The Report committee for Xin Li

Certifies that this is the approved version of the following report:

Detecting and Correcting Publication Bias in Meta-Analysis

APPROVED BY

SUPERVISING COMMITTEE:

Supervisors: _____

S. Natasha Beretvas

_____

Gary Borich

Detecting and Correcting Publication Bias in Meta-Analysis

by

Xin Li, M.S. Stat.

Report

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

December 2009

Acknowledgements

I would like to appreciate all the helps from Natasha Beretvas and Gary Borich. Thank you so much for your helpful suggestions and guidance, and I am so proud that I am one of your students!

I would like to give tons of thanks to my lovely and warm family. 奶奶, 老爸老妈, 二爸，维嬢嬢, and Rebecca, 谢谢你们! I love you all!

I would also like to thank my friends for their help and support, especially 婷婷wer and 婧J.

Detecting and Correcting Publication Bias in Meta-Analysis

by

Xin Li, M.A.

The University of Texas at Austin, 2009

SUPERVISORS: Natasha Beretvas, Gary Borich

Publication bias (PB) makes the resources for meta-analysis (M-A) unreliable in the sense of completion and accuracy, so to investigate, identify and correct PB is a very important issue in M-A. The current study proposed an empirical comparison in both detection and correcting PB, using a Monte Carlo study. Conditions to be manipulated include the number of primary studies, number of missing studies and true effect size. RANNOR in SAS will be used to generate normally distributed random variables and, for each condition, 10,000 M-As will be simulated. Type I error rates are to be calculated for the conditions with no PB and powers were estimated for the conditions with PB and adequate type I error control. Finally, a demonstration of how M-A can and should be used as a part of program evaluations was given.

TABLE OF CONTENTS

LIST OF FIGURES

Introduction

Meta-analysis is most commonly criticized for the impact that publication bias can have on the validity of its results. The primary source of publication bias is that the published results are more likely to be the statistically significant results than the non-significant results. Therefore, the non-significant results cannot be included in a meta-analytic summary.

When doing meta-analysis, there is an important assumption that the results being meta-analyzed should come from a random sample of independent studies. Publication bias introduces statistical bias into meta-analytic results because the existence of publication bias violates the above assumption. Due to publication bias, only significant effects will be accessible to the meta-analyst, not the full spectrum of the sampled values, and so the meta0analytic average of these effects will be biased in the same direction as the bias.

There are many methods used for assessing publication bias. The funnel plot (Light & Pillemer, 1984) is a simple scatter plot of each study's effect size along with some function of the study's sample size. Fail-safe number for p-values (Rosenthal, 1979) represents the number of studies with an average null effect required to change a meta-analytic result from statistical significance to non-significance. If it is not large, then this suggests the existence of publication bias. The rank correlation method (Begg, 1994) computes the rank correlation between each study's effect size and a function of the study's sample size and uses it to test for publication bias. Egger's regression method (Egger et al., 1997) regresses the standardized normal

deviate of the effect size estimate on the precision of the corresponding effect size estimate, and uses the hypothesis tests with a null hypothesis that the regression line goes through the origin to assess publication bias. Funnel plot regression method (Macaskill, Walter & Irwig, 2001) regresses the study size on the effect size estimate, and uses the result that the regression slope is zero to support a lack of publication bias. The Trim and Fill method (Duval & Tweedie's, 2000) is a more recent, non-parametric method used to assess the presence of publication bias. It is based on the inferences associated with a funnel plot, and provides the number of studies' effect size estimates required to make the funnel plot more symmetric.

Despite this plethora of available methods, many meta-analysts are not assessing publication bias. The current study involves summarizing methods used to assess publication bias in applied meta-analyses, and then simulating data to demonstrate inferential differences across those methods for publication bias assessment.

Integrative Analysis and Interpretation

Meta-Analysis

Since its first introduction in the social sciences by Glass (1976), meta-analysis has been increasingly used in many fields of research. Meta-analysis is a statistical technique that involves aggregation of the quantitative results of empirical studies. After a meta-analyst formulates her research topic, she needs to identify and retrieve the relevant research studies from the appropriate population (Lipsey & Wilson, 2001). Next, the meta-analyst should choose an appropriate effect size statistic for her research question and then calculate the relevant effect size of interest for each study. The set of studies' effect size estimates are then pooled together to obtain an overall average effect size estimate. To make the estimates of the overall effect size more accurate, weighting techniques are often used. Hedges and Olkin (1985) provided a straightforward approach, and the formula is shown below:

$$\bar{T}_{\cdot} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} \qquad (1)$$

where k is the number of studies, $\bar{T}_{\cdot}$ is the weighted average effect size estimate across the k studies, $\hat{\theta}_i$ is the effect size estimate for study i, and w_i is the weight associated with $\hat{\theta}_i$. A commonly used weight, w_i, is the inverse of the variance of the effect size estimate, $\hat{\theta}_i$, i.e. $\frac{1}{\hat{\sigma}_i^2}$. It should be noted that there are other weights that

researchers use, however, that recommended by Hedges and Olkin will be used in the current study.

Following calculation of the pooled effect size estimate, further analyses can be conducted by the meta-analyst. The meta-analyst can calculate the associated standard error of $\hat{\theta}_i$, and then conduct a test of the statistical significance of the effect size or construct a confidence interval around the estimate. Although not of interest in the current study, meta-analysts also can explore variability in studies' effects by conducting weighted regression analyses that include study and sample descriptors as predictors of effect sizes.

Because the key point in meta-analysis is the effect size statistic, the most commonly used meta-analytic effect size statistics are introduced here, which include $\hat{\delta}$ and $\hat{\rho}$. The former effect size, $\hat{\delta}$, estimates the standardized difference between two true population means. The formula for calculating the sample estimate, $\hat{\delta}$ is:

$$\hat{\delta} = \frac{\bar{X}_1 - \bar{X}_2}{S_p} \qquad (2)$$

where $\bar{X}_i$ represents the sample mean for group *i*, and $S_p = \sqrt{\frac{(df_1)(s_1^2) + (df_2)(s_2^2)}{df_1 + df_2}}$ is the square root of the average of the two samples' variances (s_1^2 and s_2^2) weighted by their associated degrees of freedom (*df*). The variance estimate of $\hat{\delta}$ can be calculated using the following equation:

$$\hat{\sigma}^2_{\hat{\delta}} = \left[\frac{n_1 + n_2}{n_1 n_2}\right] + \left[\frac{\hat{\delta}^2}{2(n_1 + n_2)}\right] \quad (3)$$

where n_i represents the sample size for group i.

The effect size statistic $\hat{\rho}$ provides an estimate of the population correlation, ρ. The formula for calculating $\hat{\rho}$ is:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (4)$$

where, here, n is the number of (X, Y) pairs in the sample data. The sample variance of $\hat{\rho}$ is:

$$\hat{\sigma}^2_{\hat{\rho}} = \frac{(1 - \hat{\rho}^2)^2}{(n_i - 1)} \quad (5)$$

However, there are some notable disadvantages of $\hat{\rho}$. First, because the correlation coefficient is only defined in the interval [-1, 1], the distribution of $\hat{\rho}$ is not symmetric for true values different than zero. Second, the variance of $\hat{\rho}$ depends on the value of ρ (see Equation 5). Therefore, many researchers use Fisher's (1928) normalizing and variance-stabilizing $\hat{\rho}$-to-$Z_{\hat{\rho}}$ transformation rather than using the ρ metric. The $\hat{\rho}$-to-$Z_{\hat{\rho}}$ transformation is as follows:

$$Z_{\hat{\rho}} = (0.5)\left[\ln\left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}}\right)\right]. \quad (6)$$

The sampling distribution of $Z_{\hat{\rho}}$ can be assumed asymptotically normal when the sample size is large, and its variance, $\hat{\sigma}^2_{Z_{\hat{\rho}}}$, depends only on sample size:

$$\hat{\sigma}^2_{Z_{\hat{\rho}}} = \frac{1}{n-3}. \quad (7)$$

Once an analysis using the effect size statistic $Z_{\hat{\rho}}$ has been conducted, the results can then be back-transformed to the ρ metric by using the formula shown below:

$$\hat{\rho} = \frac{e^{2Z_{\hat{\rho}}} - 1}{e^{2Z_{\hat{\rho}}} + 1}. \quad (8)$$

Clearly, computation of each study's effect size and its associated variance is relatively easy given the appropriate data are available. In addition, pooling the estimates together to obtain a single weighted average, $\overline{T}.$, is also quite simple (see Equation 1). Unfortunately, the problem of publication bias hampers the possible validity of the estimate obtained using only published data. The next sections describe publication bias in more detail including methods used to detect it.

Publication Bias

One of the primary advantages of meta-analysis is that it supplies researchers with a rigorous way to handle and summarize the information provided in a large number of studies. However, one of the major criticisms of meta-analysis is the impact that could result from selective publication bias on the accuracy of the results produced by a meta-analysis. According to Begg (1994), there are two categories of publication bias: subjective publication bias and selective (objective) publication bias.

Subjective publication bias refers to the situation in which a published article does not accurately represent the conclusions that are reached in the research process. This can occur when statistical analyses are subjectively influenced by the intent to present only data that supports the preferred results. In other words, selective publication bias, which is the primary source of publication bias in meta-analysis, occurs when the decision to publish, made by an author or journal editor is influenced by the results of study. This means that, in general, studies with statistically significant results (in the hypothesized direction) are more likely to get published. In addition, selective publication bias might also be called objective publication bias because it is "the objective data reported that are subject to the bias." (Begg, 1994, p. 400) Because meta-analysts more frequently encounter and thus mainly focus on selective publication bias, the term "publication bias" will be used here to refer to "selective publication bias".

Generally speaking, articles that present statistically significant results have a higher probability of getting published (Begg, 1994). In addition, results associated with larger sample sizes will have more power and thus articles involving studies that have larger sample sizes are more likely to get published (Begg, 1994). This means that the magnitude of the publication bias is positively related to the inverse of the sample size of a study (Begg, 1994). Meta-analysis focuses on the combination and comparison of results. Thus if only results from published studies are analyzed, then this will result in a dataset that could lead the meta-analytic results to be biased in the direction of the kind of results that are published – namely, those with larger effect

sizes. Therefore, in order to have accurate and valid results, publication bias needs to be considered carefully in meta-analysis.

Consideration of publication bias has two forms. First, methods for *detecting* publication bias should be used. Second, methods for *correcting* possibly identified bias should also be used if evidence indicates the presence of publication bias. The current study is designed to focus on the detection of publication bias. The next section will describe some of the procedures currently used to detect publication bias.

Publication Bias Detection

In general, there are a lot of methods that are commonly used to detect publication bias (Begg, 1994). Here, the focus will be on statistical significance tests used to identify publication bias. However, first a commonly used graphical method for detecting publication bias, called the funnel plot approach, will be introduced because it provides the basis for two of the statistical tests that will be mentioned later.

Funnel Plot Approach

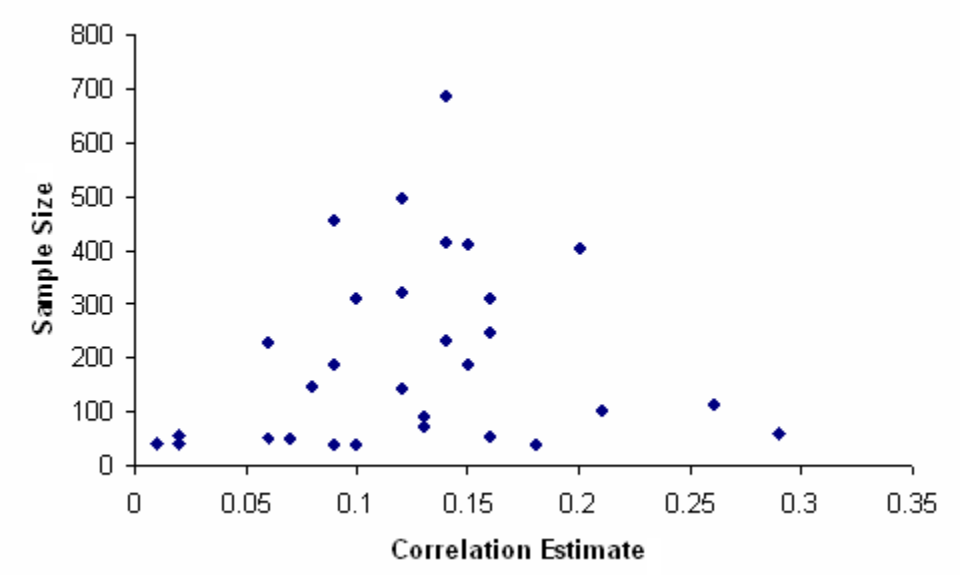
The funnel plot (Light and Pillemer, 1984) is a relatively simple method for detecting possible publication bias. It is a graphical approach involving inspection of a plot of the effect size estimates, $\hat{\theta}$, along the *X* axis against some measure of the precision of $\hat{\theta}$ (typically either the sample size on which $\hat{\theta}$ is based, or the inverse of the standard error of $\hat{\theta}$). The shape of the resulting funnel plot can be used to visually assess the potential for publication bias. An example of two funnel plots for $\hat{\rho}$ is

given below (see Figures 1 and 2). The depicted data were taken from Heller, Ilies and Watson's meta-analysis (2004). The correlation of interest represents the relationship between job satisfaction and marital satisfaction. It should be noted that two outlying points were deleted from the analysis to facilitate demonstration of the funnel plot.

Based on the first funnel plot, a meta-analyst would conclude that there is no strong evidence for publication bias. The reason for this is that the data points form a plot that is symmetric about a vertical line where $\hat{\rho} \approx 0.15$. Less variation is evident in the estimates associated with larger sample sizes (and thus smaller variances) at the top of the plot and more variability can be noted in the estimates associated with smaller sample sizes at the base of the plot (see Figure 1).

Figure 1.

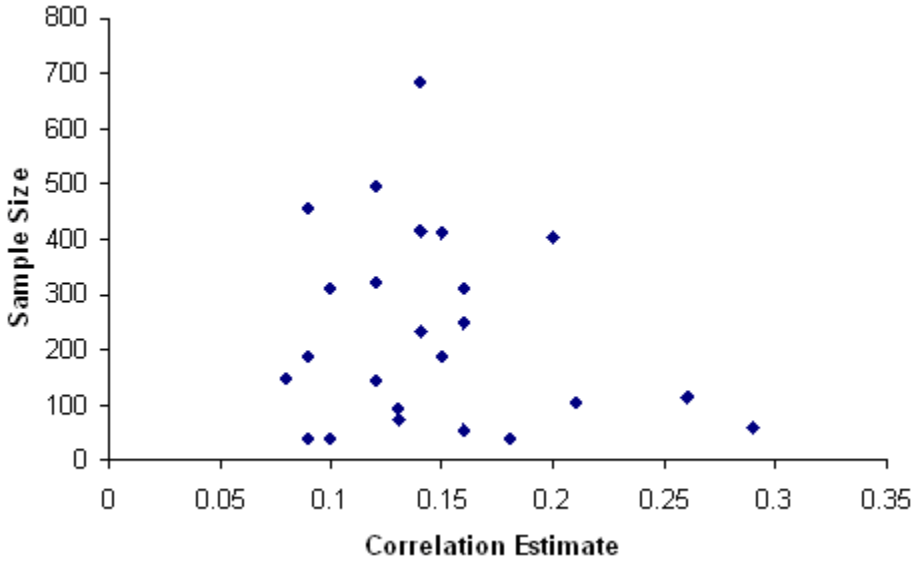
Funnel Plot of Correlation Coefficients between Job Satisfaction and Marital Satisfaction from Heller, Ilies and Watson (2004)



The second funnel plot (see Figure 2) was constructed by mimicking publication bias in that the smallest (six) $\hat{\rho}$ s were deleted. Visual inspection of the second funnel plot would lead to identification of a truncated (on the left-hand side) funnel plot and the meta-analyst would infer the possibility of publication bias.

Figure 2.

Funnel Plot of Correlation Coefficients between Job Satisfaction and Marital Satisfaction from Heller, Ilies and Watson (2004), with the Smallest Six Estimates Deleted



However, interpretation of a funnel plot is purely subjective, and it can only give a suggestion of possible publication bias, but cannot give any further information. For example, it does not indicate how many missing studies there might be nor how different the effect size estimate would be if there were no missing studies. In addition, when an asymmetric funnel plot is identified, this cannot be used as evidence solely of publication bias because it may be caused by other forms of selection bias (e.g., true heterogeneity, data irregularities, and even chance) (Egger et al, 1997). Despite the subjectivity associated with interpretation of the funnel plot's results, several methods for assessing possible publication bias, (including Egger's

regression and the Trim and Fill procedures) are based on the funnel plot. There are also several other alternative methods used to assess publication bias that are not based on the funnel plot. The indices and statistics that are commonly used to assess publication bias are presented next.

Fail-Safe Number

The fail-safe number (FSN) was derived by Rosenthal (1979). Its goal is to estimate how many additional studies, whose average observed effect is zero, are required to reverse the statistical significance of the overall effect size estimate. The number of required additional studies is denoted as k_0. If the estimated value of k_0 is too large to be reasonable based on the research domain, then the researcher can conclude that there is strong evidence against publication bias.

Suppose a dataset is formed with results from *k* (published) studies, and the pooled effect size estimate, $\overline{T}_{\cdot}$, (Equation 1) calculated using this dataset is significantly greater than zero. For each effect size estimate, its *Z*-score can be calculated by using the estimated value of the effect size statistic, $\hat{\theta}$, divided by its corresponding standard error ($s_{\hat{\theta}}$). Denoting the *Z*-score for the *i*th study by Z_i (i = 1, 2… *k*), we then have:

$$Z_i = \frac{\hat{\theta}_i}{S_{\hat{\theta}_i}} \qquad (9)$$

and the overall *Z*-score for the studies' effect size estimates is:

$$Z = \sum_{i=1}^{k} \frac{Z_i}{\sqrt{k}} \qquad (10)$$

The result is significantly greater than zero if and only if *Z* is larger than the critical *Z*-score (Z_c) for example, if $Z > Z_c$ for a one-tailed hypothesis test. Then, by the definition of k_0 as described above, the following formula can be used to derive k_0 (the number of unpublished studies' effect size estimates with an average effect size estimate of zero):

$$\sum_{i=1}^{k} \frac{Z_i}{\sqrt{k + k_0}} < Z_c \qquad (11)$$

Multiplying both sides of Equation 11 by the positive quantity $\frac{\sqrt{k + k_0}}{Z_c}$ and subtracting *k* from both sides, the formula for k_0 is easily obtained:

$$k_0 > -k + \left(\frac{\sum_{i=1}^{k} Z_i}{Z_c} \right)^2 \qquad (12)$$

The main advantage of using the FSN method is that it provides a very simple way to estimate the number of missing studies. However, there is no model underlying the FSN (Orwin, 1983). In addition, Rosenthal did not give any statistical criteria to decide whether k_0 is large enough to say there is evidence of publication bias. Instead, Rosenthal suggested a rule of thumb such that there is evidence of publication bias if k_0 is greater than 5*k*+10. Several researchers have also argued that

the assumption that, on average, the missing studies have a null effect was not reasonable (Becker, 2005). This method also does not directly consider sample size information (Sutton et al., 2000), and it does not permit modeling of heterogeneity (Iyengar and Greenhouse, 1988). To overcome some of those weakness mentioned above, many variations on the FSN have been developed, including Orwin's (1983) N_{es} and an FSN based on Fisher's (1932) work. Detailed information regarding these variations is available elsewhere (e.g., see Becker, 2005).

Another criticism of the Rosenthal's FSN is that it is not associated with a statistical test. The next few sections describe the methods for assessing publication bias that use statistical tests.

Begg's Rank Correlation Method

Begg derived a statistic that describes the strength of the rank correlation between the standardized effect size estimates, $\hat{\theta}_i^B$, and the variance of the quantity $(\hat{\theta}_i - \bar{T}_\cdot)$, $\sigma_i^{2^*}$. Begg also provided a statistic describing the rank correlation between $\hat{\theta}_i^B$ and its corresponding sample size n_i. In either case, this statistic is used to test the null hypothesis that there is no publication bias. The test statistic that is based on the rank correlation between $\hat{\theta}_i^B$ and $\sigma_i^{2^*}$ will be denoted by r_{tv}. The statistic that is based on the rank correlation between $\hat{\theta}_i^B$ and n_i is denoted by r_{tn}. The standardized effect size estimate is computed using the following equation.

$$\hat{\theta}_i^B = \frac{\hat{\theta}_i - \overline{T}_.}{\sqrt{\sigma_i^{2^*}}} \quad (13)$$

where $\overline{T}_.$ is the weighted average effect size estimate across the *k* studies (see Equation 1) being meta-analyzed and $\sigma_i^{2^*}$ is

$$\sigma_i^{2^*} = \sigma_i^2 - \frac{1}{\sum_{j=1}^{k} (\sigma_j^2)^{-1}} \quad (14)$$

Ranks are assigned to every $\hat{\theta}_i^B$ and $\sigma_i^{2^*}$ (or n_i) with 1 being assigned to the largest value and 2 being assigned to the second largest value, etc. Then, for *k* studies, consider all *k* (*k*-1)/2 pairs of ($\hat{\theta}_i^B, \sigma_i^{2^*}$) and ($\hat{\theta}_j^B, \sigma_j^{2^*}$) for studies *i* and *j* to obtain the number of concordant and discordant pairs. Here, if the pair, ($\hat{\theta}_i^B, \sigma_i^{2^*}$) and ($\hat{\theta}_j^B, \sigma_j^{2^*}$), is concordant for studies *i* and *j*, then the ranks of $\hat{\theta}_i^B$ and $\sigma_i^{2^*}$ are both higher or both lower than the corresponding ranks of $\hat{\theta}_j^B$ and $\sigma_j^{2^*}$. Otherwise, this pair is said to be discordant. Let *P* denote the number of concordant pairs and *Q* denote the number of discordant pairs. For each pair of $\hat{\theta}_i^B$ and $\sigma_i^{2^*}$, ranks are assigned with ties being allowed. Based on the rank values of $\sigma_i^{2^*}$, from the smallest to the largest, each rank value of $\hat{\theta}_i^B$ is compared to the subsequent rank values of $\hat{\theta}_i^B$, and the number of subsequent rank values that are larger than the original rank value

of $\hat{\theta}_i^B$ is counted, excluding the tied ranks and denoted by p_i. In addition, q_i (which is the number of subsequent rank values that are smaller than the original rank value of $\hat{\theta}_i^B$) is counted. Using the sums of p_i and q_i, (where $P = \sum_{i=1}^{k} p_i$ and $Q = \sum_{i=1}^{k} q_i$), a normalized z score is computed as following.

$$Z_{BV} = \frac{(P - Q)}{\sqrt{k(k-1)(2k+5)/18}} \qquad (15)$$

Z_{BV} can then be tested to assess whether there is potential publication bias.

The preceding description and Equation 15 describe the procedure for using Kendall's tau as Begg's rank correlation. However, Begg's rank correlation test, either for r_{tv} or r_{tn}, can also be based on Spearman's rho (Begg, 1994) , which involves easier computation (Colton, 1974). In the current paper, use of Kendall's tau will be assessed, because it is the version that is most commonly used.

It is assumed that a study with a small sample size and small effect size estimate is unlikely to get published. Thus, when publication bias exists, there will be none necessary correlation between the obtained effect size estimates, $\hat{\theta}_i$, and its corresponding sample size, n_i, or with its estimated sample variance, $\hat{\sigma}_i^2$. Thus, if a strong correlation is found (using Begg's rank correlation test) then this provides evidence supporting the possibility of publication bias.

Kromrey and Rendina-Gobioff (2006) assessed the Type I error rate of both r_{tn} and r_{tv}. The authors found that the Type I error control was better for larger true

effect sizes and for scenarios with a smaller numbers of studies. Type I error control was closer to the nominal alpha-level when using r_{tn} than r_{tv} and the control was best for smaller sample sizes. Begg (1994) claimed that use of this rank correlation test has been found to be low in statistical power which was supported in studies conducted by Sterne, Gavaghan and Egger (2000), Macaskill, Walter and Irwig (2001) and Kromrey and Rendina-Gobioff (2006). Macaskill, Walter and Irwig's (2001) and Kromrey and Rendina-Gobioff's (2006) studies also showed that the Begg's rank correlation test based on Kendall's tau using r_{tv}, has more statistical power than the test based on n_i, (i.e. using r_{tn}). The maximum power increased as the number of studies and sample size per study increased, and as the magnitude of publication bias increased. It should be noted that when r_{tn} was used, power was also found to increase for larger true effect sizes, whereas when r_{tv} was used, power increased for smaller true effect sizes (Kromrey and Rendina-Gobioff, 2006).

Begg's rank correlation test for assessing publication bias is not directly based on an examination of the asymmetry of a funnel plot, however, there are several other statistical tests of publication bias that are based on funnel plot asymmetry. These two methods include Egger's regression and the funnel plot regression methods that are described below.

Egger's Regression with OLS Estimation

Egger developed a publication bias assessment method that involves regressing the standardized normal deviate, $\hat{\theta}_i^E$, of the effect size estimate on the precision, $p(\hat{\theta}_i)$, of the corresponding effect size estimate, $\hat{\theta}_i$, where

$$p(\hat{\theta}_i) = \frac{1}{\sqrt{\sigma_i^2}} \tag{16}$$

and

$$\hat{\theta}_i^E = \hat{\theta}_i \; p(\hat{\theta}_i) \tag{17}$$

Egger's regression equation that is estimated to assess publication bias is then

$$\hat{\theta}_i^E = \beta_0 + \beta_1 \; \mathrm{p}(\hat{\theta}_i) \tag{18}$$

Egger's regression equation can be estimated using ordinal least squares (OLS), or weighted least squares (WLS). However, because WLS estimates have been shown to be biased (Kromrey and Rendina-Gobioff, 2006), the WLS method will not be discussed here. The test statistic used in Egger's regression method with OLS estimation is denoted by $\hat{\beta}_0$. The equation

$$\beta_0 = p(\hat{\theta}_i)(\hat{\theta}_i - \beta_1) \tag{19}$$

can be obtained by first subtracting $\beta_1 p(\hat{\theta}_i)$ from both sides of Equation 18 to obtain: $\beta_0 = \hat{\theta}_i^E - \beta_1 p(\hat{\theta}_i)$, and then substituting $\hat{\theta}_i^E$ from Equation 17. Based on Equation 19, $\hat{\beta}_0 = p(\hat{\bar{T}}_\cdot)(\hat{\bar{T}}_\cdot - \hat{\beta}_1)$ using the pooled overall effect size estimate $\hat{\bar{T}}_\cdot$, where $\hat{\bar{T}}_\cdot$ can be interpreted as the gradient of this OLS regression line (Sutton, et al.,

2002). If the data in the funnel plot were symmetric (indicating a lack of publication bias), then $\hat{\bar{T}}_{\cdot} = \hat{\beta}_1$ will result in $\hat{\beta}_0$ equaling zero (see Equation 19), that is the regression line would then go through the origin. Thus, Egger's hypothesis tests (with a null hypothesis that $\hat{\beta}_0 = 0$) can be used to assess publication bias. If the null hypothesis is rejected, evidence has been found supporting publication bias.

Macaskill, Walter and Irwig (2001) conducted a simulation study and found that even though all the publication bias identification methods that were assessed in their study showed low statistical power, Egger's test using $\hat{\beta}_0$ had higher power than both Begg's rank correlation test and the funnel plot regression method (which will be described next). However, in their simulation study, Kromrey and Rendina-Gobioff (2006) found that Egger's regression method did not provide higher power than Begg's rank correlation test, and its maximum power increased as the number of studies, sample size and magnitude of publication bias increased, and as the true effect size decreased. In terms of Type I error control, both simulation studies found that Egger's test performed poorly. Type I error control was better with larger true effect sizes, sample sizes and smaller numbers of studies (Kromrey & Rendina-Gobioff, 2006). Sterne, Gavaghan and Egger (2000) recommended Egger's regression method over Begg's rank correlation test.

Another statistical test of publication bias that is based on the funnel plot logic is the funnel plot regression method. The difference between Egger's regression method and the funnel plot regression method is that a different test statistic is used to

detect publication bias because of the use of a different predictor in the regression model.

Funnel Plot Regression

Funnel plot regression is a publication bias detection method that was suggested by Macaskill, Walter and Irwig (2001). Funnel plot regression involves use of study size n_i as the predictor with the effect size estimate $\hat{\theta}_i$ as the criterion. Weighted least squares estimation is recommended, and the weight most often used is the inverse of the variance of the effect size estimate. The intercept of the regression line for a symmetric plot of $\hat{\theta}_i$ against n_i directly corresponds with the pooled effect size estimate, that is $\hat{\beta}_0 = \hat{\bar{T}}_.$. Using matrix algebra, it can be shown that a lack of publication bias is supported when the regression slope for the predictor (study sample size) equals zero.

Kromrey and Rendina-Gobioff's simulation study (2006) only considered the case when $\frac{1}{\sigma_i^2}$ is used as the weight in funnel plot regression, and the test statistic used for this method will be denoted by $\hat{\beta}_{1,\sigma^2}$. The authors found that this method showed conservative Type I error control in most of conditions they introduced in their study, and the control improved as the number of studies and the magnitude of the true effect size increased. They also found that the power for detecting publication bias provided by funnel plot regression was low. Last, they found that power was

enhanced for a larger number of studies, larger sample sizes and effect sizes and when a larger magnitude of publication bias had been introduced.

Macaskill, Walter and Irwig's study (2001) compared two versions of the funnel plot regression method using both $\hat{\beta}_{1,\sigma^2}$ (for which the weight is just the inverse of the variance for each estimation, i.e. $\frac{1}{\sigma_i^2}$) and $\hat{\beta}_{1,\sigma_p^2}$ (for which the inverse of the pooled variance for each study is used as the weight). The authors found that the funnel plot regression method using $\hat{\beta}_{1,\sigma_p^2}$ performed well overall in terms of Type I error control, but the method using either $\hat{\beta}_{1,\sigma^2}$ or $\hat{\beta}_{1,\sigma_p^2}$ showed lower power compared to Egger's regression method and Begg's rank correlation method using r_{tv}. They asserted that the funnel plot regression method using $\hat{\beta}_{1,\sigma_p^2}$ should be the preferred approach for identifying publication bias.

One additional publication bias detection method that was investigated by Kromrey and Rendina-Gobioff (2006) but not investigated in Macaskill et al.'s (2001) study is the non-parametric trim and fill (Duval and Tweedie, 2000) method. The next section describes this method.

Trim and Fill Method

The trim and fill method was derived by Duval and Tweedie (2000a, 2000b). They developed three estimators, namely R_0, L_0 and Q_0, to estimate k_0, which represents the number of missing (assumed unpublished) studies. The closer the value

of k_0 is to zero, the fewer studies are inferred to be missing. The trim and fill method involves analyzing the rank of effect size estimates' magnitudes and is thus another non-parametric method.

The symbol, η, will be used, here, to denote the ranks of the absolute values (i.e. magnitudes) of both observed and missing studies' effect size estimates, such that η_i is the rank for $|\hat{\theta}_i|$. Publication bias is defined as occurring in meta-analysis when results from studies are not "published" (i.e., when they are unavailable for analysis). Thus, the ranks for effect size estimates of observed studies are the only available ranks. η_i^* will be used to denote the rank of $|\hat{\theta}_i^*|$, where $\hat{\theta}_i^*$ is the effect size estimate for the *i*th *observed* study. The lower case Greek letter gamma (γ) with a star, γ^*, will be used to denote the length of the rank string that contains only the ranks of the *largest, positive, observed* effect size estimates. (It is assumed here that the sign of the effect size of interest is positive. Obviously, the reverse of this procedure can be used for negative effect sizes). The last descriptor of ranks necessary for the trim and fill method is T_k, the trimmed rank test statistic, where T_k is the sum of the *p* positive ranks of *k* observed effect size estimates (i.e. for which $\hat{\theta}_i^* > 0$), such that:

$$T_k = \sum_{i=1}^{p} \eta_i^* \qquad (20)$$

where *p* is the count of $\hat{\theta}_i^*$ s greater than zero.

The rank descriptors described above (γ^* and T_k) are necessary for the formulas used in the trim and fill procedure. The formulas for the three estimators, R_0, L_0 and Q_0, are as follows:

$$R_0 = \gamma^* - 1 \tag{21}$$

$$L_0 = \frac{4T_k - k(k+1)}{2k-1}, \text{ and} \tag{22}$$

$$Q_0 = k - \frac{1}{2} - \sqrt{2k^2 - 4T_k + \frac{1}{4}} \tag{23}$$

(Duval & Tweedie, 2000). The R_0 estimator has the simplest form (see Equation 21). Both L_0 and Q_0 are more complex functions of the number of observed studies, *k*, and the trimmed rank test statistic T_k.

It should be noted that for Q_0 to be calculable, the function under the square root sign must obviously be greater than or equal to zero. Thus, for Q_0 to be well defined, the following must hold: $T_k \leq \frac{k^2}{2} + \frac{1}{16}$. When this is not the case, in practice, then the value of Q_0 is set to $\left(k - \frac{1}{2}\right)$. Even though it is not explicitly mentioned by Duval and Tweedie (2000a, 2000b), it should be emphasized that the simple one-step use of these three trim and fill estimators is based on the assumption that the population effect size is known to be zero. The trim and fill method that is used when the population effect size is *unknown* will be described later.

According to Duval and Tweedie (2000b), L_0 seems to provide smaller estimates of k_0 than Q_0 does. The authors also claimed that L_0 has a smaller mean square error than R_0 (except in the region approximated by $k_0 \geq \frac{k}{4} + 2$ for larger values of k_0 at any fixed *k* level). They also concluded that L_0 and Q_0 appear to be more robust than R_0 to the situation in which there is a relatively isolated negative effect size estimate with a very high rank. Last, they suggested using the results from the estimator L_0 when the number of estimated missing studies ($\hat{k}_0$) is greater than the number of observed studies (*k*).

Because in practice, the number of missing studies is conceptualized as a whole number, not a fraction, the trim and fill estimators should be calculated using the following forms:

$$R_0^+ = \max\{0, R_0\}, \tag{24}$$

$$L_0^+ = [\max\{0, L_0 + \frac{1}{2}\}], \tag{25}$$

$$Q_0^+ = [\max\{0, Q_0 + \frac{1}{2}\}], \tag{26}$$

where f(*x*) = [*x*] represents the function that provides the integer part of *x* (Duval and Tweedie, 2000a, 2000b).

The reason why this method is called the Trim and Fill method is that it also provides a way to correct the possible publication bias by filling in the missing points in a funnel plot (see Figures 1 and 2) in order to get a better estimate of the target

effect size. Because the current study only focuses on assessing publication bias, the correction function of the trim and fill method will not be discussed here.

The trim and fill method used when the population effect size is unknown (but assumed positive) involves iteratively estimating, trimming, and filling funnel plots (Duval and Tweedie, 2000a, 2000b). It involves the following iterative steps:

(1) Calculate a weighted average, $\overline{T}_.$, of the unknown population effect size (θ) from the observed dataset of $\hat{\theta}_i^*$ s (for i = 1, 2… k) using either a fixed effects (see Equation 1) or a random effects model (Raudenbush, 1994) and denote this first estimate by $\overline{T}_.^{(1)}$. Then, center each of the k observed effect size estimates around this average $\overline{T}_.^{(1)}$ with this centered effect size denoted by $\hat{\theta}_i^{*(1)}$, where $\hat{\theta}_i^{*(1)} = \hat{\theta}_i^* - \overline{T}_.^{(1)}$ for i = 1, 2… k.

(2) Estimate the number of missing studies, k_0, using one of the three trim and fill estimators (see Equations 20 through 22 and 23 through 25) based on the centered dataset $\hat{\theta}_i^{*(1)}$ (i = 1, 2… k), and denote the estimate by $\hat{k}_0^{(1)}$.

(3) Remove the $\hat{k}_0^{(1)}$ largest, positive observed effect size values from the original observed dataset $\hat{\theta}_i^*$, and re-calculate $\overline{T}_.$ using these $[k - \hat{k}_0^{(1)}]$ observed values to obtain $\overline{T}_.^{(2)}$. The newly centered dataset is $\hat{\theta}_i^{*(2)} = \hat{\theta}_i^* - \overline{T}_.^{(2)}$ (for i = 1, 2… k).

(4) Re-estimate k_0 based on the centered dataset $\hat{\theta}_i^{*(2)}$ using the same trim and fill estimator that was used in step (2), and denote this estimate by $\hat{k}_0^{(2)}$.

(5) Remove $\hat{k}_0^{(2)}$ largest values from the original observed dataset $\hat{\theta}_i^*$, and re-estimate $\overline{T}$. using these $[k - \hat{k}_0^{(2)}]$ observed values to obtain $\overline{T}_\cdot^{(3)}$. Construct the dataset $\hat{\theta}_i^{*(3)} = \hat{\theta}_i^* - \overline{T}^{(3)}$, i = 1, 2… k, and again estimate k_0 from $\hat{\theta}_i^{*(3)}$.

(6) Repeat the procedure until the stopping rule shown below is met, and denote the final estimate of the number of missing studies by $\hat{k}_0^{(F)}$.

The procedure has converged on a solution when: $\hat{k}_0^{(J)} = \hat{k}_0^{(J-1)} = \hat{k}_0^{(F)}$

[where $\overline{T}^{(J)} = \overline{T}^{(J-1)}$ is obtained].

Duval and Tweedie (2000a) concluded based on the iterative algorithm above that it is only possible to find $\overline{T}_\cdot^{(J)}$ in iteration J in step (6) if $k_0 \leq k - 1$. They also found that if the data are fitted using a fixed-effects model, the iteration algorithm will converge in no more than four iterations, however, if the random-effects model is used, the iteration algorithm might not converge. Based on the results of their simulation study results, the authors concluded that the iterative estimate and the non-iterative versions have essentially the same properties. However, because the non-iterative version is for use when the true effect size is known, the iterative estimator should typically be used. Duval and Tweedie concluded that using the R_0^+ estimator, for which the distributional properties are known, (R_0 is assumed to have a negative

binomial distribution), could provide a powerful test of the null hypothesis that the number of missing studies is zero. In addition, this hypothesis test should also maintain a reasonable type I error rate (Duval & Tweedie, 2000b).

For publication bias detection purposes, simulation studies with different combinations of k_0 and $k_0 + k$ were conducted. Hypothesis tests of $k_0 = 0$ and rejection regions as $\{R_0^+ >$ some reasonable integer$\}$ based on the distributional properties of R_0^+ were performed. Results showed that in the region of $\{R_0^+ > 3\}$, the nominal alpha was smaller than 5% when the null hypothesis was true. For the alternative hypotheses of $k_0 \geq 7$, the power of the test also was shown to be higher than 0.80.(Duval & Tweedie, 2000b). Kromrey and Rendina-Gobioff (2006) found that the Trim and Fill method exhibited conservative Type I error control, and showed less power than Begg's rank correlation method (both r_{tn} and r_{tv}) and Egger's regression method. In addition, in a simulation study conducted by Terrin et al.(2003), the Trim and Fill method appeared to adjust publication bias inappropriately where no publication bias was introduced, when the collection of studies estimated more than one single true effect, (i.e. in the presence of heterogeneity).

As evidenced by this summary of methodological articles, many methods for assessing publication bias have been developed and there is still little consensus about which method works best. A review was conducted of recently published applied

meta-analyses to assess which publication bias methods are most commonly used. The next section describes this review.

Review of Publication Bias Assessment Practices

A review was conducted of all the published articles in the most recent two years (including 2006 and up to September of 2007) of the *Review of Educational Research* and *Psychological Bulletin* journals to assess the degree to which applied meta-analysts assessed the possibility of publication bias. Articles involving quantitative meta-analyses were selected from a total of 128 articles, resulting in 37 articles involving 75 quantitative meta-analyses. Within those articles, only 23 mentioned anything about publication bias. Out of these articles that mentioned publication bias, seven used the Trim and Fill method, one used the Egger's Regression method, one used the FSN, and for the remaining 14 articles, the authors said that they had solved the problem of possible publication bias either by including several sources of unpublished papers, such as dissertations or including publication status as a moderator. No one mentioned using either Begg's Rank Correlation Method or Funnel Plot Regression despite results supporting these methods maintaining the nominal alpha levels. Although the Trim and Fill method seems slightly more complex than the other methods, it was more commonly used. This might be because code has been implemented in statistical software to calculate the relevant indices (e.g., Johnson, Chang & Lord, 2006; Unsworth & Engle, 2007). The lack of common method used in those applied articles that have been reviewed suggests that the lack of consensus in methods used for assessing publication bias in Meta-analysis.

Statement of Purpose

Publication bias causes a lot of problems in meta-analysis as it can negatively impact the validity of meta-analytic results. Meta-analysts should be ultimately interested in correcting for possible publication bias but the most important first step involves finding a reliable indicator of the existence of publication bias.

There are many methods that can be used to detect publication bias, including those mentioned in the previous sections. However, it appears that even though a lot of research has been done to evaluate those methods, there has been no consistent answer to which method performs better than the others. For example, Sterne, Gavaghan and Egger (2000) recommended Egger's regression method, whereas Macaskill, Walter and Irwig (2001) recommended funnel plot regression method using $\hat{\beta}_{1,\sigma_p^2}$. In addition, as found in the brief summary of previous applied meta-analyses, some methods are more popular than others, but this popularity is not because the methods are simply better than the others but possibly solely due to the easier implementation of a particular method over others. Again, the problem is that there is no consensus about which method performs the best overall. Therefore, the goal of this proposed study is to extend the literature in three ways. First, compares the performance of those mentioned methods by introducing different sample size for each study. Second, compare the performance of those mentioned methods using the $Z_{\hat{\rho}}$ metric, which can be assumed asymptotically normal when the sample size is large, instead of $\hat{\delta}$ and $\hat{\rho}$ .Third, compare inferential differences across numerical

methods for publication bias assessment when studies have unequal sample sizes and use $Z_{\hat{\rho}}$ metric, i.e. FSN versus three Trim and Fill estimators, R_0^+, L_0^+ and Q_0^+.

The methods that will be compared in the current study include Rosenthal's Fail-Safe number, Begg's rank correlation method, Egger's OLS regression method, funnel plot regression method and the Trim and Fill method. A simulation study will be conducted to assess the impact of several factors on the identification of publication bias for a meta-analysis of correlations. These different simulation conditions include the number of studies ($k + k_0$), true effect size (ρ), and the magnitude of the publication bias (k_0 values). For all cases in which no publication bias ($k_0 = 0$) is introduced, the Type I Error rate (i.e., the cases where the null hypothesis of zero missing studies is falsely rejected) will be computed and compared across methods. For all non-zero values of k_0, the statistical power associated with each scenario will be computed and compared.

Method

Manipulated Conditions.

A Monte Carlo study that simulates different sets of meta-analyses will be used to evaluate the impact of certain conditions on the identification of publication bias. Three design factors will be manipulated and will be discussed in detail.

Number of Observed Studies.

The number of observed studies, *k*, included in a meta-analysis will be investigated at three levels representing a small, moderate and large number of studies. Specifically, the sizes of 20, 50 and 100 will be used.

Number of Missing Studies.

The number of missing studies in a meta-analysis will be investigated using four levels, specifically, zero, small, moderate and large values for k_0. The small value for k_0 is 10% of the number of observed effect sizes in a meta-analysis, (i.e., $k_0 = 0.1k$). The moderate value for k_0 will be 20% and the large value will be 40% of k.

True Effect Size.

The true value of ρ will also be investigated using four different true values (including 0, 0.2, 0.5 and 0.8). Because there are some notable disadvantages of ρ, Fisher's (1928) normalizing (see Equation 6) and variance-stabilizing (see Equation 7) ρ-to-Z_ρ transformation will be used in this simulation. Specifically, data will be generated in the form of Z_ρ and all analyses will be conducted using the Z_ρ metric.

In summary, there will be three different k values, four different k_0 values and four different Z_ρ values in this fully crossed design that results in forty-eight different combinations of conditions. For each combination of conditions, 1,000 replications will be generated

Data Generation.

For each study *i*, in each meta-analysis entailing *k* effect sizes, the effect size estimate and its associated variance must be sampled from the relevant sampling distributions. To accomplish this, first the sample size for study *i*, n_i, needs to be generated. Previous methodological research on publication bias (Kromrey & Gobioff, 2006, and Macaskill, Walter & Irwig, 2001) has typically used the same sample size for each study in the simulated meta-analyses. In an attempt to make this study better mimic an applied meta-analysis, sample size per study within each meta-analysis will be modeled to vary. Following Hafdahl (2007), n_i will be generated based on the following equation

$$n_i = \left[\left(\frac{\bar{n}}{2} \right) \left(\frac{X_i - 3}{\sqrt{6}} \right) + \bar{n} \right] \qquad (26)$$

where $\bar{n}$ = 100, X_i is a value sampled from a χ^2 distribution with three degrees of freedom, and, as in Equation 25, f(*x*) = [*x*] represents the function that extracts the integer part of *x*. After obtaining the value of n_i, Equation 7 can be used to obtain the resulting value for $\hat{\sigma}^2_{Z_{ri}}$. Next Z_{ri} can be generated by taking a random sample from a

normal distribution with a mean of Z_{ρ_i} and a standard deviation of $\sigma_{Z_{ri}}$. For the case that a meta-analysis includes k observed studies and k_0 missing studies, $k + k_0$ sample sizes will be generated to obtain n_j, for j = 1 to $k + k_0$. Then, $k + k_0$ associated values of Z_r s will next be generated.

Parameter Values and Evaluation

For each simulated meta-analysis, the following publication bias detection methods: FSN method, Begg's rank correlation method, Egger's regression method with OLS estimation, funnel plot regression method and the Trim and Fill method will be used. For each of these methods, the decision about whether to reject the null hypothesis that there is no publication bias will be recorded. For the FSN and the Trim and Fill methods, the resulting estimates of k_0 will be obtained (see Equations 12, 24, 25 and 26).

For each combination of conditions, 1,000 meta-analytic datasets will be generated. For all the methods, the proportion of the replications that publication bias is identified when the true k_0 is zero will be tallied to provide a Type I error rate. Similarly, the proportion of replications for which publication bias is correctly identified will be summarized when the true k_0 is not zero to provide a measure of empirical power. In addition to these proportions, the bias in estimation of k_0 for the

FSN and Trim and Fill methods will be gathered when there is no publication bias.

Last, the relative bias $\frac{\hat{k}_0 - k_0}{k_0}$ will be assessed when the true k_0 is nonzero.

Expected Results

Publication bias remains as one of the major criticisms in meta-analysis. Even though there are many methods developed for assessing the presence of publication bias in meta-analysis, not many applied meta-analysts use these methods. In addition, some methods were identified as being preferred over others in the short review of applied meta-analyses despite their associated performance as discovered in the relevant simulation studies. The purpose of this proposed study is to contribute to the assessment of the performance of publication bias detection methods by including consideration of scenarios in which each study might be based on a different sample size.

Type I error rate

Type I error rate will be considered only when no publication bias is introduced, that is when the number of true missing studies is zero ($k_0 = 0$). In general, for all the methods, except FSN and the Trim and Fill method, it is expected that the larger the true value of effect size ρ is, the better the Type I error control. This was showed by the simulation studies conducted by Kromrey and Rendina-Gobioff (2006) and Macaskill, Walter and Irwig (2001). However, depending on which method is used, increasing the number of studies, *k*, is not always expected to improve Type I error control. The FSN is expected to exhibit good Type I error control when the true value of ρ is zero. However it is expected to perform extremely badly when ρ is nonzero regardless of the size of *k*. This is because that the FSN

assumes that the missing studies have a null effect that is the average effect of those additional studies is zero. Begg's rank correlation method, both based on the variance $\sigma_i^{2^*}$ and on the sample size n_i, is expected to demonstrate liberal Type I error control in all cases. Begg's rank correlation method based on the variance $\sigma_i^{2^*}$, showed liberal Type I error control in Kromrey and Rendina-Gobioff's (2006) simulation study, and Begg's rank correlation method based on the sample size was said to have Type I error control close to nominal alpha level in both the simulation studies conducted by Kromrey and Rendina-Gobioff (2006) and Macaskill, Walter and Irwig (2001). It is also expected that, using Begg's rank correlation method, there will be a slightly decrement in Type I error control when the number of studies increases, The two regression methods, Egger's regression method and funnel plot regression method, are expected to have liberal Type I error rates in most conditions when the number of studies is not small. Both of these regression methods are expected to have increasing control as the number of studies increases. The Trim and Fill method is expected to be conservative in terms of its Type I error control in most conditions. It is also expected that Type I error control will increase as the number of studies increases. Overall, (1)Begg's rank correlation method is expected to perform the best among all the methods in Type I error control in all the conditions, except that FSN is expected to be better when the true value of the effect size is zero; and (2) Egger's regression method is expected to be better than other methods in most conditions, especially when the number of studies is not small.

Empirical power

Power will be considered only when publication bias is introduced, that is when the number of missing studies is not zero. Based on the results provided by the two simulation studies conducted by Kromrey and Rendina-Gobioff (2006) and Macaskill, Walter and Irwig (2001), and the design for this current simulation study, namely the unequal sample sizes and the use of the $Z_{\hat{\rho}}$ metric, in general, the power of using all the methods is not expected to be high. Power is expected to be relatively high in the conditions that the number of studies is large and the number of missing studies is also large. When using FSN, it is expected that when the true value of the effect size is zero, power will be higher than when the true value of the effect size is nonzero. Begg's rank correlation method, both based on $\sigma_i^{2^*}$ and n_i, is expected to show a better performance in power as the true value of the effect size gets smaller. Egger's regression method is expected to have the same pattern in power as does Begg's rank correlation method, whereas funnel plot regression and Trim and fill methods are expected to show an opposite pattern that the power would be higher in the conditions that the true value of the effect size is larger. Overall, Begg's rank correlation and Trim and Fill methods are expected to be better in power than the other methods.

Limitations and Future Research

Detecting publication bias in meta-analysis is crucial, because publication bias affects the validity of the resulting pooled effect size estimate. However, once evidence has been found for publication bias, it behooves the meta-analyst to try to correct for this bias. In the current study, however, the main focus is to compare methods used to *assess* rather than *correct for* publication bias. Methods for correcting publication bias, such as the Trim and Fill method and selection modeling, should be addressed and compared in future studies. In addition to that, developing new methods for assessing and correcting publication bias are encouraged as a direction for future research.

In the current study, publication bias was introduced in the manipulated conditions by using different percentages of the number of the observed studies. By using a fixed numbers of missing studies (i.e., of k_0), this provided a way to compare the estimated number of missing studies to the known number of missing studies. However, this restricted the pattern of missing studies to only those that had the smallest k_0 effect size which probably does not mimic all real-world meta-analytic scenarios.

Several methods that are used for detecting publication bias were described and compared in the current study. Except for the FSN and Begg's rank correlation method, all of those methods (including Egger's regression, funnel plot regression and the Trim and Fill methods) assess publication bias based on examining the asymmetry of a funnel plot. Even though, the manipulated conditions for the current

study introduce publication bias as the only cause for funnel plot asymmetry, however, a weakness is that researchers have pointed out that publication bias is not the only possible reason for funnel plot asymmetry. Thus, conditions modeling other sources of funnel plot asymmetry, such as heterogeneity of data, need be considered in data generation procedures and introduced into the comparison of methods in future studies.

One more possible direction for future study is to compare those methods in publication assessment using different meta-analytic effect size statistics. The current study only considered the different conditions with the use of the $Z_{\hat{\rho}}$ metric, because it can be assumed asymptotically normal with large sample size and no simulation studies have worked on the $Z_{\hat{\rho}}$ metric. For future research, simulation study can be conducted to compare the performance of publication bias detecting methods using the $Z_{\hat{\rho}}$ metric versus $\hat{\delta}$ metric, for example.

Publication bias is one of the major criticism for meta-analysis, and it will remain as one of the most commonly criticism for the impact it can have on the validity of meta-analytic results. By including those missing studies' effect size estimates, usually the relative small ones in most areas of studies, in a meta-analysis, the inference would change from significance to non-significance. The change of inference may cause a lot of problems and have impact on many serious issues. In the Educational Psychology field, for example, to adapt a new teaching method may cost a lot of money and time. More important, the new adapted teaching method may

affect students' motivation, strategic thinking, and so on, in learning. If the decision of adapting a new teaching method were based on the result of some meta-analysis, and there were publication bias associated with this meta-analytic result, which means in fact this new method is not as good as the old one, people's lives might have bad impact. In sum, meta-analysts should pay heavy attention to publication bias in their practice in meta-analysis.

Addendum

Meta-analysis is a statistical technique consisting of a set of quantitative methods that accumulates multiple results across studies designed to assess a set of related research hypotheses. It provides an aggregated estimate of an effect size parameter that is more comprehensive and thus more powerful than those derived from single studies. The generalizability of meta-analytic results makes meta-analysis a technique that can and should be used as a part of program evaluations. Use of meta-analysis could benefit evaluators by helping them enhance the design of their evaluation studies (e.g., specifying program inputs and constraints), by helping them choose data analysis procedures and then by helping evaluators select the appropriate variables to be measured. An artificial example will be used to illustrate how meta-analysis could be used to aid program evaluation.

Let Xin be a program evaluator, and as an evaluator she was asked to evaluate the effectiveness of a training program that was designed to benefit participants' lives. One of the stakeholders asked specifically whether this training program improved one's life satisfaction . For this particular question, Xin's first thought to use a paired-samples t statistic, involving pre-training and post-training life satisfaction scores, which would be measured using some well-established life satisfaction assessment. After carrying out the analysis, however, Xin found that the results of the statistical test were not significant. After showing the non-significant result to her supervisor, Xin was required to find out the reason(s) as well as to design and run another analysis.

Xin went back to her evaluation design, and had several hypotheses such that cultural background, personality and gender would be considered as constraints and then other data analysis techniques should be used instead of the paired-t statistic, repeated measures design. Unfortunately, data on hand was limited as these variables were not assessed, and so these hypotheses could not be tested directly. In addition, for financial considerations, Xin must provide some strong evidence that her hypotheses were supported before calling back those participants to obtain their scores on these three variables.

To find support for her hypotheses, Xin first conducted a literature search to find articles that assessed the relationship among cultural background, personality, gender and life satisfaction, and she found many studies over the past few years that were aimed at this topic. However, because those studies were conducted under different circumstances, including use of different personality constructs, different assessments for life satisfaction as well as different sample sizes, and for different groups of participants, the results were not consistent nor could they be adapted directly to answer Xin's research question. Lucky, Xin remembered that meta-analysis, as mentioned above, can be used to generalize results by controlling for study characteristics, and, therefore, she thought meta-analysis could be used for obtaining evidence for her hypotheses.

Detailed procedures of how to carry out a meta-analysis was described at the beginning of the Literature Review, and in general, a researcher needs to (1) search for literatures, (2) choose relevant studies according to the researcher's incorporation

criteria among these literatures, (3) decide appropriate effect size statistic, (4) calculate the relevant effect size for each study, (5) pool together the effect sizes to obtain an overall average effect size estimate under the assumption of homogeneous effects, and then if homogeneity cannot be assumed (6) pool together the effect sizes to obtain an overall average effect size estimate under the assumption of heterogeneous effects. Last, but not least, the potential for publication bias must be assessed and potentially corrected when conducting steps (5) and (6). For simplicity, let's consider only one of Xin's hypotheses, namely, that females tend to be more satisfied with their lives than males.

According to her research hypothesis, Xin was interested in summarizing past research investigating the different views of life satisfaction between females (coded as 1) and males (coded as 2) using different life satisfaction assessments. Suppose she finished her literature search, including results from the relevant studies, determined the effect size statistic (i.e., the standardized difference in means) and obtained the results contained in Table 1.

Table 1.

Summary of the Data Set

Study ID	Mean1	SD1	N1	Mean2	SD2	N2	$\hat{\delta}_i$	v_i
1	0.56	0.29	133	0.38	0.44	151	0.476	0.015
2	0.66	0.25	54	0.69	0.24	44	-0.121	0.041
3	0.55	0.24	73	0.44	0.34	68	0.374	0.029
4	0.71	0.16	73	0.57	0.29	68	0.600	0.030
5	0.37	0.34	46	0.28	0.30	41	0.277	0.047
6	0.60	0.27	46	0.60	0.25	41	0.000	0.046
7	33.08	0.69	64	29.14	0.64	81	5.916	0.149
8	15.18	5.23	243	13.56	3.76	220	0.352	0.009
9	15.68	4.84	203	13.76	3.23	217	0.469	0.010
10	44.42	7.54	243	38.95	6.44	220	0.776	0.009
11	41.02	8.00	203	39.35	6.49	217	0.230	0.010
12	5.12	0.88	172	4.41	0.86	120	0.812	0.015
13	34.88	8.69	120	32.03	7.18	118	0.356	0.017

To calculate the desired effect size estimate $\hat{\delta}_i$ for each study, Xin first computed Hedge's *g* for each study using Equation 2, and then

$$\hat{\delta}_i = \hat{g}_i(1-\frac{3}{4m_i-1}), \qquad (27)$$

where $m_i = (N1)_i + (N2)_i - 2$. For example, the first study was consisted of 133 females and 151 males, and the average life satisfaction scores for the female and male groups were 0.56 and 0.38 according, with associated standard deviations of 0.29 and 0.44. So, the calculations Xin did in order to obtain the effect size estimate $\hat{\delta}_1$ and its associated variance in Table 1 could be specified step by step as following:

(i) Find the pooled variance for study 1 as

$$S_{p_1}^2 = \frac{[(N1)_1 - 1]SD1^2 + [(N2)_1 - 1]SD2^2}{(N1)_1 + (N2)_1 - 1}$$

$$= \frac{(133-1)0.29^2 + (151-1)0.44^2}{133+151-1} = \frac{(132)(0.0841) + (150)(0.1936)}{283}$$

$$= 0.141842$$

(ii) Compute Hedge's g for study 1 using Equation 2 as

$$\hat{g}_1 = \frac{(Mean1)_1 - (Mean2)_1}{\sqrt{S_{p_1}^2}}$$

$$= \frac{0.56 - 0.38}{\sqrt{0.141842}} = \frac{0.18}{0.376619} = 0.477937$$

(iii) Find the unbiased estimate of δ for study 1 using Equation 27 as

$$\hat{\delta}_1 = \hat{g}_1(1 - \frac{3}{4m_1 - 1})$$

$$= 0.477937\left(1 - \frac{3}{4(133+151-2)-1}\right)$$

$$= 0.477937(1 - 0.002662) = 0.476665 \approx 0.477$$

(iv) The associated variance is calculated by Equation 3

$$v_1 = \tilde{n} + \frac{\hat{\delta}_1^2}{2(N1)_1 + 2(N2)_1} = \frac{133+151}{(133)(151)} + \frac{0.477^2}{2(133)+2(151)}.$$

$$= 0.01414 + 0.0004 = 0.014541 \approx 0.015$$

Answers from steps (iii) and (iv) were then recorded into the last two entries in Table 1 for study 1. For the remaining 12 studies, similar computations were carried out using steps (i). For demonstration purposes, suppose also that Xin had assessed for publication bias and found there's no evidence of such bias, and therefore, she continued her analysis based on those 13 studies. She computed the overall pooled

point estimate $\bar{\delta}_{\cdot}$ under the assumption of homogeneous effects using Equation 1 and a step-by-step illustration is provided as following:

(i) Compute the weight for study 1 as

$$w_1 = \frac{1}{v_1} = \frac{1}{0.015} = 68.7695.$$

(ii) Compute the weighted effect size estimate for each study as

$$wd_1 = w_1\hat{\delta}_1 = (68.7695)(0.477) = 32.7799.$$

(iii) Find the weights and weighted effect size estimates for the rest 12 studies with the procedures showed in steps (v) and (vi).

(iv) Sum up the weights and weighted effect size estimates across all 13 studies.

(v) Compute the overall pooled estimate as

$\bar{\delta}_{\cdot} = \frac{\sum wd}{\sum w} = \frac{377.908}{762.558} = 0.494$, with associated variance

$$v_{\cdot} = \frac{1}{\sum w} = \frac{1}{762.558} = 0.001.$$

Next, she computed the value of the test statistic $z = \frac{\bar{\delta}_{\cdot}}{\sqrt{v_{\cdot}}} = \frac{0.494}{\sqrt{0.001}} = 13.632$ with the relevant critical test statistic. Xin tested her hypothesis at α = 0.05, and concluded that the conception of life satisfaction for females was significantly higher than that for males in the population under the assumption of homogeneous effects.

Before making her conclusion, Xin also tested the assumption of homogeneous effects at $\alpha = 0.05$ using *Q*-statistic. To find the value of the Q-statistic, a *q* value was first computed for each study, and using study 1 again as an example, its *q* value was $q_1 = \frac{\left(\hat{\delta}_1 - \bar{\delta}_.\right)^2}{v_.} = \frac{\left(0.477 - 0.494\right)^2}{0.001} = 0.3005$. Therefore the value of the *Q*-statistic then was $Q = \Sigma q_i = 240.02$, which is assumed to have a chi-square distribution with 12 (i.e., the number of included studies minus 1) degrees of freedom. Thus, Xin inferred that the null hypothesis of homogeneous effects sizes should be rejected and assumed instead a random-effects model.

The steps of how to estimate a random-effect model are outlined as following:

(i) Find the variance for the 13 effect size estimates $\hat{\delta}_i$, which was 2.424 for Xin's data set (contained in Table 1).

(ii) Find the average for the 13 variances associated with the effect size estimates $\hat{\delta}_i$, i.e. v_i ,and for the illustrated data set it was 0.033.

(iii) Calculate the variability between these 13 studies using the values found in steps (x) and (xi), and $\hat{\sigma}^2_{\hat{\delta}_i} = 2.424 - 0.033 = 2.391$.

(iv) Add the newly obtained between variability to the individual variability for each study found in step (iv), and as for study 1 the variability for the random-effect model was then $v_1^R = v_1 + \hat{\sigma}^2_{\hat{\delta}_i} = 0.015 + 2.424 = 2.439$.

(v) Repeat steps (v) to (ix) to obtain the overall pooled estimate by replacing

all the v_i s by the v_i^R s, and the sum of the 13 weighted effect size estimates using random-effect model was 4.240 while the sum of the 13 weights was 5.363. Therefore, the overall pooled point estimate using the random-effect model was $\bar{\delta}_{\cdot} = \frac{\sum wd_{new}}{\sum w_{new}} = \frac{4.240}{5.364} = 0.790$, with associated variance $v_{\cdot} = \frac{1}{\sum w_{new}} = \frac{1}{5.364} = 0.186.$

Therefore, the value of the test statistic using the random-effect model was $z = \frac{\bar{\delta}_{\cdot}}{\sqrt{v_{\cdot}}} = \frac{0.790}{\sqrt{0.186}} = 1.831$, which led to the conclusion that the conception of life satisfaction for females was significantly higher than that for males in the population under the assumption of heterogeneous effects, too.

Given both the random- and fixed-effects pooled estimates of the effect size were found to be statistically significant, Xin could conclude that females had significantly higher life satisfaction than males. Thus, she had strong evidence to support her proposal requesting participants' gender and could then re-analyze her data including gender as a covariate. In a similar fashion, meta-analyses can and should be used to provide support for evidence-based practice.

In summary, meta-analysis is a powerful statistic technique, which has been used in many different disciplines. Its advantages of generalizability to relevant populations and providing powerful aggregated estimate over estimates from a single study also provide program evaluators with an alternative way to help evaluation

design, as well as for helping evaluators choose data analysis procedures and variables to be measured, especially when on-hand resources are limited.

References

Ackerman, L. P., Beier, E. M., & Boyle, O. M. (2005). Working memory and intelligence: the same or different constructs. *Psychological Bulletin*, 131, 30-60

Arthur, W., Bennett, W., & Huffcutt, A. (2001). *Conducting meta-analysis using SAS.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Begg, C. B. (1994). Publication Bias. In Cooper, H.& Hedges, L.V. (Eds.), *The handbook of research synthesis*. (pp. 399-409). New York, NY: Russell Sage Foundation.

Bernard, et al. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature analysis of the empirical literature. *Review of Educational Research*, 74, 379-439.

Cheung A., & Slavin R. (2005) A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research*, 75, 247-284.

Cooper, H.& Hedges, L.V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Duval, S. & Tweedie, R. (2000). A nonparametric “trim and fill” method of ccounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95, 89-98.

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56,

455-463.

Duval, S., Tweedie, R., Abrams, K.R. & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal*, 320, 1574-1577.

Fisher, R. A. (1928). *Statistical methods for research workers* (2nd Ed.). London: Olover & Boyd.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, 5, 3-8.

Hedege, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic.

Heller, D., Ilies, R., & Watson, D. (2004). The role of person versus situation in life satisfaction: a critical examination. *Psychological Bulletin*, 130, 574-600.

Kromrey, J. D., & Rendina-Gobioff, G. (2006). On knowing what we do not know. *BEducational and Psychological Measurement*, 66, 357-373.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: the science of reviewing research*. Cambridge, MA: Havard University Press.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20, 641-654.

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455-463.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L.V. Hedges

(Eds.), *The handbook of research synthesis* (pp. 301-321). New York, NY: Russell Sage Foundation.

Rosenthal, R. (1979). The “file-drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638-641.

Sterne, J., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, 54, 1046-1055.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22, 2113-2126.

VITA

Xin Li was born in Chengdu, Sichuan, China, the daughter of Li, S. R. and Ren, X. F. After completing her high school in Jinjinag High School in 1998, she entered the University of Tennessee at Chattanooga in 2000. She received the degree of Bachelor of Science from the University of Tennessee at Chattanooga in 2004. In September 2004, she entered the graduate school at the University of Texas at Austin.

Permanent Address: #12 Gong Nong Yuan Street B-101

Chengdu, Sichuan 610000

China

This report was typed by the author.