicated that different sample size across studies did not affect the statistical power (solid lines and dashed lines).

Then, the influence of unbalanced design on statistical power was investigated. It was known that in primary studies, the unbalanced design decreased the statistical power. Compared with equal sample size between groups, the unbalanced design was associated with lower statistical power in meta-analysis as well. The difference of the fixed-effects model can be checked by comparing the power estimates from Table 4.3 and Table 4.7. The difference of the random-effects model can be checked by comparing Table 4.4 and Table 4.8. The power differences were listed in Table 4.12 for both models. The power of unbalanced design was always lower than the power of balanced design. The largest difference was .057 (population effect size: .2, average sample size: 30; number of studies: 20). The power decreased around .04 to .05 in many cells. Interestingly, when all the parameters were small, the power did not decrease a lot. Instead, the large power drop occurred when the one of the parameters increased but not large enough to avoid the discrepancy. There was no power difference if the parameters were large enough to

generate perfect power (close to 1). The power curves were drawn for population effect size .1 to show similar conclusions. Average sample size ratio of 1:2 and 1:4 were both included to show that power decreased with more unbalanced design. Figure 4.8 and Figure 4.9 indicated that higher degree of unbalanced sample size between both groups, design imbalance could substantially lower statistical power. The degree of decreasing was different under different conditions.

Finally, the influence of both factors on statistical power were considered. The power differences were calculated using power estimates from Table 4.3, Table 4.4, Table 4.9, and Table 4.10. The results were shown in Table 4.13. The results indicated the power decreased under most of the conditions. Surprisingly, power estimates increased in the random-effects model when the number of studies was 5 and population effect size was .1. Again, power curves were drawn to show the results more directly. Figure 4.10 (fixed-effects model) and Figure 4.11 (random-effects model) indicated that the statistical power was decreased as studies became more unbalanced and more varied in sample size.

The power curves of four designs were drawn in one paragraph to check the power differences at the end. Figure 4.12 (fixed-effects model) and Figure 4.13 (random-effects model) further indicated that the solid lines (square and plus symbols for equal sample size and balance design and unequal sample size and balanced design) were close to each other, and that the dotted lines (circle and cross symbols for equal sample size and unbalanced design and unequal sample size and unbalanced design) were close to each other. Thus, the decrease of statistical power was largely due to the unbalanced design rather than the unequal sample size across studies.

As stated in the power discrepancy discussion, power estimates were close to 1 under certain conditions (e.g., large population effect size). The discussion of unequal sample size across studies and unbalanced design was not necessary for those conditions.

The study presents the results of a thorough simulation of conditions that may influence power of different meta-analysis methods. The analytical power were generated to match the conditions in the simulation. The study provided a broader insight into the power estimates of meta-analysis procedures. Three predictions were generally supported:

(1) Discrepancies between analytical power and simulated power were identified. Of selected conditions, all the discrepancies in the fixe-effects model were below .05. A few discrepancy values in the random-effects model were above .05.

(2) Unbalanced design decreases the statistical power, while unequal sample size across studies does not.

(3) There is no systematic bias in analytical power. As shown from the power tables. Underestimation and overestimation were both identified. However, larger discrepancy in power estimates (around .05) indicated the underestimation of power.

Table 4.1 *Type I Error Rates of Three Models – Equal Sample size and Balanced Design*

Fixed-Effects Model					
Average Sample Size	Number of Studies				
	5	10	20	50	80
30	.046	.048	.047	.050	.049
40	.050	.049	.053	.050	.045
50	.047	.046	.049	.050	.047
60	.049	.046	.052	.047	.048
80	.045	.045	.048	.047	.052
100	.049	.049	.053	.050	.052
Random-Effects Model – 1					
Average Sample Size	Number of Studies				
	5	10	20	50	80
30	.041	.038	.043	.044	.047
40	.041	.042	.046	.045	.043
50	.045	.046	.047	.049	.052
60	.049	.046	.049	.054	.045
80	.050	.046	.048	.050	.049
100	.051	.051	.053	.054	.053
Random-Effects Model – 2					
Average Sample Size	Number of Studies				
	5	10	20	50	80
30	.151	.090	.070	.057	.058
40	.154	.091	.072	.058	.052
50	.154	.095	.071	.059	.060
60	.158	.095	.071	.065	.051
80	.154	.090	.067	.058	.055
100	.155	.096	.071	.061	.057

Table 4.2 *Type I Error Rates of Fixed-Effects model with Varied Population Effect Sizes*

Population Effect Size (SD=.1)					
Average Sample Size	Number of Studies				
	5	10	20	50	80
30	.058	.052	.055	.056	.057
40	.056	.058	.058	.061	.058
50	.063	.062	.063	.065	.066
60	.069	.066	.066	.071	.062
80	.074	.068	.072	.069	.072
100	.078	.077	.079	.080	.079
Population Effect Size (SD=.2)					
Average Sample Size	Number of Studies				
	5	10	20	50	80
30	.080	.075	.079	.079	.082
40	.087	.091	.092	.094	.088
50	.102	.103	.105	.103	.106
60	.122	.114	.118	.121	.109
80	.139	.136	.136	.136	.139
100	.161	.157	.162	.162	.161

Table 4.3 *Statistical Power of the Fixed-effects Model (Equal Sample Size and Balanced Design)*

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
	Population Effect Size = .1									
30	0.093	0.138	0.228	0.485	0.676	0.094	0.139	0.232	0.490	0.687
40	0.108	0.169	0.281	0.609	0.799	0.109	0.170	0.293	0.608	0.807
50	0.117	0.198	0.353	0.696	0.886	0.124	0.201	0.352	0.705	0.885
60	0.136	0.229	0.405	0.775	0.934	0.139	0.232	0.410	0.781	0.933
80	0.162	0.286	0.519	0.884	0.977	0.170	0.293	0.516	0.885	0.979
100	0.196	0.347	0.601	0.941	0.994	0.201	0.352	0.608	0.942	0.994
	Population Effect Size = .2									
30	0.229	0.410	0.689	0.972	0.998	0.231	0.408	0.686	0.972	0.998
40	0.286	0.516	0.797	0.994	1.000	0.292	0.514	0.806	0.994	1.000
50	0.345	0.601	0.881	0.999	1.000	0.351	0.607	0.884	0.999	1.000
60	0.411	0.689	0.933	1.000	1.000	0.408	0.686	0.933	1.000	1.000
80	0.510	0.806	0.980	1.000	1.000	0.514	0.806	0.979	1.000	1.000
100	0.604	0.882	0.993	1.000	1.000	0.607	0.884	0.994	1.000	1.000
	Population Effect Size = .3									
30	0.447	0.737	0.952	1.000	1.000	0.447	0.734	0.955	1.000	1.000
40	0.556	0.854	0.987	1.000	1.000	0.560	0.847	0.988	1.000	1.000
50	0.651	0.915	0.997	1.000	1.000	0.655	0.916	0.997	1.000	1.000
60	0.741	0.956	0.999	1.000	1.000	0.734	0.955	0.999	1.000	1.000
80	0.845	0.989	1.000	1.000	1.000	0.847	0.988	1.000	1.000	1.000
100	0.915	0.997	1.000	1.000	1.000	0.916	0.997	1.000	1.000	1.000
	Population Effect Size = .5									
30	0.862	0.990	1.000	1.000	1.000	0.854	0.989	1.000	1.000	1.000
40	0.940	0.999	1.000	1.000	1.000	0.936	0.999	1.000	1.000	1.000

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
50	0.977	1.000	1.000	1.000	1.000	0.973	1.000	1.000	1.000	1.000
60	0.989	1.000	1.000	1.000	1.000	0.989	1.000	1.000	1.000	1.000
80	0.998	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Population Effect Size = .8										
30	0.998	1.000	1.000	1.000	1.000	0.997	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.4 *Statistical Power of the Random-effects Model (Balanced Design and Equal Sample Size across Studies)*

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = .1										
30	.076	.116	.200	.439	.644	.083	.121	.203	.443	.638
40	.085	.142	.236	.544	.755	.093	.146	.252	.550	.755
50	.100	.167	.305	.632	.842	.104	.168	.300	.638	.838
60	.118	.195	.351	.712	.887	.114	.190	.346	.711	.893
80	.152	.237	.430	.812	.950	.135	.232	.427	.816	.952
100	.166	.274	.504	.873	.978	.153	.274	.500	.882	.979
Population Effect Size = .2										
30	.188	.343	.609	.954	.997	.186	.341	.612	.952	.996
40	.234	.439	.726	.988	1.000	.228	.431	.732	.986	1.000
50	.280	.518	.823	.997	1.000	.272	.509	.817	.996	1.000
60	.333	.590	.877	.999	1.000	.312	.577	.877	.999	1.000
80	.422	.701	.943	1.000	1.000	.392	.688	.944	1.000	1.000
100	.487	.777	.975	1.000	1.000	.457	.773	.975	1.000	1.000
Population Effect Size = .3										
30	.377	.657	.922	1.000	1.000	.356	.642	.920	1.000	1.000
40	.462	.774	.973	1.000	1.000	.444	.763	.972	1.000	1.000
50	.549	.852	.989	1.000	1.000	.527	.844	.990	1.000	1.000
60	.620	.904	.996	1.000	1.000	.596	.898	.997	1.000	1.000
80	.738	.961	.999	1.000	1.000	.714	.957	1.000	1.000	1.000
100	.804	.981	1.000	1.000	1.000	.793	.982	1.000	1.000	1.000
Population Effect Size = .5										
30	.780	.976	1.000	1.000	1.000	.751	.971	1.000	1.000	1.000
40	.870	.993	1.000	1.000	1.000	.855	.993	1.000	1.000	1.000

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
50	.925	.998	1.000	1.000	1.000	.920	.999	1.000	1.000	1.000
60	.954	.999	1.000	1.000	1.000	.955	1.000	1.000	1.000	1.000
80	.984	1.000	1.000	1.000	1.000	.987	1.000	1.000	1.000	1.000
100	.994	1.000	1.000	1.000	1.000	.996	1.000	1.000	1.000	1.000
Population Effect Size = .8										
30	.986	1.000	1.000	1.000	1.000	.986	1.000	1.000	1.000	1.000
40	.996	1.000	1.000	1.000	1.000	.998	1.000	1.000	1.000	1.000
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	.986	1.000	1.000	1.000	1.000

Table 4.5 *Statistical Power of the Fixed-effects Model (Maximum sample size: Average sample size * 3)*

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
	Population Effect Size = .1									
30	.096	.133	.227	.478	.676	.094	.139	.232	.490	.687
40	.104	.168	.294	.606	.805	.109	.170	.293	.608	.807
50	.120	.196	.348	.699	.885	.124	.201	.352	.705	.885
60	.142	.223	.418	.775	.929	.139	.232	.410	.781	.933
80	.165	.292	.510	.884	.978	.170	.293	.516	.885	.979
100	.197	.344	.603	.941	.993	.201	.352	.608	.942	.994
	Population Effect Size = .2									
30	.225	.403	.682	.970	.999	.231	.408	.686	.972	.998
40	.289	.499	.804	.993	1.000	.292	.514	.806	.994	1.000
50	.351	.618	.885	.998	1.000	.351	.607	.884	.999	1.000
60	.416	.677	.933	1.000	1.000	.408	.686	.933	1.000	1.000
80	.505	.809	.978	1.000	1.000	.514	.806	.979	1.000	1.000
100	.601	.879	.993	1.000	1.000	.607	.884	.994	1.000	1.000
	Population Effect Size = .3									
30	.445	.737	.952	1.000	1.000	.447	.734	.955	1.000	1.000
40	.557	.840	.989	1.000	1.000	.560	.847	.988	1.000	1.000
50	.646	.916	.997	1.000	1.000	.655	.916	.997	1.000	1.000
60	.738	.956	.999	1.000	1.000	.734	.955	.999	1.000	1.000
80	.845	.988	1.000	1.000	1.000	.847	.988	1.000	1.000	1.000
100	.916	.997	1.000	1.000	1.000	.916	.997	1.000	1.000	1.000
	Population Effect Size = .5									
30	.863	.991	1.000	1.000	1.000	.854	.989	1.000	1.000	1.000
40	.931	.999	1.000	1.000	1.000	.936	.999	1.000	1.000	1.000

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
50	.975	1.000	1.000	1.000	1.000	.973	1.000	1.000	1.000	1.000
60	.990	1.000	1.000	1.000	1.000	.989	1.000	1.000	1.000	1.000
80	.998	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Population Effect Size = .8										
30	.997	1.000	1.000	1.000	1.000	.997	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.6 *Statistical Power of the Random-effects Model (Maximum sample size: Average sample size * 3)*

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
	Population Effect Size = .1									
30	.077	.112	.189	.432	.628	.083	.122	.203	.443	.638
40	.085	.143	.244	.547	.755	.093	.145	.252	.549	.757
50	.098	.169	.295	.643	.834	.104	.167	.300	.637	.837
60	.119	.188	.343	.710	.894	.114	.190	.346	.710	.893
80	.138	.240	.428	.813	.951	.133	.233	.427	.816	.953
100	.169	.285	.502	.880	.977	.152	.274	.499	.882	.979
	Population Effect Size = .2									
30	.189	.348	.611	.956	.995	.186	.343	.612	.953	.996
40	.237	.437	.728	.984	1.000	.228	.429	.733	.986	1.000
50	.286	.521	.821	.996	1.000	.271	.507	.818	.996	1.000
60	.334	.582	.877	.999	1.000	.311	.578	.877	.999	1.000
80	.414	.692	.942	1.000	1.000	.388	.690	.944	1.000	1.000
100	.482	.776	.974	1.000	1.000	.455	.774	.975	1.000	1.000
	Population Effect Size = .3									
30	.372	.655	.921	1.000	1.000	.356	.643	.920	1.000	1.000
40	.460	.773	.972	1.000	1.000	.443	.760	.972	1.000	1.000
50	.535	.853	.991	1.000	1.000	.525	.842	.990	1.000	1.000
60	.618	.900	.997	1.000	1.000	.594	.898	.997	1.000	1.000
80	.725	.957	.999	1.000	1.000	.708	.957	1.000	1.000	1.000
100	.802	.981	1.000	1.000	1.000	.790	.982	1.000	1.000	1.000
	Population Effect Size = .5									
30	.773	.974	1.000	1.000	1.000	.751	.971	1.000	1.000	1.000
40	.864	.993	1.000	1.000	1.000	.854	.993	1.000	1.000	1.000

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
50	.922	.998	1.000	1.000	1.000	.919	.998	1.000	1.000	1.000
60	.951	.999	1.000	1.000	1.000	.954	1.000	1.000	1.000	1.000
80	.982	1.000	1.000	1.000	1.000	.986	1.000	1.000	1.000	1.000
100	.993	1.000	1.000	1.000	1.000	.996	1.000	1.000	1.000	1.000
Population Effect Size = .8										
30	.987	1.000	1.000	1.000	1.000	.986	1.000	1.000	1.000	1.000
40	.995	1.000	1.000	1.000	1.000	.998	1.000	1.000	1.000	1.000
50	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.7 *Statistical Power of the Fixed-effects Model (Average sample size ratio: 1:2)*

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
	Population Effect Size = .1									
30	.089	.126	.207	.439	.624	.089	.129	.211	.446	.636
40	.101	.157	.255	.556	.751	.102	.156	.266	.559	.760
50	.109	.180	.319	.646	.850	.116	.184	.319	.654	.846
60	.125	.210	.367	.726	.905	.129	.211	.372	.733	.904
80	.148	.258	.473	.844	.963	.156	.266	.470	.846	.965
100	.179	.316	.551	.911	.987	.184	.319	.559	.915	.988
	Population Effect Size = .2									
30	.208	.370	.632	.953	.995	.211	.371	.635	.954	.996
40	.259	.470	.751	.987	1.000	.265	.469	.758	.988	1.000
50	.312	.550	.837	.997	1.000	.319	.557	.845	.997	1.000
60	.371	.638	.908	1.000	1.000	.371	.635	.903	.999	1.000
80	.464	.761	.965	1.000	1.000	.469	.758	.964	1.000	1.000
100	.555	.844	.986	1.000	1.000	.557	.845	.988	1.000	1.000
	Population Effect Size = .3									
30	.406	.687	.925	1.000	1.000	.407	.684	.932	1.000	1.000
40	.510	.808	.975	1.000	1.000	.512	.804	.978	1.000	1.000
50	.599	.883	.993	1.000	1.000	.605	.882	.994	1.000	1.000
60	.688	.930	.998	1.000	1.000	.684	.932	.998	1.000	1.000
80	.800	.981	1.000	1.000	1.000	.804	.978	1.000	1.000	1.000
100	.882	.993	1.000	1.000	1.000	.882	.994	1.000	1.000	1.000
	Population Effect Size = .5									
30	.818	.982	1.000	1.000	1.000	.813	.981	1.000	1.000	1.000
40	.913	.997	1.000	1.000	1.000	.908	.996	1.000	1.000	1.000

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
50	.961	1.000	1.000	1.000	1.000	.957	.999	1.000	1.000	1.000
60	.983	1.000	1.000	1.000	1.000	.981	1.000	1.000	1.000	1.000
80	.996	1.000	1.000	1.000	1.000	.996	1.000	1.000	1.000	1.000
100	.999	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000	1.000
Population Effect Size = .8										
30	.995	1.000	1.000	1.000	1.000	.994	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000	1.000
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.8 *Statistical Power of the Random-effects Model (Average sample size ratio: 1:2)*

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = .1										
30	.072	.106	.182	.398	.594	.079	.114	.186	.404	.590
40	.079	.130	.215	.499	.712	.089	.135	.231	.505	.708
50	.091	.155	.279	.586	.799	.098	.156	.274	.591	.797
60	.109	.178	.317	.668	.855	.107	.176	.316	.664	.859
80	.141	.215	.394	.769	.930	.126	.214	.392	.775	.932
100	.155	.249	.467	.839	.965	.143	.252	.462	.849	.967
Population Effect Size = .2										
30	.173	.309	.565	.931	.992	.171	.311	.565	.929	.992
40	.212	.399	.673	.978	.999	.209	.394	.685	.976	.999
50	.254	.473	.783	.994	1.000	.249	.467	.774	.992	1.000
60	.300	.545	.842	.998	1.000	.286	.533	.841	.998	1.000
80	.387	.657	.921	1.000	1.000	.359	.643	.921	1.000	1.000
100	.445	.738	.962	1.000	1.000	.421	.731	.961	1.000	1.000
Population Effect Size = .3										
30	.341	.603	.891	.999	1.000	.324	.593	.889	.999	1.000
40	.420	.729	.958	1.000	1.000	.405	.716	.955	1.000	1.000
50	.505	.817	.983	1.000	1.000	.484	.803	.982	1.000	1.000
60	.576	.874	.994	1.000	1.000	.550	.865	.993	1.000	1.000
80	.695	.943	.999	1.000	1.000	.669	.936	.999	1.000	1.000
100	.768	.971	1.000	1.000	1.000	.751	.971	1.000	1.000	1.000
Population Effect Size = .5										
30	.736	.962	1.000	1.000	1.000	.703	.954	1.000	1.000	1.000
40	.834	.988	1.000	1.000	1.000	.814	.988	1.000	1.000	1.000

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
50	.898	.995	1.000	1.000	1.000	.890	.997	1.000	1.000	1.000
60	.936	.999	1.000	1.000	1.000	.934	.999	1.000	1.000	1.000
80	.975	1.000	1.000	1.000	1.000	.978	1.000	1.000	1.000	1.000
100	.991	1.000	1.000	1.000	1.000	.992	1.000	1.000	1.000	1.000
Population Effect Size = .8										
30	.978	1.000	1.000	1.000	1.000	.977	1.000	1.000	1.000	1.000
40	.993	1.000	1.000	1.000	1.000	.995	1.000	1.000	1.000	1.000
50	.998	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000	1.000
60	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.9 *Statistical Power of the Fixed-effects Model (Average sample size ratio – 1:2; Maximum sample size: Average sample size * 3)*

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = .1										
30	.090	.124	.208	.436	.625	.089	.129	.211	.446	.636
40	.097	.154	.266	.555	.757	.102	.156	.266	.559	.760
50	.112	.178	.314	.641	.847	.116	.184	.319	.654	.846
60	.132	.203	.381	.726	.898	.129	.211	.372	.733	.904
80	.151	.266	.467	.846	.965	.156	.266	.470	.846	.965
100	.182	.309	.550	.912	.986	.184	.319	.559	.915	.988
Population Effect Size = .2										
30	.208	.365	.632	.951	.996	.211	.371	.635	.954	.996
40	.263	.456	.756	.986	1.000	.265	.469	.758	.988	1.000
50	.318	.566	.845	.997	1.000	.319	.557	.845	.997	1.000
60	.377	.629	.901	.999	1.000	.371	.635	.903	.999	1.000
80	.462	.758	.963	1.000	1.000	.469	.758	.964	1.000	1.000
100	.550	.839	.987	1.000	1.000	.557	.845	.988	1.000	1.000
Population Effect Size = .3										
30	.401	.687	.928	1.000	1.000	.407	.684	.932	1.000	1.000
40	.513	.794	.979	1.000	1.000	.512	.804	.978	1.000	1.000
50	.597	.884	.993	1.000	1.000	.605	.882	.994	1.000	1.000
60	.689	.930	.998	1.000	1.000	.684	.932	.998	1.000	1.000
80	.798	.980	1.000	1.000	1.000	.804	.978	1.000	1.000	1.000
100	.881	.993	1.000	1.000	1.000	.882	.994	1.000	1.000	1.000
Population Effect Size = .5										
30	.819	.984	1.000	1.000	1.000	.813	.981	1.000	1.000	1.000

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
40	.903	.997	1.000	1.000	1.000	.908	.996	1.000	1.000	1.000
50	.958	1.000	1.000	1.000	1.000	.957	.999	1.000	1.000	1.000
60	.982	1.000	1.000	1.000	1.000	.981	1.000	1.000	1.000	1.000
80	.997	1.000	1.000	1.000	1.000	.996	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000	1.000
Population Effect Size = .8										
30	.994	1.000	1.000	1.000	1.000	.994	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000	1.000
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.10 *Statistical Power of the Random-effects Model (Average sample size ratio – 1:2; Maximum sample size: Average sample size * 3)*

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = .1										
30	.084	.115	.184	.405	.591	.078	.111	.179	.384	.561
40	.090	.141	.235	.513	.716	.087	.131	.221	.479	.679
50	.105	.167	.278	.603	.795	.096	.150	.262	.562	.766
60	.124	.184	.326	.671	.864	.105	.170	.301	.634	.832
80	.142	.231	.404	.776	.932	.122	.206	.373	.746	.913
100	.170	.275	.476	.852	.965	.138	.242	.439	.823	.955
Population Effect Size = .2										
30	.187	.331	.578	.936	.990	.166	.301	.543	.914	.988
40	.234	.416	.692	.976	.999	.203	.377	.662	.968	.998
50	.280	.496	.787	.992	1.000	.241	.448	.752	.988	1.000
60	.324	.554	.847	.998	1.000	.276	.514	.820	.996	1.000
80	.398	.662	.924	.999	1.000	.344	.623	.905	1.000	1.000
100	.460	.747	.962	1.000	1.000	.405	.710	.951	1.000	1.000
Population Effect Size = .3										
30	.357	.623	.897	1.000	1.000	.314	.577	.873	.999	1.000
40	.442	.742	.958	1.000	1.000	.392	.694	.945	1.000	1.000
50	.510	.821	.986	1.000	1.000	.467	.782	.977	1.000	1.000
60	.592	.871	.994	1.000	1.000	.533	.848	.991	1.000	1.000
80	.696	.941	.998	1.000	1.000	.645	.926	.998	1.000	1.000
100	.776	.972	1.000	1.000	1.000	.732	.965	1.000	1.000	1.000
Population Effect Size = .5										
30	.739	.963	1.000	1.000	1.000	.688	.947	.999	1.000	1.000

Average Sample Size	Number of Studies									
	Power Simulation					Power Function				
	5	10	20	50	80	5	10	20	50	80
40	.837	.988	1.000	1.000	1.000	.800	.984	1.000	1.000	1.000
50	.899	.996	1.000	1.000	1.000	.877	.995	1.000	1.000	1.000
60	.935	.999	1.000	1.000	1.000	.924	.999	1.000	1.000	1.000
80	.975	1.000	1.000	1.000	1.000	.972	1.000	1.000	1.000	1.000
100	.989	1.000	1.000	1.000	1.000	.990	1.000	1.000	1.000	1.000
Population Effect Size = .8										
30	.973	1.000	1.000	1.000	1.000	.973	1.000	1.000	1.000	1.000
40	.994	1.000	1.000	1.000	1.000	.994	1.000	1.000	1.000	1.000
50	.999	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000	1.000
60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.11 *Power Difference between Equal Sample size and Unequal Sample Size*

Average Sample Size	Number of Studies									
	Fixed-effects Model					Random-effects Model				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = .1										
30	-.003	.005	.001	.007	0	.001	-.004	-.011	-.007	-.016
40	.004	.001	-.013	.003	-.006	0	.001	.008	.003	0
50	-.003	.002	.005	-.003	.001	-.002	.002	-.01	.011	-.008
60	-.006	.006	-.013	0	.005	.001	-.007	-.008	-.002	.007
80	-.003	-.006	.009	0	-.001	-.014	.003	-.002	.001	.001
100	-.001	.003	-.002	0	.001	.003	.011	-.002	.007	-.001
Population Effect Size = .2										
30	.004	.007	.007	.002	-.001	.001	.005	.002	.002	-.002
40	-.003	.017	-.007	.001	0	.003	-.002	.002	-.004	0
50	-.006	-.017	-.004	.001	0	.006	.003	-.002	-.001	0
60	-.005	.012	0	0	0	.001	-.008	0	0	0
80	.005	-.003	.002	0	0	-.008	-.009	-.001	0	0
100	.003	.003	0	0	0	-.005	-.001	-.001	0	0
Population Effect Size = .3										
30	.002	0	0	0	0	-.005	-.002	-.001	0	0
40	-.001	.014	-.002	0	0	-.002	-.001	-.001	0	0
50	.005	-.001	0	0	0	-.014	.001	.002	0	0
60	.003	0	0	0	0	-.002	-.004	.001	0	0
80	0	.001	0	0	0	-.013	-.004	0	0	0
100	-.001	0	0	0	0	-.002	0	0	0	0
Population Effect Size = .5										
30	-.001	-.001	0	0	0	-.007	-.002	0	0	0
40	.009	0	0	0	0	-.006	0	0	0	0

Average Sample Size	Number of Studies									
	Fixed-effects Model					Random-effects Model				
	5	10	20	50	80	5	10	20	50	80
50	.002	0	0	0	0	-.003	0	0	0	0
60	-.001	0	0	0	0	-.003	0	0	0	0
80	0	0	0	0	0	-.002	0	0	0	0
100	0	0	0	0	0	-.001	0	0	0	0

Table 4.12 *Power Difference between Balanced Design and Unbalanced Design*

Average Sample Size	Number of Studies									
	Fixed-effects Model					Random effects Model				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = .1										
30	.004	.012	.021	.046	.052	.004	.01	.018	.041	.05
40	.007	.012	.026	.053	.048	.006	.012	.021	.045	.043
50	.008	.018	.034	.050	.036	.009	.012	.026	.046	.043
60	.011	.019	.038	.049	.029	.009	.017	.034	.044	.032
80	.014	.028	.046	.04	.014	.011	.022	.036	.043	.02
100	.017	.031	.050	.03	.007	.011	.025	.037	.034	.013
Population Effect Size = .2										
30	.021	.040	.057	.019	.003	.015	.034	.044	.023	.005
40	.027	.046	.046	.007	0	.022	.04	.053	.01	.001
50	.033	.051	.044	.002	0	.026	.045	.04	.003	0
60	.04	.051	.025	0	0	.033	.045	.035	.001	0
80	.046	.045	.015	0	0	.035	.044	.022	0	0
100	.049	.038	.007	0	0	.042	.039	.013	0	0
Population Effect Size = .3										
30	.041	.05	.027	0	0	.036	.054	.031	.001	0
40	.046	.046	.012	0	0	.042	.045	.015	0	0
50	.052	.032	.004	0	0	.044	.035	.006	0	0
60	.053	.026	.001	0	0	.044	.03	.002	0	0
80	.045	.008	0	0	0	.043	.018	0	0	0
100	.033	.004	0	0	0	.036	.01	0	0	0
Population Effect Size = .5										
30	.044	.008	0	0	0	.044	.014	0	0	0
40	.027	.002	0	0	0	.036	.005	0	0	0

Average Sample Size	Number of Studies									
	Fixed-effects Model					Random effects Model				
	5	10	20	50	80	5	10	20	50	80
50	.016	0	0	0	0	.027	.003	0	0	0
60	.006	0	0	0	0	.018	0	0	0	0
80	.002	0	0	0	0	.009	0	0	0	0
100	.001	0	0	0	0	.003	0	0	0	0

Table 4.13 *Power Difference between Equal Sample Size, Balanced Design and Unequal Sample Size, Unbalanced Design*

Average Sample Size	Number of Studies									
	Fixed-effects Model					Random effects Model				
	5	10	20	50	80	5	10	20	50	80
Population Effect Size = .1										
30	.003	.014	.02	.049	.051	-.008	.001	.016	.034	.053
40	.011	.015	.015	.054	.042	-.005	.001	.001	.031	.039
50	.005	.02	.039	.055	.039	-.005	0	.027	.029	.047
60	.004	.026	.024	.049	.036	-.006	.011	.025	.041	.023
80	.011	.02	.052	.038	.012	.01	.006	.026	.036	.018
100	.014	.038	.051	.029	.008	-.004	-.001	.028	.021	.013
Population Effect Size = .2										
30	.021	.045	.057	.021	.002	.001	.012	.031	.018	.007
40	.023	.06	.041	.008	0	0	.023	.034	.012	.001
50	.027	.035	.036	.002	0	0	.022	.036	.005	0
60	.034	.06	.032	.001	0	.009	.036	.03	.001	0
80	.048	.048	.017	0	0	.024	.039	.019	.001	0
100	.054	.043	.006	0	0	.027	.03	.013	0	0
Population Effect Size = .3										
30	.046	.05	.024	0	0	.02	.034	.025	0	0
40	.043	.06	.008	0	0	.02	.032	.015	0	0
50	.054	.031	.004	0	0	.039	.031	.003	0	0
60	.052	.026	.001	0	0	.028	.033	.002	0	0
80	.047	.009	0	0	0	.042	.02	.001	0	0
100	.034	.004	0	0	0	.028	.009	0	0	0
Population Effect Size = .5										
30	.043	.006	0	0	0	.041	.013	0	0	0
40	.037	.002	0	0	0	.033	.005	0	0	0

Average Sample Size	Number of Studies									
	Fixed-effects Model					Random effects Model				
	5	10	20	50	80	5	10	20	50	80
50	.019	0	0	0	0	.026	.002	0	0	0
60	.007	0	0	0	0	.019	0	0	0	0
80	.001	0	0	0	0	.009	0	0	0	0
100	0	0	0	0	0	.005	0	0	0	0

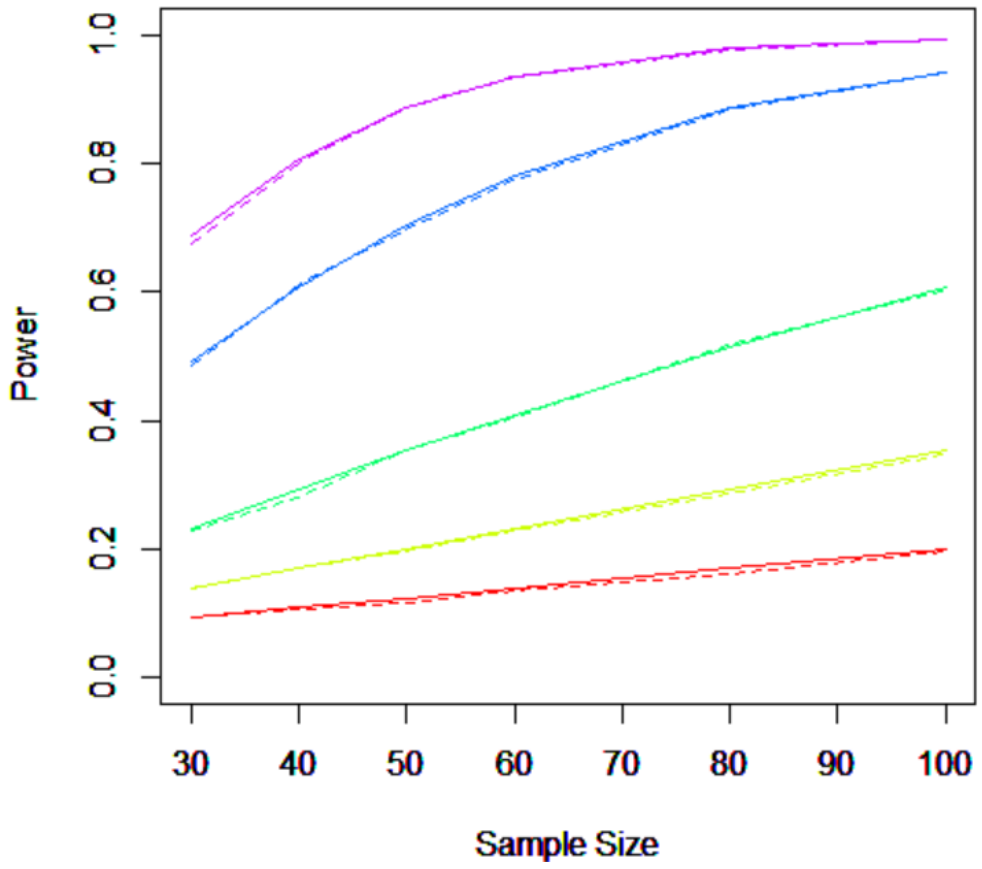

Figure 4.1 Power curves by sample size and number of studies (fixed-effects model equal sample size and balanced design)

Note: purple, blue, green, yellow, and red: large to small number of studies; solid lines: analytical power; dashed lines: simulated power; population effect size .1.

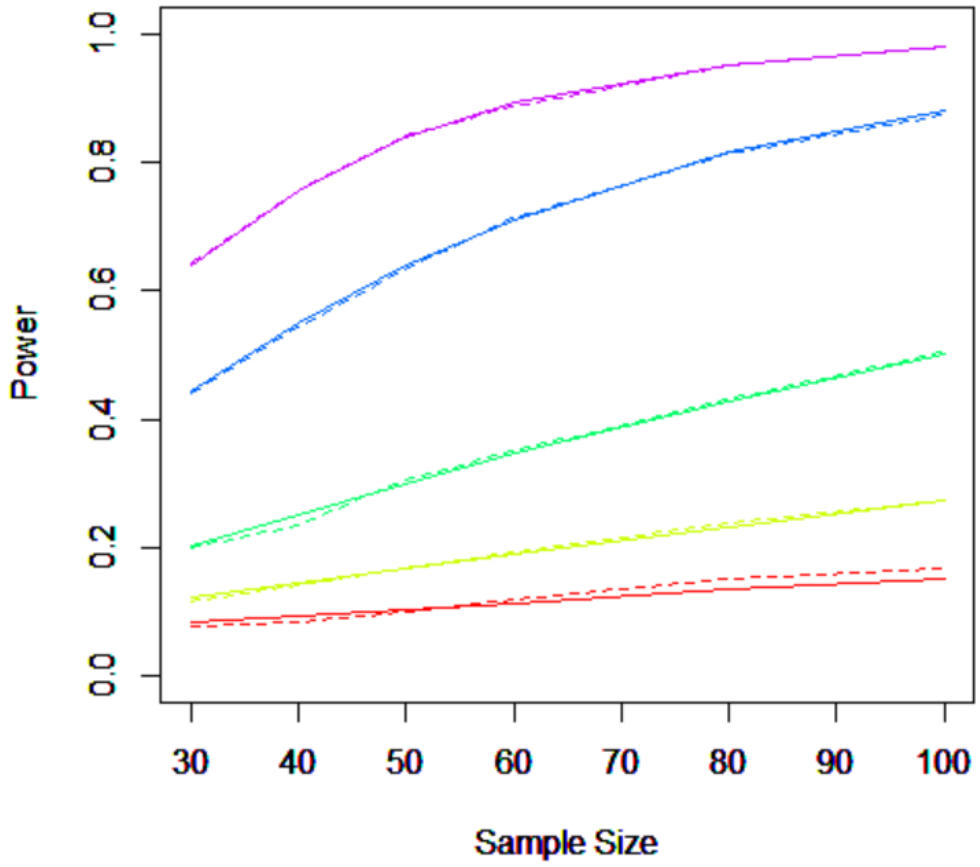

Figure 4.2 Power curves by sample size and number of studies (random-effects model equal sample size and balanced design)

Note: purple, blue, green, yellow, and red: large to small number of studies; dashed lines: simulated power; solid lines: analytical power; population effect size .1.

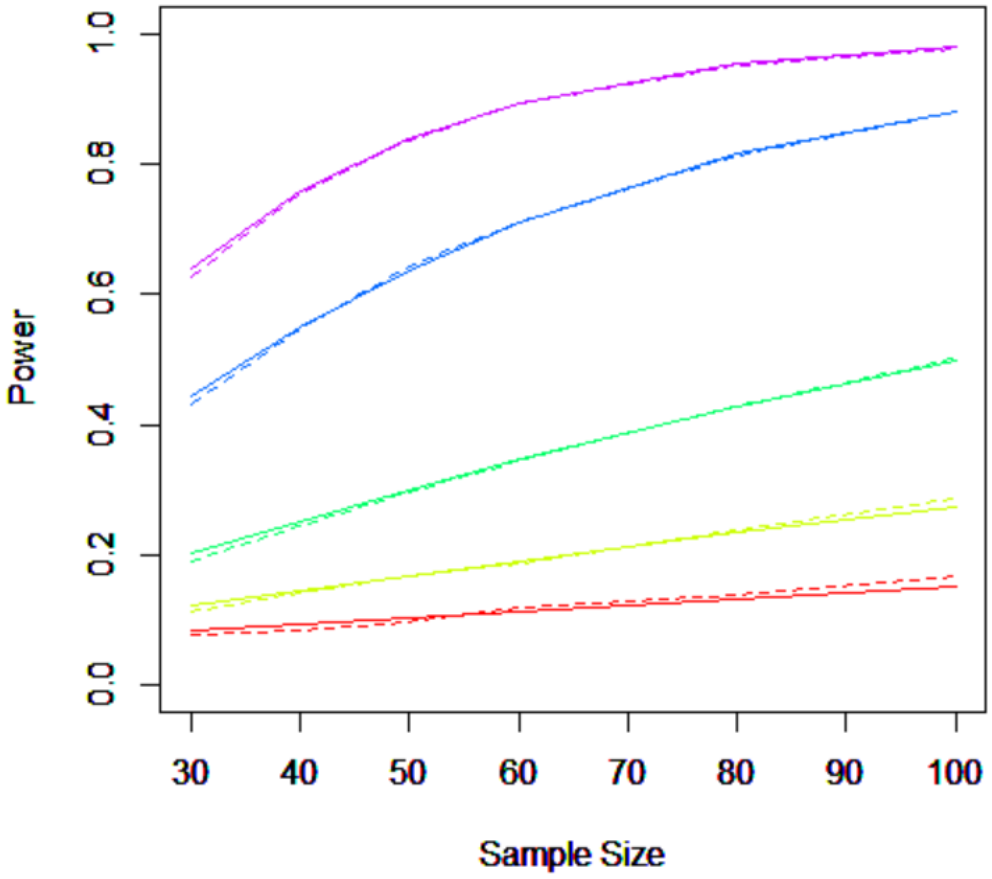

Figure 4.3 Power curves by sample size and number of studies (random-effects model unequal sample size and balanced design)

Note: purple, blue, green, yellow, and red: large to small number of studies; dashed lines: simulated power; solid lines: analytical power; population effect size .1.

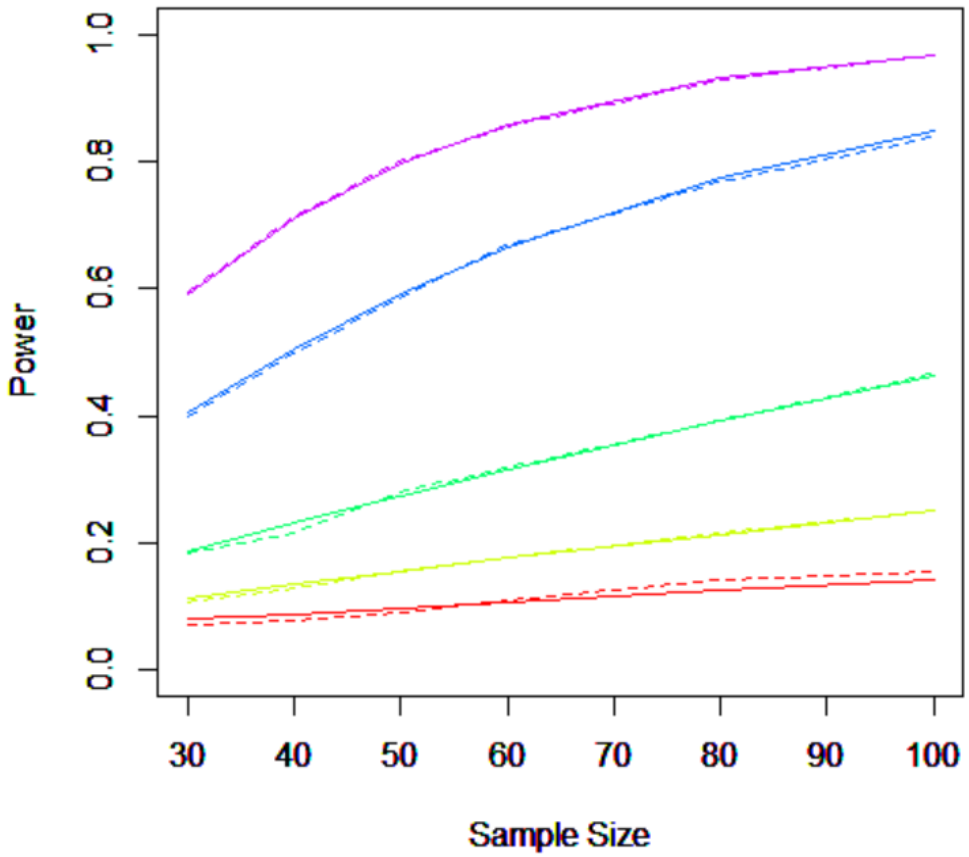

Figure 4.4 Power curves by sample size and number of studies (random-effects model equal sample size and unbalanced design)

Note: purple, blue, green, yellow, and red: large to small number of studies; dashed lines: simulated power; solid lines: analytical power; population effect size .1.

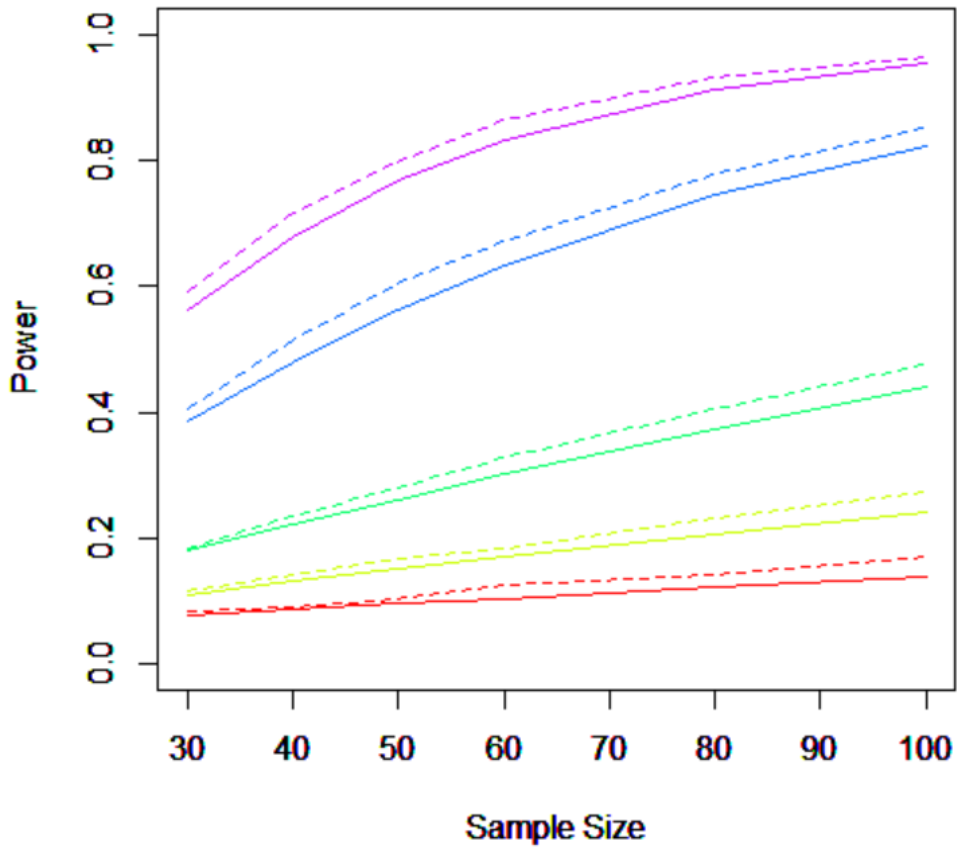

Figure 4.5 Power curves by sample size and number of studies (random-effects model unequal sample size and unbalanced design)

Note: purple, blue, green, yellow, and red: large to small number of studies; dashed lines: simulated power; solid lines: analytical power; population effect size .1.

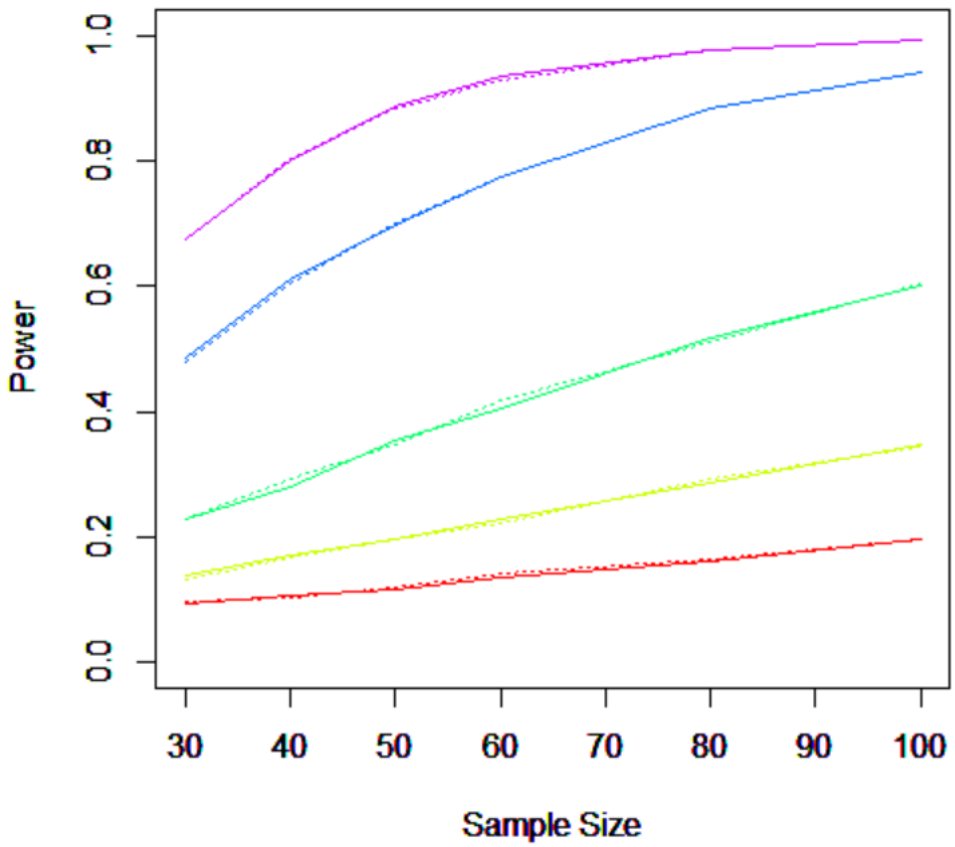

Figure 4.6 Power curves of the fixed-effects model

Note: purple, blue, green, yellow, and red: large to small number of studies; equal sample size – solid lines vs unequal sample size across studies – dotted lines; population effect size .1.

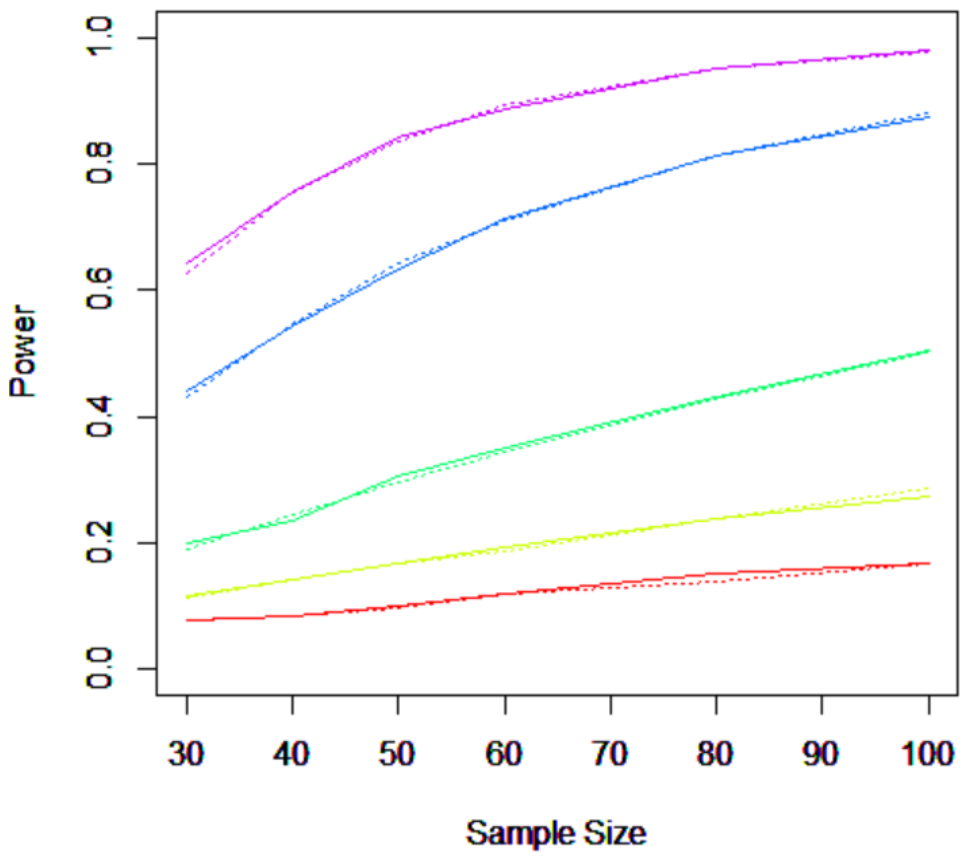

Figure 4.7 Power curves of the random-effects model

Note: purple, blue, green, yellow, and red: large to small number of studies; equal sample size – solid lines vs unequal sample size across studies – dotted Lines; population effect size .1.

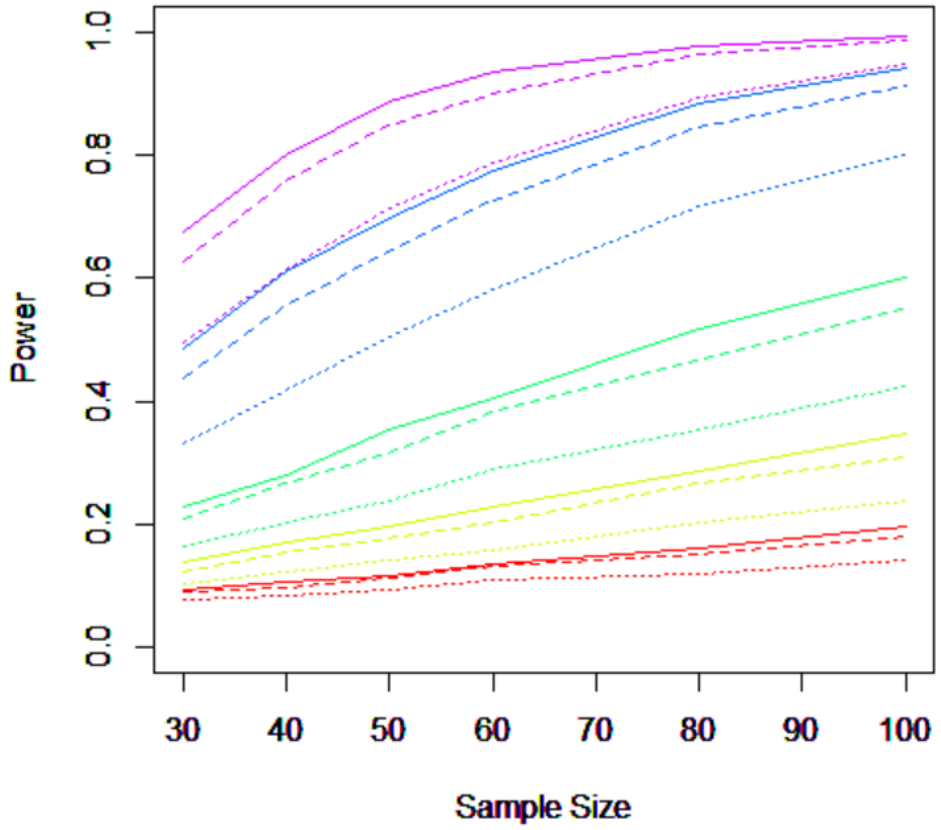

Figure 4.8 Power curves of the fixed-effects model

Note: purple, blue, green, yellow, and red: large to small number of studies; balanced design – solid lines; average sample size ratio: 1:2 – dashed lines; average sample size ratio: 1:4 – dotted lines; population effect size .1.

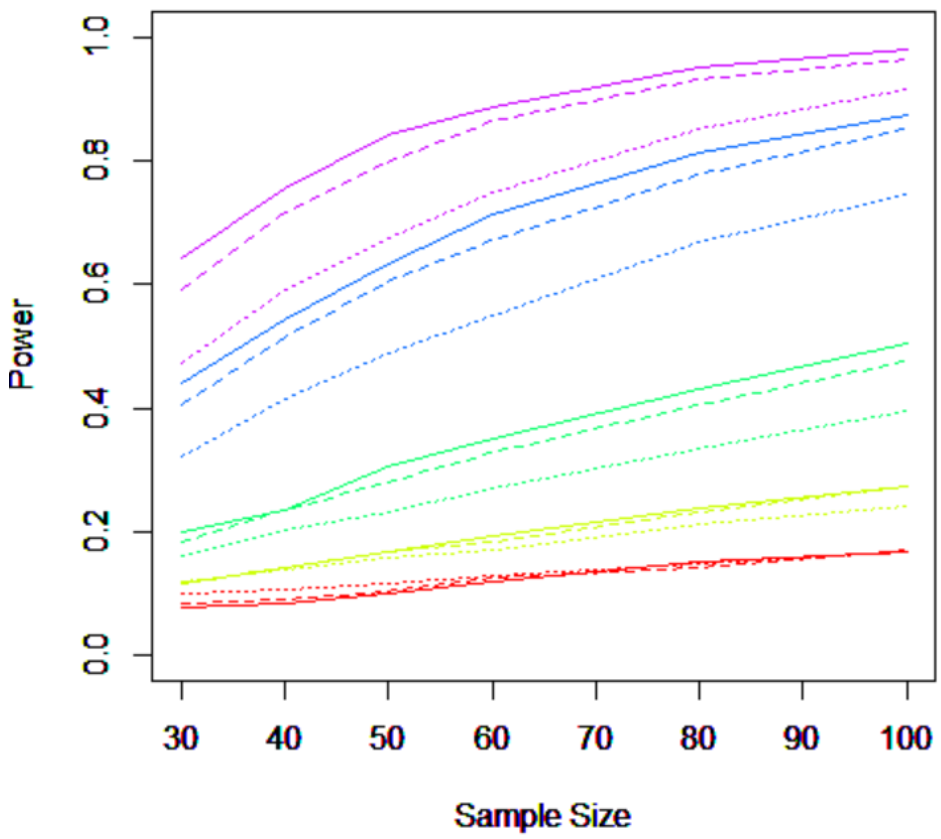

Figure 4.9 Power curves of the random-effects model

Note: purple, blue, green, yellow, and red: large to small number of studies; equal sample size – solid lines; average sample size ratio: 1:2 – dashed lines; average sample size ratio: 1:4 – dotted lines; population effect size .1.

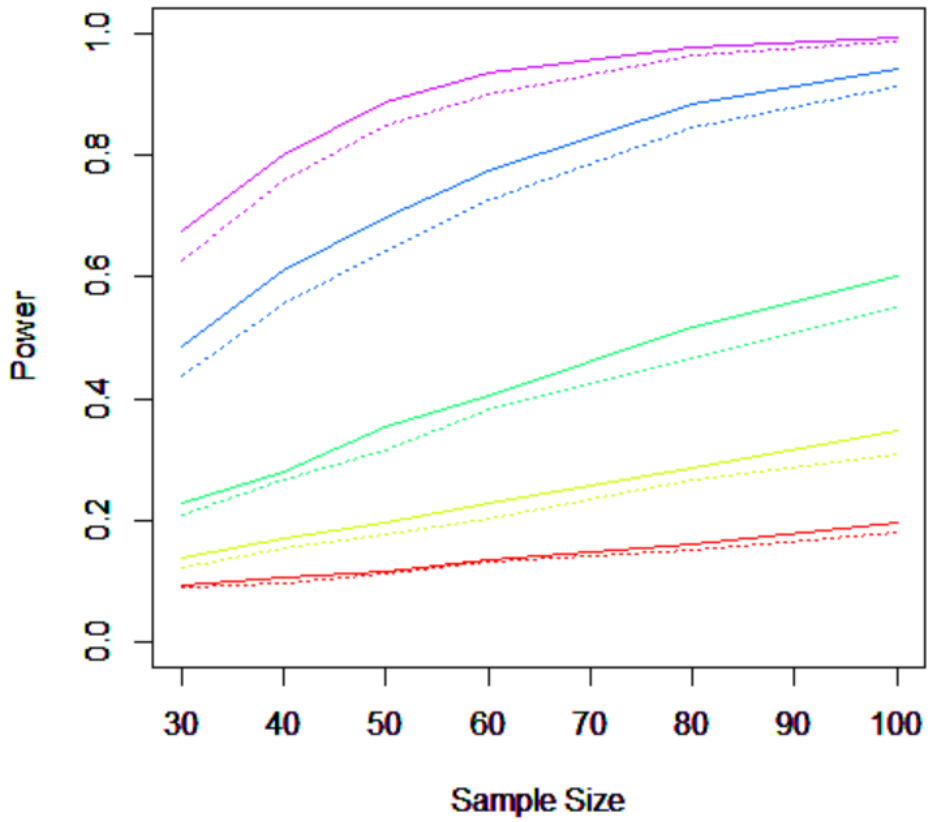

Figure 4.10 Power curves of the fixed-effects model

Note: purple, blue, green, yellow, and red: large to small number of studies; equal sample size – solid lines vs unequal sample size across studies and unbalanced design – dotted lines; population effect size .1.

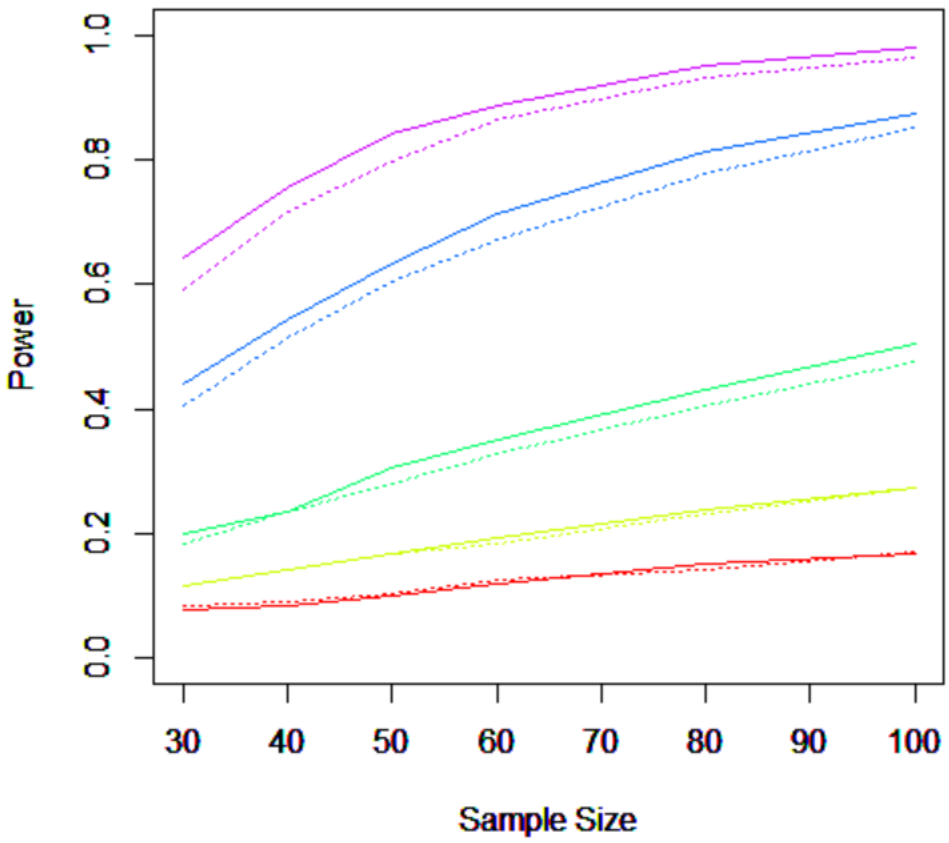

Figure 4.11 Power curves of the random-effects model

Note: purple, blue, green, yellow, and red: large to small number of studies; equal sample size – solid lines vs unequal sample size across studies and unbalanced design – dotted lines; population effect size .1.

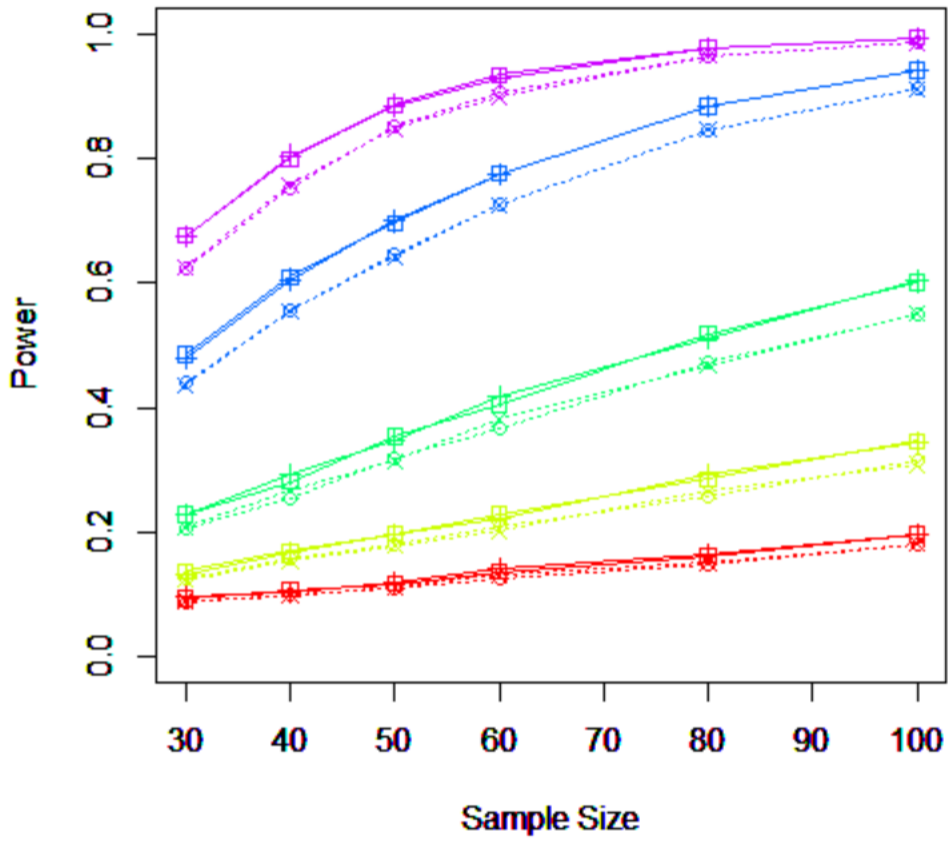

Figure 4.12 Power curves of the fixed-effects model

Note: purple, blue, green, yellow, and red: large to small number of studies; equal sample size and balanced design, unequal sample size across studies and balanced design – solid lines; equal sample across studies and unbalanced design, unequal sample size across studies and unbalanced design – dotted lines); population effect size .1.

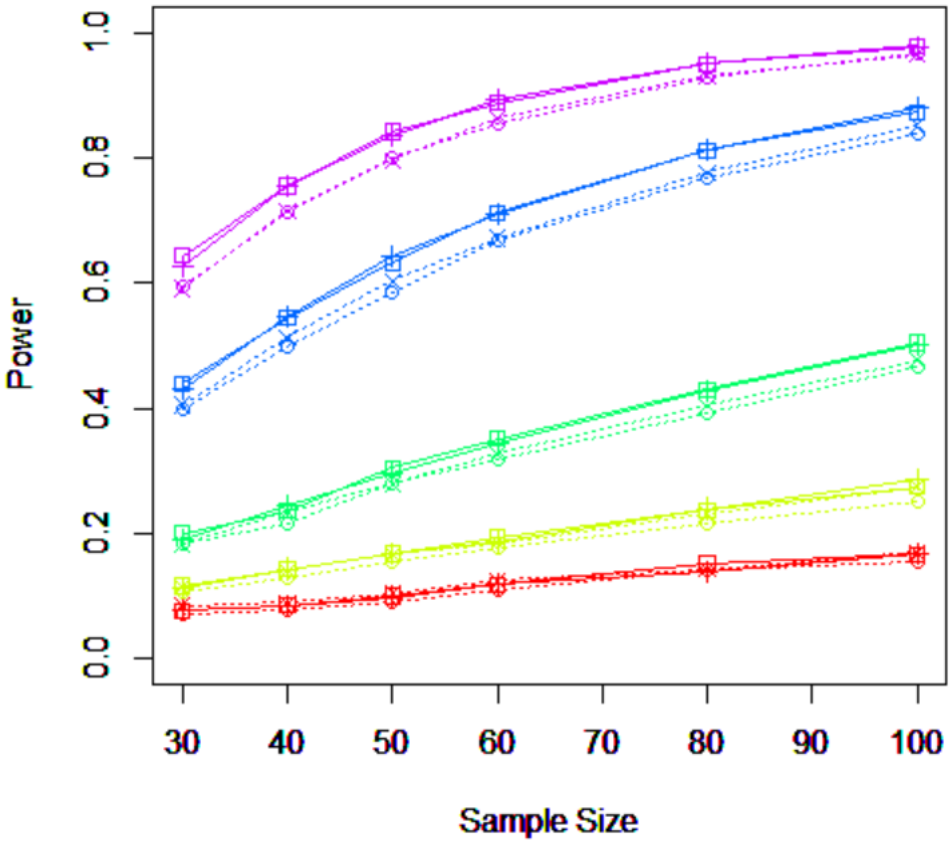

Figure 4.13 Power curves of the random-effects model

Note: purple, blue, green, yellow, and red: large to small number of studies; equal sample size and balanced design, unequal sample size across studies and balanced design – solid lines; equal sample across studies and balanced design, unequal sample size across studies and unbalanced design – dotted lines); population effect size .1.

CHAPTER 5

CONCLUSION

Meta-analysis has been used to synthesize research results of similar nature for several decades. There has been an increasing interest in using meta-analysis because it enables researchers to reconcile inconsistent findings from small studies on the same topic and reach a definitive answer to the research question of interest. The meta-analysis can overcome the limitation of small studies, which often lack sufficient statistical power.

There exists some literature on statistical power in meta-analysis. Problems with statistical power in meta-analysis have been addressed by researchers (e.g., Stern, Gavaghan & Egger, 2000). The current study investigated the discrepancy between the simulated power and analytical approximate power for the Hedge's g (corrected from Cohen's effect size d) under various conditions (i.e., varying average sample size, number of studies, and population effect size), using both the fixed and random-effects models. The influence of unequal sample size across studies and unbalanced design within studies on statistical power was analyzed and examined. The findings can potentially inform educational researchers about the actual statistical power in a planned meta-analysis.

The potential factors that influence statistical power are model selection, population effect size, number of involved studies, sample sizes of the studies, and design balance of those studies. The current study produced new findings about meta-analysis

and statistical power. A few findings not directly related to the research questions are briefly discussed as these may be useful information for researchers.

(1) The Hunter-Schmidt method manages Type I error poorly when the number of studies in the meta-analysis is small. This does not appear surprising because the method weighs the effect size by the number of studies. The smaller number of studies make it difficult to correctly reject the null hypothesis.

(2) When the population effect sizes greatly vary, the random-effect model should be used. Otherwise, the Type I error rate will not be controlled properly. This is especially true when there is a large amount of variation in the effect sizes among studies. Selecting an appropriate model is critical for the correct estimation of power in a meta-analysis.

(3) In the pilot run, Hedge's g does help decreasing the difference between the simulated power and the analytical power under certain conditions compared with the Cohen's d. Thus, Hedge's g was selected as the effect size index. However, the simulated power between Cohen's d and Hedge's g is similar overall. It does not influence the main conclusions of the current study.

(4) The power discrepancies between simulated and analytical power in the fixed-effects model were minimal (.01 or below). Thus, the power formulas for the fixed effects model should be able to provide accurate estimates. However, the simulated power and analytical power may show noticeable discrepancies in certain selected conditions in the random-effects model. Certain adjustment can be made to address the discrepancy (i.e., employ power simulation). In the random-effects model, the power discrepancy is negligible when the power is high enough under certain conditions.

(5) The unbalanced design does influence the statistical power in meta-analysis as it does in a primary single study. This is more pronounced with high design imbalance than with low design imbalance. The latter case only shows minor change in statistical power, compared with that of balanced design. The influence of unequal sample size across studies is minimal.

There are other considerations when planning a meta-analysis. The first is to decide whether effect sizes should be treated as fixed or random.

First, researchers need to choose a fixed-effects model or random-effects model for the meta-analysis. The literature review reveals that the random-effects models have become increasingly popular recently (Hall & Brannick, 2002). As cited by Field (2001), it is more likely to have datasets with varied effect sizes across studies. The assumption of fixed population effect size is tenable only when researchers do not intend to generalize the results beyond the datasets. For example, the researchers include most of the representative datasets in their meta-analysis, and they do not need to generalize the results. Researchers may choose a fixed-effects model or random-effects model by calculating the Q statistics, which can be used as a reference to decide if the population effect sizes are fixed across studies, but it should be considered in conjunction with other criteria, such as the generalizability of the meta-analysis results. Researchers could opt to conduct power analysis, using both fixed and random-effects model. By doing so, they can make an informed decision if they are not sure about heterogeneity of the dataset.

Secondly, researchers need to collect and estimate the parameter values necessary for power analysis, after deciding the appropriate statistical model. The average sample size of the individual studies and the number of studies are easy to estimate, as long as

researchers have access to the original datasets. As mentioned in the introduction, one difficulty in power analysis is the correct estimation of the population effect size. In theory, researchers cannot obtain 100% accurate population effect size, but a relatively accurate estimate can be obtained. There are reference books and research articles on different research topics. They often report the effect sizes from the previous research. For example, Hattie (2009) synthesized over 800 meta-analysis studies related to student achievements of various kinds. The effect sizes on different outcomes were obtained and discussed in detail. The effect sizes in the relevant literature varied greatly, ranging from negative values to large positive values. If researchers are interested in achievement related studies, Hattie's book offers a good resource to estimate the population effect size from related studies. Estimating the population effect size from the dataset researchers analyze is not a good practice. An alternative way is to report the confidence interval of the effect size estimates from the dataset. The upper and lower bound can be used to calculate the statistical power. After all the parameters are estimated, power analysis can be performed, according to the formulas in Chapter 2.

When conducting power analysis, researchers can consider varying population effect size, number of studies, and average sample size. Low statistical power in meta-analysis exists when all the parameters are small as shown in the power tables in Chapter 3 and Chapter 4. They also need to consider the influence of unbalanced design on statistical power by calculating the average sample size ratio between two groups. More unbalanced design will lead to lower statistical power. The following recommendations were made. If researchers are certain about the large population effect size in a meta-analysis project (.8 or above), researchers are likely to attain sufficient statistical power

no matter what other parameters they have in their studies. They do not need to consider the probability of making Type II errors. Low statistical power may be a concern when the population effect size is .5 or below. Only one design – unequal sample size across studies and unbalanced design – was considered, because this design was mostly close the real situation. The following table included the settings that are needed to achieve power of .8, which is the general cut-off score of ideal power. The recommendations are rough numbers based on the selected conditions in the current study. The analytical power and simulated power generally indicated similar conclusions due to the minimal discrepancies (Table 5.1). Values lower than 30 indicated that 30 was large enough to receive power .8; values higher than 100 indicated that 100 was not large enough to receive power .8. Slightly larger sample size is needed for the random-effects model under certain conditions (e.g., population effect size: .5 and number of studies: 5). Admittedly, these were not exactly the same conditions as what we have in practice. The developed simulation code can be employed to analyze statistical power in meta-analysis for different parameter values, varying degree of design balance and unequal sample sizes.

It should be noted that the current study does not include all possible scenarios in terms of average sample size, number of studies, or population effect size. However, this limitation can be easily overcome by initiating a simulation study that incorporates any new considerations. As demonstrated in the current study, simulation has proved to be a very efficient way to study and understand the performance of statistical power in a real meta-analysis.

The effect sizes are generated from the *t* distribution. The assumption of *t* distribution may not be holding true for small sample size under 30 in small studies. Thus, the results of average sample size of fewer than 30 were not considered in the current study. However, studies with unequal sample size may still have studies with lower sample size. More simulation studies are needed to understand the effect sizes that have other distribution properties. In addition, the current study uses an average sample size ratio between the two groups for design balance. In reality, the sample size ratio between the two groups can vary from one individual study to another. This situation can be examined in future simulation studies. Future studies in this area can extend to new possible study configurations as they arise from a meta-analysis. The developed R code can be adapted to accommodate those new considerations. Another practical thing is to develop a SAS macro that can simulate and calculate statistical power. Practical researchers can assign parameters and receive two power estimates simultaneously.

It is hoped that the current study helps to motivate further research aiming at examining statistical power in more complicated meta-analyses. Given the urge for meta-analysis in social science research, the current study essentially offers a stepping stone for more advanced analysis. Further research can examine statistical power in testing moderator effects and publication bias effect in meta-analysis. For example, there are differences in math achievement between female and male students, but such differences may depend on the grade levels. The moderating effect of grade level on gender difference can be of great interest, and so is the statistical power for testing the moderating effect. Analyzing power for testing a moderating effect can lead to a new line of research in this area. Another promising area will be the power for testing

publication bias in meta-analysis, which refers to the fact that studies with significant results are more likely to be published. This issue often surfaces in meta-analyses.

Table 5.1 *Sample Size Needed to Receive Power of .8*

Fixed-Effects Model					
Population Effect Size	Number of Studies				
	5	10	20	50	80
0.5	30	< 30	< 30	< 30	< 30
0.3	80	40	< 30	< 30	< 30
0.2	> 100	100	50	< 30	< 30
0.1	> 100	> 100	> 100	80	50
Random-Effects Model					
Average Sample Size	Number of Studies				
	5	10	20	50	80
0.5	40	< 30	< 30	< 30	< 30
0.3	> 100	50	< 30	< 30	< 30
0.2	> 100	> 100	60	< 30	< 30
0.1	> 100	> 100	> 100	100	60

REFERENCES

Bhattacherjee, A. (2012). *Social science research: principles, methods, and practices*. *Textbooks Collection.* Book 3. From http://scholarcommons.usf.edu/oa_textbooks/3

Brezau, S. and R. Graves (2001), Statistical power and effect sizes of clinical neuropsychology research. Journal of Clinical and Experimental Neuropsychology, 23(3): 399-406.

Borenstein, M., Hedges, L. V., Higgins, J.P.T, & Rothstein, H.R. (2009). *Introduction to meta-analysis*. Chichester, U.K. : John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1, 97-111.

Cafri, G., Kromrey, J. D., & Brannick, M.T. (2010). A Meta-Meta-Analysis: Empirical Review of Statistical Power, Type I Error Rates, Effect Sizes, and Model Selection of Meta-Analyses Published in Psychology, *Multivariate Behavioral Research, 45*(2), 239-27.

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

Cohen, J. (1990). "Things I have learned (so far)," *American Psychologist*, 45(12): 1304-1312.

Cohen, J. (1992). Statistical Power analysis .Current Directions in *Psychological Science*, 1(3), 98–101.

Cohen, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, 8(3), 243–253.

Cook, D.A., & Hatala, R. (2014). Got power? A systematic review of sample size adequacy in health professions education research. *Advances in Health Sciences Education*. From DOI: 1.1007/s10459-014-9509-5

Ellis, P.D. (2010). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*. New York, NY : Cambridge University Press.

Else-Quest, N. M., Higgins, A., Allison, C., & Morton, L. C. (2012). Gender differences in self-conscious emotional experience: A meta-analysis. *Psychological Bulletin, 138,* 947-981.

Field. A.P. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6(2). 161–18.

Field. A.P. (2003). The problems of using fixed-effects models of meta-analysis on real-world data. *Psychological Methods*, 6(2). 161–18.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. Educational Researchers, 5, 3-8.

Hall, S. M., & Brannick, M. T. (2002). *Comparison of Two Random-Effects Methods of Meta-Analysis*. *87*(2), 377-389.

Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. *New York: Routledge.*

Hsu, L., (1994). Unbalanced Designs to Maximize Statistical Power in Psychotherapy Efficacy Studies. *Psychotherapy Research. 4*(2), 95-106.

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., & Olkin, I. (1985). *Statistical models for meta-analysis*. New York: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6, 203–217.

Hedges, L. V., & Vevea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504.

Hogan, J.W., Roy, J., & Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine,* 23:1455–1497 from DOI: 1.1002/sim.1728.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed-effects vs. random-effects meta-analysis models: Implications for cumulative knowledge in psychology. *International journal of Selection and Assessment, 8*(4), 275-292.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research findings (2nd ed.). Newbury Park, CA: Sage.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random-effects models in meta-analysis. Psychological Methods, 3, 486-501.

Hutchinson, S.R. & Bandalos, D.L. (1997). A guide to Monte Carlo simulations for applied researchers. *Journal of Vocational Education Research, 22(4),* 233-245.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33, 14–26. doi:10.3102/0013189X033007014

Kraemer, H.C., Yesavage, J., & Brooks, J.O. (1998). The advantages of excluding under-powered studies in meta-analysis: Inclusions vs exclusionist viewpoints. *Psychological Methods,* 3(1): 23-31.

Liu, X. (2013). *Statistical Power Analysis for the Social and Behavioral Sciences: Basic and Advanced Techniques*. New York, NY: Routledge.

Lindsay, R.M. (1993). Incorporating statistical power into the test of significance procedure: A methodological and empirical inquiry. *Behavioral Research in Accounting*, 5: 211-236.

Lipsey, M., & Hurley, S. (2009). Design sensitivity: Statistical power for applied experimental research. In L. Bickman, & D. Rog (Eds.), *The SAGE handbook of applied social research methods.* (2nd ed., pp. 44-77). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/1.4135/9781483348858.n2

Mone, M.A., Mueller, G.C.,&Mauland,W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49*, 103-12.

Mooney, C. Z. (1997). *Monte Carlo simulation* [Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-116]. Thousand Oaks, CA: Sage.

Murphy, K.R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (2nd Ed).* Mahwah, NJ: Lawrence Erlbaum.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods,* 5, 241-301.

Normand, S.T. (1999). Tutorial in Biostatistics: Meta-analysis: formulating, evaluating, combining and reporting. *Statistics in Medicine*. 18. 321-359.

Rossi, J.S. (1990). Statistical power of psychological research: what have we gained in 20 years?" *Journal of Consulting and Clinical Psychology*, 58(5): 646-656.

Sterne, J.A.,Gavagham, S., & Egger, M.(2000). Publication and related bias in meta-analysi" Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*. 53. 1119-1129.

Trochim, W. (2000). The Research Methods Knowledge Base, 2nd Edition. Atomic Dog Publishing, Cincinnati, OH.

Thoms, L. (1997). Retrospective power analysis. *Conservation Biology*, 11(1): 276-28.

Voyer, D. (2011). Time limits and gender differences on paper and-pencil tests of mental rotation: a meta-analysis. *Psychonomic Bulletin Review*. 18. 267–277 from DOI 1.3758/s13423-010-0042-0

APPENDIX A

R CODE

Basic Power Simulation (Chapter 2)

```
PopulationEffect<-0.2
SD<-1
Simultime<-1000
Samplesize<-100
pv<-rep(NA,Simultime)
for (i in 1: Simultime)
{print (i)
SimuValues<-rnorm(Samplesize, PopulationEffect, SD)
 pv[i]<-t.test(SimuValues,alternative= "two.sided",
mu=0)$p.value
     }
mean(pv<.05)
```

Meta-analysis Application (Chapter 3)

```
# Read in the dataset
Mydata<-read.csv("DIRECTORY OF THE FILE",header=TRUE)
Fnumber<-Mydata[,1]
Mnumber<-Mydata[,2]
ES<- Mydata[,3]
NumberStudy<-6

# Meta-analysis in the fixed-effects model
Variancewithin<-
((Fnumber+Mnumber)/(Fnumber*Mnumber))+((ES*ES*.5)/(Fnumber+
Mnumber))
Weight<-1/Variancewithin
SumWeight<-sum(Weight)
SumWd<-sum(Weight*ES)
WeightedD<- SumWd/SumWeight
SEM<-sqrt(1/SumWeight)
Zstat<- WeightedD/SEM
```

```
p.value<- 2*pnorm(-abs(Zstat))
Upper<- WeightedD+1.96* SEM
Lower<- WeightedD-1.96* SEM

# Meta-analysis in the random-effects model-HP
Qstat<- SumWdsquare-(SumWd* SumWd)/SumWeight
Qstat<-10
Cstat<-SumWeight-( SumWsquare/ SumWeight)
df<- NumberStudy -1
if(Qstat-df>0){Tsquare<-(Qstat-df)/Cstat} else {Tsquare<-0}
BetweenStudyVariance<-rep(Tsquare, NumberStudy)
VarianceTotal<- BetweenStudyVariance+ Variancewithin
WeightRandom<- 1/ VarianceTotal
SumWeightRandom<-sum(WeightRandom)
SumWeightRandomd<-sum(WeightRandom*ES)
WeightdRandom<- SumWeightRandomd/SumWeightRandom
SEMRandom<-sqrt(1/SumWeightRandom)
ZstatRandom<- WeightdRandom/SEMRandom
p.valueRandom<- 2*pnorm(-abs(ZstatRandom))
UpperRandom<- WeightdRandom+1.96* SEMRandom
LowerRandom<- WeightdRandom-1.96* SEMRandom

# Meta-analysis in the random-effects model-HS
WeightRandomHS<- Fnumber+Mnumber
SumWeightRandomHS<-sum(WeightRandomHS)
SumWeightRandomdHS<-sum(WeightRandomHS*ES)
WeightdRandomHS<- SumWeightRandomdHS/ SumWeightRandomHS
WeightdRandomHSmatrix<- rep(WeightdRandomHS, NumberStudy)
NomiVariance<-sum((WeightRandomHS)*(ES-
WeightdRandomHSmatrix)* (ES- WeightdRandomHSmatrix))
DenomiVariance<- SumWeightRandomHS
VarianceHS<-(NomiVariance/ DenomiVariance)/NumberStudy
SEMRandomHS<-sqrt(VarianceHS)
ZstatRandomHS<- WeightdRandomHS/SEMRandomHS
p.valueRandomHS<- 2*pnorm(-abs(ZstatRandomHS))
UpperRandomHS<- WeightdRandomHS +1.96* SEMRandomHS
LowerRandomHS<- WeightdRandomHS - 1.96* SEMRandomHS

# Power Functions
# Define number of studies, population effect size and
sample size

#Fixed-effects model
 Fix=function(N1, N2, NumberStudy, PES)
    {
    Vtotal<- (N1+ N2)/( N1* N2)+.5* PES^2/( N1+ N2)
    lamda<-sqrt(NumberStudy)* PES/ sqrt(Vtotal)
```

```
   power<-pnorm(lamda-qnorm(1-.05/2))+pnorm(qnorm(.05/2)-
lamda)
     return(power)
    }

#Random-effects model (Tau square)
    RandomTsquare=function(N1, N2, NumberStudy, PES)
    {
    Vtotal<- (N1+ N2)/( N1* N2)+.5* PES^2/(N1+ N2)+Tsquare
    lamda<-sqrt(NumberStudy)*PES/sqrt(Vtotal)
   power<-pnorm(lamda-qnorm(1-.05/2))+
pnorm(qnorm(.05/2)-lamda)
    return(power)
    }

#Random-effects model (Ratio)
#Define it as the ratio of within group variance and
between group variance
#Define it as small & medium & large(.33, .67, 1.0)
   p<-c(1.33,1.67,2)
   RandomRatio=function (N1, N2, NumberStudy,PES,p)
   {
     Vtotal<- ((N1+ N2)/( N1* N2)+.5* PES^2/( N1+ N2))*p
     lamda<-sqrt(NumberStudy)* PES/ sqrt(Vtotal)
     power<-pnorm(lamda-qnorm(1-.05/2))+
pnorm(qnorm(.05/2)-lamda)
    return(power)
    }
```

Simulation and Analytical Power – Equal Sample Size and Balanced Design

```
rm(list=ls())
#Define all the parameters. Same for all designs
#sample size
possible.ns <- c(30,40,50,60,80,100)
#Number of studies
I.ns <- c(5,10,20,50,80)
# Set Type I error rate as .05(fixed)
alpha <- 0.05
# number of simulation iterations(fixed)
sims <- 10000
#Population effect size (set as 0,.1,.2,.3,.5,.8)
PES <-0.1
```

Fixed-Effects Model

```
#Define the seed to receive the same results in each run
set.seed(1000)
#####################################################
main<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)

# set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
p.value<-as.numeric(rep(NA,sims))

#looping at different average sample size
for (j in 1:n){
N<- possible.ns[j]
#looping at different number of studies
for (k in 1:s){
I<- I.ns[k]
#Simulation loop
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Sample size across studies equal in this condition
Nvary<-rep(N,I)
# Simulate the effect size using t distribution
# Sample size between two groups in each study are equal
d0 <- rt(I,Nvary-2)*2*sqrt(1/Nvary)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PES
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-(4/Nvary)*(1+0.125*ES*ES)
Varianceg<-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
SumWd<-sum(Weight*ES)
WeightedD<- SumWd/SumWeight
SEM<-sqrt(1/SumWeight)
Zstat<- WeightedD/SEM
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(Zstat))
significant.experiments[i] <- ifelse(p.value[i] <=
```

```
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)
}
}
out <- list(prob)
names(out) <- c("Real Type I error rate & Power")
out
}
FIX<-main(possible.ns,I.ns,PES,sims,alpha)
FIX

#Power function
FixPowfunction<-function(possible.ns, I.ns, PES)
    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
   power <- array(rep(NA,n*s),dim=c(n,s))
   #looping at different sample size
   for (j in 1:n){
   N <- possible.ns[j]
   #looping at different number of studies
   for (k in 1:s){
   I<- I.ns[k]
   Vtotal<-(4/N)*(1+0.125*PES*PES)
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
}
}
   return(powerround)
    }
FixPowerFunction<-FixPowfunction(possible.ns,I.ns, PES)
FixPowerFunction
```

Random-effects Model

```
#Hedges & Colleagues Method

set.seed(1000)
##########################################################
Random_1<-function(possible.ns, I.ns, PES, sims, alpha)
{
```

```
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)

# Set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
Tsquare_ave<- array(rep(NA,n*s),dim=c(n,s))
Tsquare.array<-as.numeric(rep(NA,sims))
p.value<-as.numeric(rep(NA,sims))

#loop for different average sample size
for (j in 1:n){
N <- possible.ns[j]
#loop for different number of studies
for (k in 1:s){
I<- I.ns[k]
# Simulation loop
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Sample size across studies equal in this condition
Nvary<-rep(N,I)
# Simulate the effect size using t distribution
# Sample size between two groups in each study are equal
d0 <- rt(I,Nvary-2)*2*sqrt(1/Nvary)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
# Vary the population effect size of each study to meet the
random-effects model assumption
PESVARY<-rnorm(I,PES,0.1)
ES<- g + PESVARY
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-(4/Nvary)*(1+0.125*ES*ES)
Varianceg<-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
SumWd<-sum(Weight*ES)
SumWdsquare<-sum(Weight*ES*ES)
SumWsquare<-sum(Weight*Weight)
Qstat<- SumWdsquare-(SumWd*SumWd)/SumWeight
Cstat<-SumWeight-(SumWsquare/SumWeight)
df<- I -1
#Use if function to define Tsquare (Between-study variance)
if(Qstat-df>0){Tsquare<-(Qstat-df)/Cstat} else {Tsquare<-0}
```

```
if(Qstat-df>0){Tsquare.array[i]<-(Qstat-df)/Cstat} else
{Tsquare.array[i]<-0}
BetweenStudyVariance<-rep(Tsquare,I)
VarianceTotal<- BetweenStudyVariance+ Varianceg
WeightRandom<- 1/VarianceTotal
SumWeightRandom<-sum(WeightRandom)
SumWeightRandomd<-sum(WeightRandom*ES)
WeightdRandom<- SumWeightRandomd/SumWeightRandom
SEMRandom<-sqrt(1/SumWeightRandom)
ZstatRandom<- WeightdRandom/SEMRandom
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandom))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)
Tsquare_ave[j,k]<- mean(Tsquare.array)
}
}
out <- list(prob,Tsquare_ave)
names(out) <- c("Real Type I error rate & Power","average T
square")
out
}
Random_HP<- Random_1(possible.ns,I.ns,PES,sims,alpha)
Random_HP[[1]]

#Hunter&Schmidt Method

set.seed (1000)
############################################################
Random_2<-function(possible.ns, I.ns, PES, sim, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)

# set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
p.value<-as.numeric(rep(NA,sims))

#loop for different average sample size
```

```
for (j in 1:n){
N <- possible.ns[j]
#looping for different number of studies
for (k in 1:s){
I<- I.ns[k]
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Sample size across studies equal in this condition
Nvary<-rep(N,I)
# Simulate the effect size using t distribution
# Sample size between two groups in each study are equal
d0 <- rt(I,Nvary-2)*2*sqrt(1/Nvary)
# Vary the population effect size of each study to meet the
random-effects model assumption
PESVARY<-rnorm(I,PES,0.1)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PESVARY
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
WeightRandomHS<- Nvary
SumWeightRandomHS<-sum(WeightRandomHS)
SumWeightRandomdHS<-sum(WeightRandomHS*ES)
WeightdRandomHS<- SumWeightRandomdHS/SumWeightRandomHS
WeightdRandomHSmatrix<-rep(WeightdRandomHS,I)
NomiVariance<-sum((WeightRandomHS)*(ES-
WeightdRandomHSmatrix)* (ES- WeightdRandomHSmatrix))
DenomiVariance<-SumWeightRandomHS
VarianceHS<-(NomiVariance/DenomiVariance)/I
SEMRandomHS<-sqrt(VarianceHS)
ZstatRandomHS<- WeightdRandomHS/SEMRandomHS
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandomHS))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)

}
}
   out <- list(prob)
names(out) <- c("Real Type I error rate & Power")
out
}
```

```
Random_HS<- Random_2(possible.ns,I.ns,PES,sims,alpha)
Random_HS[[1]]

#Power function

    AveTsquare<-Random_HP[[2]]
    RandomPowfunction<-function(possible.ns, I.ns, PES)
    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
      power <- array(rep(NA,n*s),dim=c(n,s))
   #looping at different sample size
   for (j in 1:n){
   N <- possible.ns[j]
   #looping at different number of studies
   for (k in 1:s){
   I<- I.ns[k]
   Variancewithin<-(4/N)*(1+0.125*PES*PES)
   Tsquare<- AveTsquare[j,k]
   Vtotal<-Variancewithin+Tsquare
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
}
}
   return(powerround)
    }
RandomPowerFunction<-
RandomPowfunction(possible.ns,I.ns,PES)
RandomPowerFunction
```

Simulation and Analytical Power – Unequal Sample Size and Balanced Design

Fixed-effects Model

```
set.seed(1000)
##########################################################
main<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)
```

```
# set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
p.value<-as.numeric(rep(NA,sims))

#loop for different average sample size
for (j in 1:n){
N<- possible.ns[j]
#loop for different number of studies
for (k in 1:s){
I<- I.ns[k]
#simulation loop
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Use the truncated binomial distribution to simulate
sample size
try<-function(p,m,c){
   (p/(1-(1-p)^c) - m/c )^2}
#Can vary the maximum value (standard deviation varies)
MaxN<-N*3
#Get the p.value (0 point)
p.to.use<-
optimize(try,interval=c(0.0001,0.9999),m=N,c=MaxN)$minimum
#Simulate the sample size (Nvary)
Nvary<-rbinom(I, MaxN,p.to.use)
nb<-sum(Nvary ==0)
while (nb>0){
Nvary [Nvary ==0]<-rbinom(nb,maxss,p.to.use)
nb<-sum(Nvary ==0)}
# Simulate the effect size using t distribution
# Sample size between two groups in each study are equal
d0 <- rt(I,Nvary-2)*2*sqrt(1/Nvary)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PES
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-(4/Nvary)*(1+0.125*ES*ES)
Varianceg<-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
SumWd<-sum(Weight* ES)
WeightedD<- SumWd/SumWeight
SEM<-sqrt(1/SumWeight)
Zstat<- WeightedD/SEM
#Return the p values of all simulations
```

```
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(Zstat))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)
}
}
out <- list(prob)
names(out) <- c("Real Type I error rate & Power")
out
}
FIX<-main(possible.ns,I.ns,PES,sims,alpha)
FIX

# Power function

FixPowfunction<-function(possible.ns, I.ns, PES)
    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
   power<- array(rep(NA,n*s),dim=c(n,s))
   #loop for different average sample size
   for (j in 1:n){
   N <- possible.ns[j]
   #loop for different number of studies
   for (k in 1:s){
   I<- I.ns[k]
   Vtotal<-(4/N)*(1+0.125*PES*PES)
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
}
}
   return(powerround)
    }
FixPowerFunction<-FixPowfunction(possible.ns,I.ns, PES)
FixPowerFunction
```

Random-effects Model

```
#Hedges & Colleagues Method
```

```
set.seed(1000)
############################################################
Random_1<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)

# set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
Tsquare_ave<- array(rep(NA,n*s),dim=c(n,s))
Tsquare.array<-as.numeric(rep(NA,sims))
p.value<-as.numeric(rep(NA,sims))

#loop for different average sample size
for (j in 1:n){
N <- possible.ns[j]
#loop for different number of studies
for (k in 1:s){
I<- I.ns[k]
#simulation loop
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Use the truncated binomial distribution to simulate
sample size
try<-function(p,m,c){
   (p/(1-(1-p)^c) - m/c )^2}
#Can vary the maximum value (standard deviation varies)
MaxN<-N*3
#Get the p.value (0 point)
p.to.use<-
optimize(try,interval=c(0.0001,0.9999),m=N,c=MaxN)$minimum
Nvary<-rbinom(I, MaxN,p.to.use)
nb<-sum(Nvary ==0)
while (nb>0){
Nvary [Nvary ==0]<-rbinom(nb,maxss,p.to.use)
nb<-sum(Nvary ==0)}
# Simulate the effect size using t distribution
# Sample size between two groups in each study are equal
d0 <- rt(I,Nvary-2)*2*sqrt(1/Nvary)
# Vary the population effect size of each study to meet the
random-effects model assumption
PESVARY<-rnorm(I,PES,0.1)
J<-1-(3/(4*(Nvary-2)-1))
```

```
g<- d0*J
ES<-g+ PESVARY
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-(4/Nvary)*(1+0.125*ES*ES)
Varianceg<-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
SumWd<-sum(Weight*ES)
SumWdsquare<-sum(Weight*ES*ES)
SumWsquare<-sum(Weight*Weight)
Qstat<- SumWdsquare-(SumWd*SumWd)/SumWeight
Cstat<-SumWeight-(SumWsquare/SumWeight)
df<- I -1
#Use if function to define Tsquare(between-study variance)
if(Qstat-df>0){Tsquare<-(Qstat-df)/Cstat} else {Tsquare<-0}
if(Qstat-df>0){Tsquare.array[i]<-(Qstat-df)/Cstat} else
{Tsquare.array[i]<-0}
BetweenStudyVariance<-rep(Tsquare,I)
VarianceTotal<- BetweenStudyVariance+ Varianceg
WeightRandom<- 1/VarianceTotal
SumWeightRandom<-sum(WeightRandom)
SumWeightRandomd<-sum(WeightRandom*ES)
WeightdRandom<- SumWeightRandomd/SumWeightRandom
SEMRandom<-sqrt(1/SumWeightRandom)
ZstatRandom<- WeightdRandom/SEMRandom
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandom))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)
Tsquare_ave[j,k]<- mean(Tsquare.array)
}
}
   out <- list(prob,Tsquare_ave)
names(out) <- c("Real Type I error rate & Power","average T
square")
out
}
Random_HP<- Random_1(possible.ns,I.ns,PES,sims,alpha)
Random_HP[[1]]

#Hunter&Schmidt Method
set.seed (1000)
```

```
#############################################################
Random_2<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)


# set the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
p.value<-as.numeric(rep(NA,sims))

#loop for different average sample size
for (j in 1:n){
N <- possible.ns[j]
#loop for different number of studies
for (k in 1:s){
I<- I.ns[k]
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Use the truncated binomial distribution to simulate
sample size
try<-function(p,m,c){
   (p/(1-(1-p)^c) - m/c )^2}
#Can vary the maximum value (standard deviation varies)
MaxN<-N*3
#Get the p.value (0 point)
p.to.use<-
optimize(try,interval=c(0.0001,0.9999),m=N,c=MaxN)$minimum
Nvary<-rbinom(I, MaxN,p.to.use)
nb<-sum(Nvary ==0)
while (nb>0){
Nvary [Nvary ==0]<-rbinom(nb,maxss,p.to.use)
nb<-sum(Nvary ==0)}
# Simulate the effect size using t distribution
# Sample size between two groups in each study are equal
d0 <- rt(I,Nvary-2)*2*sqrt(1/Nvary)
# Vary the population effect size of each study to meet the
random-effects model assumption
PESVARY<-rnorm(I,PES,0.1)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<-g+ PESVARY
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
```

```
WeightRandomHS<- Nvary
SumWeightRandomHS<-sum(WeightRandomHS)
SumWeightRandomdHS<-sum(WeightRandomHS*ES)
WeightdRandomHS<- SumWeightRandomdHS/SumWeightRandomHS
WeightdRandomHSmatrix<-rep(WeightdRandomHS,I)
NomiVariance<-sum((WeightRandomHS)*(ES-
WeightdRandomHSmatrix)* (ES- WeightdRandomHSmatrix))
DenomiVariance<-SumWeightRandomHS
VarianceHS<-(NomiVariance/DenomiVariance)/I
SEMRandomHS<-sqrt(VarianceHS)
ZstatRandomHS<- WeightdRandomHS/SEMRandomHS
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandomHS))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)

}
}
   out <- list(prob)
names(out) <- c("Real Type I error rate & Power")
out
}

Random_HS<- Random_2(possible.ns,I.ns,PES,sims,alpha)
Random_HS[[1]]

#Power function

AveTsquare<-Random_HP[[2]]

RandomPowfunction<-function(possible.ns, I.ns, PES)

    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
   power <- array(rep(NA,n*s),dim=c(n,s))
   for (j in 1:n){
   N <- possible.ns[j]
   #added here for looping at number of studies
   for (k in 1:s){
   I<- I.ns[k]
   Variancewithin<-(4/N)*(1+0.125*PES*PES)
```

```
   Tsquare<- AveTsquare[j,k]
   Vtotal<-Variancewithin+Tsquare
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
}
}
   return(powerround)
    }
RandomPowerFunction<-
RandomPowfunction(possible.ns,I.ns,PES)
RandomPowerFunction
```

Simulated and Analytical Power – Equal Sample Size and Unbalanced Design

Fixed-effects Model

```
set.seed (1000)
############################################################
main<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)

# set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
p.value<-as.numeric(rep(NA,sims))

#loop for different average sample size
for (j in 1:n){
N<- possible.ns[j]
#loop for different number of studies
for (k in 1:s){
I<- I.ns[k]
#simulation loop
for (i in 1:sims){
# In each simulation, perform the meta-analysis
Nvary<-rep(N,I)
# Simulate the effect size using t distribution
# Vary the sample size within each study
N1<-Nvary*(1/3)
N2<-Nvary*(2/3)
d0<-rt(I,Nvary-2)*sqrt(1/N1+1/N2)
```

```
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PES
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-Nvary/(N1*N2)+(ES*ES*0.5)/Nvary
Varianceg <-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
SumWd<-sum(Weight* ES)
WeightedD<- SumWd/SumWeight
SEM<-sqrt(1/SumWeight)
Zstat<-WeightedD/SEM
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(Zstat))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)
}
}
out <- list(prob)
names(out) <- c("Real Type I error rate & power")
out
}
FIX<-main(possible.ns,I.ns,PES,sims,alpha)
FIX


# Power function

FixPowfunction<-function(possible.ns, I.ns, PES)
    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
power <- array(rep(NA,n*s),dim=c(n,s))
   #looping for different average sample size
   for (j in 1:n){
   N <- possible.ns[j]
   #looping for different number of studies
   for (k in 1:s){
   I<- I.ns[k]
N1<-N*(1/3)
```

```
N2<-N*(2/3)
Vtotal<-N/(N1*N2)+ (PES*PES*0.5)/N
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
}
}
   return(powerround)
    }
FixPowerFunction<-FixPowfunction(possible.ns,I.ns, PES)
FixPowerFunction
```

Random-effects Model

```
#Hedges & Colleagues Method

set.seed (1000)
############################################################
Random_1<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)

#set the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
Tsquare_ave<- array(rep(NA,n*s),dim=c(n,s))
Tsquare.array<-as.numeric(rep(NA,sims))
p.value<-as.numeric(rep(NA,sims))

#loop for different average sample size
for (j in 1:n){
N <- possible.ns[j]
#loop for different number of studies
for (k in 1:s){
I<- I.ns[k]
#simulation loop
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Sample size across studies equal in this condition
Nvary<-rep(N,I)
# Simulate the effect size using t distribution
```

```
# Sample size between two groups in each study are equal
# Simulate the effect size using t distribution
# Vary the sample size within each study
N1<-Nvary*(1/3)
N2<-Nvary*(2/3)
d0 <-rt(I,Nvary-2)*sqrt(1/N1+1/N2)
# Vary the population effect size of each study to meet the
random-effect model assumption
PESVARY<-rnorm(I,PES,0.1)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PESVARY
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-Nvary/(N1*N2)+ (ES*ES*0.5)/Nvary
Varianceg <-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
SumWd<-sum(Weight*ES)
SumWdsquare<-sum(Weight*ES*ES)
SumWsquare<-sum(Weight*Weight)
Qstat<- SumWdsquare-(SumWd*SumWd)/SumWeight
Cstat<-SumWeight-(SumWsquare/SumWeight)
df<- I -1
#Use if function to define Tsquare
if(Qstat-df>0){Tsquare<-(Qstat-df)/Cstat} else {Tsquare<-0}
if(Qstat-df>0){Tsquare.array[i]<-(Qstat-df)/Cstat} else
{Tsquare.array[i]<-0}
BetweenStudyVariance<-rep(Tsquare,I)
VarianceTotal<- BetweenStudyVariance+Varianceg
WeightRandom<- 1/VarianceTotal
SumWeightRandom<-sum(WeightRandom)
SumWeightRandomd<-sum(WeightRandom*ES)
WeightdRandom<- SumWeightRandomd/SumWeightRandom
SEMRandom<-sqrt(1/SumWeightRandom)
ZstatRandom<- WeightdRandom/SEMRandom
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandom))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)
Tsquare_ave[j,k]<- mean(Tsquare.array)
}
}
```

```
   out <- list(prob,Tsquare_ave)
names(out) <- c("Real Type I error rate & Power","average T
square")
out
}
Random_HP<- Random_1(possible.ns,I.ns,PES,sims,alpha)
Random_HP[[1]]


# Hunter & Schimt Method

set.seed (1000)
############################################################
Random_2<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)

#set the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
p.value<-as.numeric(rep(NA,sims))

#looping for the average sample size
for (j in 1:n){
N <- possible.ns[j]
#looping for the number of studies
for (k in 1:s){
I<- I.ns[k]
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Sample size across studies equal in this condition
Nvary<-rep(N,I)
# Simulate the effect size using t distribution
# Vary the sample size within each study
N1<-Nvary*(1/3)
N2<-Nvary*(2/3)
d0 <-rt(I,Nvary-2)*sqrt(1/N1+1/N2)
# Vary the population effect size of each study to meet the
random-effect model assumption
PESVARY<-rnorm(I,PES,0.1)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PESVARY
```

```
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
WeightRandomHS<- Nvary
SumWeightRandomHS<-sum(WeightRandomHS)
SumWeightRandomdHS<-sum(WeightRandomHS*ES)
WeightdRandomHS<- SumWeightRandomdHS/SumWeightRandomHS
WeightdRandomHSmatrix<-rep(WeightdRandomHS,I)
NomiVariance<-sum((WeightRandomHS)*(ES-
WeightdRandomHSmatrix)* (ES- WeightdRandomHSmatrix))
DenomiVariance<-SumWeightRandomHS
VarianceHS<-(NomiVariance/DenomiVariance)/I
SEMRandomHS<-sqrt(VarianceHS)
ZstatRandomHS<- WeightdRandomHS/SEMRandomHS
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandomHS))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)

}
}
   out <- list(prob)
names(out) <- c("Real Type I error rate & Power")
out
}

Random_HS<- Random_2(possible.ns,I.ns,PES,sims,alpha)
Random_HS[[1]]


#Power function

AveTsquare<-Random_HP[[2]]

RandomPowfunction<-function(possible.ns, I.ns, PES)
    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
   power <- array(rep(NA,n*s),dim=c(n,s))
   #looping for the average sample size
   for (j in 1:n){
   N <- possible.ns[j]
```

```
   #looping for different number of studies
   for (k in 1:s){
   I<- I.ns[k]
   N1<-N*(1/3)
   N2<-N*(2/3)
   Variancewithin<-N/(N1*N2)+ (PES*PES*0.5)/N
   Tsquare<- AveTsquare[j,k]
   Vtotal<-Variancewithin+Tsquare
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
}
}
   return(powerround)
    }
RandomPowerFunction<-
RandomPowfunction(possible.ns,I.ns,PES)
RandomPowerFunction
```

Simulated and Analytical Power – Unequal Sample Size and Unbalanced Design

Fixed-effects Model

```
set.seed (1000)
############################################################
main<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)

# set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
pvalue.array<-array(0,dim=c(sims,n,s))
p.value<-as.numeric(rep(NA,sims))

#looping for the average sample size
for (j in 1:n){
N<- possible.ns[j]
#looping for the number of studies
for (k in 1:s){
I<- I.ns[k]
#simulation loop
```

```
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Use the truncated binomial distribution to simulate
sample size
try<-function(p,m,c){
   (p/(1-(1-p)^c) - m/c )^2}
#Can vary the maximum value (standard deviation varies)
MaxN<-N*3
#Get the p.value (0 point)
p.to.use<-
optimize(try,interval=c(0.0001,0.9999),m=N,c=MaxN)$minimum
Nvary<-rbinom(I, MaxN,p.to.use)
nb<-sum(Nvary ==0)
while (nb>0){
Nvary [Nvary ==0]<-rbinom(nb,maxss,p.to.use)
nb<-sum(Nvary ==0)}
# Simulate the effect size using t distribution
# Vary the sample size within each study
N1<-Nvary*(1/3)
N2<-Nvary*(2/3)
d0 <-rt(I,Nvary-2)*sqrt(1/N1+1/N2)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PES
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-Nvary/(N1*N2)+ (ES*ES*0.5)/Nvary
Varianceg <-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
SumWd<-sum(Weight* ES)
SumWdsquare<-sum(Weight*ES*ES)
SumWsquare<-sum(Weight*Weight)
WeightedD<- SumWd/SumWeight
SEM<-sqrt(1/SumWeight)
Zstat<- WeightedD/SEM
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(Zstat))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)
}
}
out <- list(prob)
```

```
names(out) <- c("Real Type I error rate & Power")
out
}
FIX<-main(possible.ns,I.ns,PES,sims,alpha)
FIX[[1]]

# Power function

FixPowfunction<-function(possible.ns, I.ns, PES)
    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
   power <- array(rep(NA,n*s),dim=c(n,s))
   #looping for the average sample size
   for (j in 1:n){
   N <- possible.ns[j]
   #looping for the number of studies
   for (k in 1:s){
   I<- I.ns[k]
   N1<-N*(1/3)
   N2<-N*(2/3)
   Vtotal<-N/(N1*N2)+ (PES*PES*0.5)/N
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
}
}
   return(powerround)
    }
FixPowerFunction<-FixPowfunction(possible.ns,I.ns, PES)
FixPowerFunction
```

Random-effects Model

```
#Hedges & Colleagues Method

set.seed (1000)
############################################################
Random_1<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
```

```
s <- length(I.ns)

#set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
Tsquare_ave<- array(rep(NA,n*s),dim=c(n,s))
Tsquare.array<-as.numeric(rep(NA,sims))
p.value<-as.numeric(rep(NA,sims))

#looping for the average sample size
for (j in 1:n){
N <- possible.ns[j]
#looping for the number of studies
for (k in 1:s){
I<- I.ns[k]
#simulation loop
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Use the truncated binomial distribution to simulate
sample size
try<-function(p,m,c){
   (p/(1-(1-p)^c) - m/c )^2}
#Can vary the maximum value (standard deviation varies)
MaxN<-N*3
#Get the p.value (0 point)
p.to.use<-
optimize(try,interval=c(0.0001,0.9999),m=N,c=MaxN)$minimum
Nvary<-rbinom(I, MaxN,p.to.use)
nb<-sum(Nvary ==0)
while (nb>0){
Nvary [Nvary ==0]<-rbinom(nb,maxss,p.to.use)
nb<-sum(Nvary ==0)}
# Simulate the effect size using t distribution
# Vary the sample size within each study
N1<-Nvary*(1/3)
N2<-Nvary*(2/3)
d0 <-rt(I,Nvary-2)*sqrt(1/N1+1/N2)
PESVARY<-rnorm(I,PES,0.1)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PESVARY
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
Variancewithin<-(4/Nvary)*(1+0.125*ES*ES)
Varianceg <-J*J*Variancewithin
Weight<-1/Varianceg
SumWeight<-sum(Weight)
```

```
SumWd<-sum(Weight*ES)
SumWdsquare<-sum(Weight*ES*ES)
SumWsquare<-sum(Weight*Weight)
Qstat<- SumWdsquare-(SumWd*SumWd)/SumWeight
Cstat<-SumWeight-(SumWsquare/SumWeight)
df<- I -1
#Use if function to define Tsquare
if(Qstat-df>0){Tsquare<-(Qstat-df)/Cstat} else {Tsquare<-0}
if(Qstat-df>0){Tsquare.array[i]<-(Qstat-df)/Cstat} else
{Tsquare.array[i]<-0}
BetweenStudyVariance<-rep(Tsquare,I)
VarianceTotal<- BetweenStudyVariance+ Varianceg
WeightRandom<- 1/VarianceTotal
SumWeightRandom<-sum(WeightRandom)
SumWeightRandomd<-sum(WeightRandom*ES)
WeightdRandom<- SumWeightRandomd/SumWeightRandom
SEMRandom<-sqrt(1/SumWeightRandom)
ZstatRandom<- WeightdRandom/SEMRandom
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandom))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)
Tsquare_ave[j,k]<- mean(Tsquare.array)
}
}
   out <- list(prob,Tsquare_ave)
names(out) <- c("Real Type I error rate & Power","average T
square")
out
}
Random_HP<- Random_1(possible.ns,I.ns,PES,sims,alpha)
Random_HP[[1]]

#Hunter&Schmit Method
set.seed (1000)
############################################################
Random_2<-function(possible.ns, I.ns, PES, sims, alpha)
{
# number of sample size vector
n <- length(possible.ns)
# number of studies vector
s <- length(I.ns)
```

```
#set up the output
prob <- array(rep(NA,n*s),dim=c(n,s))
significant.experiments <- rep(NA, sims)
p.value<-as.numeric(rep(NA,sims))

#looping for the average sample size
for (j in 1:n){
N <- possible.ns[j]
#looping for the number of studies
for (k in 1:s){
I<- I.ns[k]
for (i in 1:sims){
# In each simulation, perform the meta-analysis
# Use the truncated binomial distribution to simulate
sample size
try<-function(p,m,c){
   (p/(1-(1-p)^c) - m/c )^2}
#Can vary the maximum value (standard deviation varies)
MaxN<-N*3
#Get the p.value (0 point)
p.to.use<-
optimize(try,interval=c(0.0001,0.9999),m=N,c=MaxN)$minimum
Nvary<-rbinom(I,MaxN,p.to.use)
nb<-sum(Nvary ==0)
while (nb>0){
Nvary [Nvary ==0]<-rbinom(nb,maxss,p.to.use)
nb<-sum(Nvary ==0)}
# Simulate the effect size using t distribution
# Vary the sample size within each study
N1<-Nvary*(1/3)
N2<-Nvary*(2/3)
d0 <-rt(I,Nvary-2)*sqrt(1/N1+1/N2)
PESVARY<-rnorm(I,PES,0.1)
J<-1-(3/(4*(Nvary-2)-1))
g<- d0*J
ES<- g + PESVARY
#Calculate the Z-test statistics - get combined effect size
and variance of all studies
WeightRandomHS<- Nvary
SumWeightRandomHS<-sum(WeightRandomHS)
SumWeightRandomdHS<-sum(WeightRandomHS*ES)
WeightdRandomHS<- SumWeightRandomdHS/SumWeightRandomHS
WeightdRandomHSmatrix<-rep(WeightdRandomHS,I)
NomiVariance<-sum((WeightRandomHS)*(ES-
WeightdRandomHSmatrix)* (ES- WeightdRandomHSmatrix))
DenomiVariance<-SumWeightRandomHS
VarianceHS<-(NomiVariance/DenomiVariance)/I
```

```
SEMRandomHS<-sqrt(VarianceHS)
ZstatRandomHS<- WeightdRandomHS/SEMRandomHS
#Return the p values of all simulations
#Return the significant test result (retain/reject the null
hypothesis)
p.value[i]<- 2*pnorm(-abs(ZstatRandomHS))
significant.experiments[i] <- ifelse(p.value[i] <=
alpha,1,0)
}
prob[j,k] <- mean(significant.experiments)

}
}
   out <- list(prob)
names(out) <- c("Real Type I error rate & Power")
out
}

Random_HS<- Random_2(possible.ns,I.ns,PES,sims,alpha)
Random_HS[[1]]


#Power function

AveTsquare<-Random_HP[[2]]
RandomPowfunction<-function(possible.ns, I.ns, PES)
    {
    # number of sample size vector
   n <- length(possible.ns)
   # number of studies vector
   s <- length(I.ns)
   power <- array(rep(NA,n*s),dim=c(n,s))
   #looping for the average sample size
   for (j in 1:n){
   N <- possible.ns[j]
   #looping for the number of studies
   for (k in 1:s){
   I<- I.ns[k]
   N1<-N*(1/3)
   N2<-N*(2/3)
   Variancewithin<-N/(N1*N2)+(PES*PES*0.5)/N
   Tsquare<- AveTsquare[j,k]
   Vtotal<-Variancewithin+Tsquare
   lamda<-sqrt(I)*PES/sqrt(Vtotal)
   power[j,k]<-pnorm(lamda-qnorm(1-
0.05/2))+pnorm(qnorm(0.05/2)-lamda)
   powerround<-round(power, digits=4)
```

```
}
}
   return(powerround)
    }
RandomPowerFunction<-
RandomPowfunction(possible.ns,I.ns,PES)
RandomPowerFunction
```

Graph Functions

```
#Graph Function(Power Curve)
   plotfun<-
function(mat,sampnvec,studynvec,lwdo=1,ltyo=1,newo=F){
#Use lwd=2 in the plot statement for wider lines (or other
numbers)
#Use lty=2 in the plot statement for dashed lines (or other
numbers)
#mat is the output called FIX/Random_HP
#sampnvec is the possible.ns - assumes these are sorted
correctly already and match rows of FIX/Random_HP
#studynvec is the I.ns - assumes these are sorted correctly
already and match FIX/Random_HP
#rows are average sample size
#columns are number of studies
#with different curves for each number of studies
nsamp<-nrow(mat)
nstudies<-ncol(mat)
colorvec<-rainbow(nstudies)
par(new=newo)
plot(sampnvec,mat[,1],col=colorvec[1],type="l",
       xlim=c(min(sampnvec),max(sampnvec)),ylim=c(0,1),
       xlab="Sample Size",ylab="Power",lwd=lwdo,lty=ltyo,
main="Power by Sample Size and Number of Studies")
for (j in 2:nstudies){
   par(new=T)
   plot(sampnvec,mat[,j],col=colorvec[j],type="l",
       xlim=c(min(sampnvec),max(sampnvec)),ylim=c(0,1),
       xlab="Sample Size",ylab="Power",lwd=lwdo,lty=ltyo,
main="Power by Sample Size and Number of Studies")
}
}
#Perform the plotfun multiple times and use newo=T to add
new curves
#Use symbolvec to define different symbols on the power
curves.
```