NORTHWESTERN UNIVERSITY

Methods for Generalizing from Experiments

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Statistics

By

Laura Elizabeth Tipton

EVANSTON, ILLINOIS

June 2011

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

Methods for Generalizing from Experiments

Laura Elizabeth Tipton

This dissertation is a collection of three papers that address issues of generalization and external validity through a statistical lens.

Chapter 1 addresses the problem of non-random sample selection bias in randomized experiments. This paper extends the use of propensity score matching methods that are often used in observational studies to address the problem of sample selection bias in experiments. Using a potential outcomes framework, I delineate the assumptions required to use this method, propose a subclassification estimator, and develop benchmarks for the expected amounts of average bias reduction and average variance inflation. For a well-defined population, I also provide a method for choosing a sub-population such that the experimental estimate is less biased.

Chapter 2 addresses the problem of parameterizing heterogeneous treatment effects (HTE) in randomized experiments. In addition to the standardized mean difference, this paper proposes a standardized individual treatment effect variance and argues that both of these parameters are needed – either separately or as a ratio –to summarize how well a treatment works for a particular population. After introducing the definitions of the parameters, I develop estimators of these parameters and their large-sample variances. I provide a simulation study to investigate small sample bias and interval coverage probabilities under three different treatment and control distribution pairs.

Chapter 3 addresses the problem of robust variance estimation in meta-analysis when the effect sizes are functions of correlated binary variables. I report the results of a large simulation study which focuses on the risk difference, log risk ratio, and log odds ratio effect sizes. This simulation study examines the accuracy of 95% confidence intervals constructed using a robust variance estimator when the number of studies in the meta-analysis is small. I report results for both estimation of the mean effect (intercept) and the estimation of a slope.

Acknowledgements

I am extremely grateful to Larry Hedges, Bruce Spencer, and Tom Cook for their support throughout my time at Northwestern and their service on my dissertation committee. Many of the ideas found in this dissertation – particularly in Chapters 1 and 3 – grew out of projects with my advisor, Larry Hedges. While I have benefited greatly from his kindness, encouragement, and friendship throughout all of my years of graduate study, in particular his advice and mentorship this past year have been invaluable. Additionally, I am indebted to Patti Ferguson for always looking out for me -- for finding time in a busy schedule, helping me navigate administrative details, and for her general enthusiasm and support.

Vivian Wong and Nate Jones have been like family, and without their friendship this dissertation may never have been completed. Our time working together in the library and attic space has proven that productivity and laughter can and should go hand in hand. Additionally, thanks are due to my friends and classmates Yuan Liao, Chris Rhoads, and Lili Yao – without them I would not have made it through the early years of my graduate coursework. Finally, Judy Hedges has been a continued source of support and friendship throughout my many years of graduate study, and for that I am grateful.

My parents and parents-in-law have been constant sources of inspiration and support during my graduate studies – thanks to Priscilla and Greg Pieratt, Tom Tipton, Beverly Lewis, and Collins Lewis. I will be forever thankful that I convinced my sister – Sarah Tipton – to move to Chicago three years ago. Most importantly, thanks are due to Jason Lewis, my husband and partner in life. Without his love and support, I certainly would never have started this Ph.D. program, much less finished it.

Table of Contents

List of Tables

List of Figures

Introduction

In the fields of education, medicine, and the social sciences, randomized experiments are conducted to test the effects of treatments or interventions on populations. The results of these experiments inform policy and practice at the societal, institutional, and individual levels. The scope of these generalizations highlights the role that external validity plays in experiments – where by external validity I mean the interpretation of the results of an experiment or study beyond the units, treatments, outcomes, and settings in the study (Shadish, Cook, and Campbell, 2002). This dissertation is a collection of three papers that address issues of generalization and external validity through a statistical lens. I focus on a subset of these concerns, including bias from non-random sample selection (Chapter 1), the interpretation and estimation of treatment effects under heterogeneity (Chapter 2), and the combination of both experimental and non-experimental results across studies via robust meta-analysis (Chapter 3).

Chapter 1 addresses the problem of sample selection bias in experiments (Hedges and O'Muircheartaigh, 2010). While within experiments, units are randomly assigned to treatment conditions (insuring high internal validity), units are rarely randomly sampled into the experiment from a well-defined population. This paper develops a method which applies propensity score matching methods – commonly used to reduce bias from treatment selection in observational studies– to the problem of reducing bias from sample selection in experiments. The utility of this work is that it increases the precision of the generalization process. The method developed here begins with an explicitly defined population and identifies its associated estimate of the population ATE and its standard

error, based on the experimental sample. In cases in which the population is not well represented by the experimental sample, the standard error of the estimate of the population ATE is very large. This means that the results of an experiment may generalize to some populations better (with greater precision) than others. This paper provides an estimator of the population ATE with less bias than the conventional estimator and an analysis of the amount of bias reduction and variance inflation that may occur when used in practice.

Chapter 2 addresses the problem of measuring unit treatment effect variance. The results of experiments are primarily summarized through estimates of the average treatment effect. However, policy makers are often interested in the degree to which the effect of a treatment varies across units. For example, Treatments A and B may have respective effect size of 0.20 and 0.10 – indicating a preference for A – but if Treatment A also has more variable effects then for some units Treatment B may be favored. In this paper, I introduce a standardized treatment variance parameter to supplement the standardized mean difference, as well as a combined measure, the coefficient of variation (and its inverse) for the treatment effect. The technical problem this paper addresses is that the unit treatment effect variance can never be fully identified, since it depends on the correlation between potential outcomes. I develop a method for addressing this problem, derive parameter and variance estimators, and provide results from a simulation study focused on the small sample properties of these estimators in terms of both bias and confidence interval coverage.

Chapter 3 differs from these first two in that it does not limit its focus to experiments. One method for determining if the treatment effect found in a study generalizes is if it can be replicated; this problem is addressed through research synthesis and meta-analysis. This paper builds on the results of a previous paper I co-authored (Hedges, Tipton, and Johnson, 2010a,b) which provides a new robust variance estimation method for meta-analyses with dependent effects. Dependencies occur when either each study contributes multiple correlated effect sizes (e.g. both reading and math scores are collected from each individual in the study) or when effect sizes are nested in research groups (a hierarchical meta-analysis). Traditional methods for dealing with correlated effect sizes have included: exactly modeling the dependency structure—a method which requires information not often reported in studies; data reduction – using only one randomly selected outcome from each study or their average; or incorrectly assuming independence. In this paper, I focus on the case in which the individual level outcome is a dichotomous variable – e.g., high school dropout – and the outcomes of interest are functions of proportions, including the log odds ratio, log risk ratio, and the risk difference. This paper presents simulation results that test the small sample properties of the method in these cases, under a variety of conditions and offers guidelines regarding sample size and confidence interval coverage.

CHAPTER 1

Improving the External Validity of Randomized Experiments Using Propensity Score Subclassification

While prized for their high internal validity, social experiments typically have low external validity (Shadish, Cook, and Campbell, 2002), making it difficult to generalize from the treatment effect estimated in the experiment to the effect expected in a larger population. For example, this problem occurs when policymakers need to generalize from the results of an experiment conducted on a convenience sample of schools to the effect expected for the population of schools in a particular state. The goal of this paper is to formalize the process of moving from a non-random sample in hand to making inferences about the estimate and standard error of a treatment effect in a well-defined, policy relevant population.

The problem of generalization is rarely addressed in the experimental design literature (Shadish, 2010), and in practice the process commonly employed is like the two-bridges analogy described by Cornfield and Tukey (1956). They argue that making generalizations to a population of interest from an experiment involves two steps or bridges. The first bridge of generalization is a statistical span from the experiment to some putative population "like" it, whereas the second bridge is a subject matter span to the population truly of interest; importantly, this second bridge is largely astatistical. In practice, the decision to generalize is often framed dichotomously: if the convenience sample seems "like" the population of interest, policymakers assume the experimental estimate is unbiased for the population, whereas if the sample is "unlike" the population they decide not to use the experimental estimate at all. Notably, the definition of "like" is usually not made clear.

This paper addresses the problem of external validity in experiments by developing a method for making this second bridge in generalization a statistical bridge. This method is a modification to propensity score matching methods commonly used in quasi-experiments and observational studies to address treatment selection bias (Rosenbaum and Rubin, 1983). Hedges and O'Muircheartaigh (in press) and Stuart, Cole, Bradshaw, & Leaf (in press) recently proposed methods for adapting propensity scores to improve the generalizability of results from experiments. The current paper extends these methods by developing new rules of thumb and theoretical results for the sample selection bias case. While Stuart et al. (in press) develop a method based on inverse-propensity score weighting, we focus here on the propensity score subclassification estimator introduced by Hedges & O'Muircheartaigh (in press).

After first delineating the assumptions needed in the sample selection case, benchmarks for bias reduction and variance inflation (as compared to a conventional estimator) are developed; these results are extensions to Cochran (1968) and Rosenbaum and Rubin (1984). In particular, we show that the propensity score subclassification estimator can be used to both reduce bias in the estimation of a population average treatment effect (PATE) and to identify the portion of a population for which an experiment can generalize with fewer costs in terms bias, variance, and extrapolation. This method also helps in indicating population units that do not have cases "like" them in the experiment. We conclude with an example based on a mathematics intervention aimed at middle-school students in the state of Texas.

Finally, it is important to note that the method we propose focuses exclusively on the process of generalizing from a sample to a population. In the Cronbach (1982) *UTOS* (i.e. Units, Treatments, Outcomes, and Settings) external validity framework, we focus here on generalizing from a sample of units, *u,* to the population of units, *U,* or to a new population of units, *U**. We do not focus on or address other problems in external validity – for example, generalizing across time, changes in the treatment or general equilibrium issues (e.g. Cook, 1993; Schneider & McDonald, 2007). These issues are certainly important, but are beyond the scope of the work presented here.

I. Introduction to the problem of generalization

Let *P* be a well-defined population composed of *N* units for which an estimate of the average treatment effect for a particular intervention is of interest. The statistical ideal for making these inferences would involve first drawing a sample *S* of *n* units randomly from these *N* population units and then to assign *n/2* of these units randomly to the treatment condition and the remaining *n/2* units to the control condition. This dual randomization process, however, rarely occurs in practice and is often infeasible (Bloom, 2005; Rubin, 1974; Shadish, Cook, and Campbell, 2002). Notably, even when this process is feasible, random sampling allows unbiased estimation only for this one well-defined population, and does not solve the problem of generalizing to a new or different population (Cronbach's *U**). While we will assume throughout that random sampling is infeasible, we focus here on the case in which random assignment within the experiment is possible. This leads to the problem of sample selection bias.

In order to examine the problem of (and provide a solution to) sample selection bias more carefully, information on the population and experimental units is required. Therefore a well-defined population frame consists of either a census or probability survey of the population of interest, with information on a variety of important measures on the population units (*P*). Similar information must also be available for units in the experiment (*S*). There are two types of potential relationships between the experiment and population data sets:

(1) the sample is a *subset* of the population data set ($S \subset P$); or

(2) the sample and population are combined into one *stacked* data set ($S \cup P$).

The first case occurs when the population frame is a state administrative data system, the experiment was conducted in the same state, and the experimental units are locatable in the state administrative system. The second case occurs when the population data set is a sample survey, or the experiment was not conducted in the population of interest. While both of these relationships can be addressed by the method we develop here, the type of relationship will have implications for estimators and their properties. It should be noted that in the *subset* case, the full data set will be of size N, whereas in the *stacked* case, the data set will be of size $N + n$.

The problems of generalization and sample selection bias can be situated in the potential outcomes framework. Let $W=1$ if a unit is assigned to the treatment. Then assume that for each unit in the population there exist two potential outcomes, $Y(0) = Y(W=0)$ and $Y(1) = Y(W=1)$, where $Y(1)$ is the unit's potential outcome under treatment and $Y(0)$ is the unit's potential outcome under some specified alternative condition. These are referred to

as potential outcomes since they are theoretical quantities. The Fundamental Problem of Causal Inference arises because both outcomes can never be observed for a particular unit (Holland, 1986). Instead, at most $Y = W\,Y(1) + (1 - W)\,Y(0)$ is observed, which for units in the experiment is either $Y(1)$ if the unit is assigned to the treatment or $Y(0)$ if the unit is not. Note that these Y's are only observed for units in the experiment. Additionally, for each unit in the population P and in the sample S, a vector of covariates $\boldsymbol{X}$ that are associated with the outcomes Y is also observed.

For each unit in the population and sample, we can define a theoretical unit treatment effect, $\Delta = Y(1) - Y(0)$. While this unit effect cannot be observed, average treatment effects (ATE) can be estimated. First, let the random variable $Z = 1$ if a unit is in the experiment. Note that in the *subset* case the unit is also in the population, whereas in *stacked* case, the unit is not. We can define the following estimands:

(1) (Population) PATE = $\tau^P = E[Y(1) - Y(0)]$ in the subset case, or

$= E[Y(1) - Y(0)|\, Z=0]$ in the stacked case; and

(2) (Sample) SATE = $\tau^S = E[Y(1) - Y(0) \mid Z=1]$ in both the stacked and subset cases.

When the experimental sample is not randomly selected from the population, there is no reason to believe that the SATE and PATE will be identical, except in the case of constant treatment effects (Imai, King, Stuart, 2008; Rubin, 1974). This is counter to the observational studies case in which the data collected is either itself considered the full population or is assumed to be a random sample from the population of interest (Imai, King, Stuart, 2008; Imbens, 2004).

Finally, note that the conventional estimator of the treatment effect is the simple difference in means,

$$T = E[Y \mid Z=1, W=1] - E[Y \mid Z=1, W=0] = \bar{Y}_T - \bar{Y}_C$$

where $\bar{Y}_T$ is the mean outcome in the treatment group ($W=1$) and $\bar{Y}_C$ is the mean outcome in the control group ($W=0$) in the experiment. Importantly, this is an unbiased estimator of the SATE. However, this estimator (and more complicated versions in cluster randomized or multi-site trials) is also commonly used to estimate the PATE. In the remainder of this paper, we will refer to this as the conventional or naïve estimator of the PATE. The focus here will be on developing a new estimator and comparing it to T in terms of changes in bias and sampling variance.

A. Assumptions needed for statistical generalization

In order to obtain an unbiased estimate of the PATE using data from an experimental sample S, the following three assumptions will be needed. Importantly, these same assumptions would be needed even in the case in which random sampling from the population of interest was employed.

(A1) *Stable Unit Treatment Value Assumptions:*

SUTVA (Rubin 1978, 1990) must be met for all units in the experiment as well as all units in the population of interest. That is, the effect of a treatment for units in both the population and sample must be *stable* in the following senses:

(a) SUTVA (S): The effect of the treatment on a unit i does not depend on the

treatment assignment of a unit *j* within the *sample*.

(b) SUTVA (P): The effect of the treatment on a unit *i* does not depend on the treatment status of a unit *j* in the *population*.

There are cases in which SUTVA holds in either *S* or *P* but not both. For example, SUTVA (P) would be violated if a treatment inherently had multiple versions but the sample was drawn so as to meet SUTVA (S). One particular example is the problem of peer effects. If peer effects for a particular intervention only occur in mixed-ability classrooms, the population contains both mixed- and segregated- ability classrooms, and the sample is chosen to contain only classrooms that are segregated by ability, then SUTVA (S) will be met but SUTVA (P) will not. This is because SUTVA is about the stability of the potential outcomes; in order to generalize, this stability is required both within the population of interest and the sample used for estimation.

(A2) *Strongly ignorable treatment assignment*:

Let $Z=1$ if a unit is in the sample and, for units with $Z=1$, let W=1 if the unit is assigned to the treatment condition. Then the treatment assignment is strongly ignorable if

$$[Y(0),Y(1)] \perp W \mid Z = 1 \text{ and } 0 < Pr(W=1|Z=1) < 1.$$

This condition means that within the experiment, the probability of receiving either treatment condition must be non-zero for all units and the treatment assignment must not be confounded with the potential outcomes. Note that this condition is met by the random assignment process typically used in experiments, which is the focus here.

The fact that this condition is met unconditionally, however, does not imply that it is met conditionally. In randomized experiments every unit has a non-zero probability of receiving the treatment, yet for particular subgroups this may not be true. More formally,

$$Pr(W=1|Z=1) = \int Pr(W=1|Z=1,\mathbf{X}=\mathbf{x})\, dP_x = E_x\{Pr(W=1|Z=1,\mathbf{X}=\mathbf{x})\}.$$

While A2 requires $0<Pr(W=1|Z=1)<1$ it does not require that $0 < Pr(W=1|Z=1, \mathbf{X}=\mathbf{x}) < 1$; indeed it is likely that this condition will not be met for particular values of **X=x**. This fact will become important when conditional treatment effects are defined.

(A3) *Sampling ignorability:*

The sampling process and the unit treatment effects are conditionally independent given the covariate vector **X**,

$$\Delta = [Y(1) - Y(0)] \perp Z \mid \mathbf{X}.$$

In order to meet this condition **X** must include any covariates that both explain variation in the treatment effects and differ in distribution in the population and experiment. This is a smaller set of variables than those associated with the outcome, which is the requirement for matching in observational studies. For example, in observational studies local matches are encouraged since many contextual variables are related to educational outcomes (Cook, Shadish, &Wong, 2008); however, so long as the treatment effect does not vary in relation to these contextual variables, they need not be accounted for in the sample selection bias adjustment case addressed here. Indeed, if the effect of a treatment is contingent on local, contextual variables, then generalizations outside the experimental sample are impossible without a great deal of modeling.

Finally, when A1 to A3 have been met, the sample selection process is referred to as *weakly ignorable* given the covariate vector $\boldsymbol{X}$. These three assumptions make clear the role that covariate information in $\boldsymbol{X}$ in both the experiment and population play in the process of adjusting for non-random sampling bias. When $\boldsymbol{X}$ is of high dimension, however, matching the experimental units to population units can lead to many cases with no matches. In observational studies, this multivariate matching problem has been addressed through the propensity score which models the probability that a unit receives the treatment (Rosenbaum & Rubin, 1983). Propensity scores are also used in survey sampling to model the probability that a unit does not respond to a survey (Little, 1986). Here we will use the propensity score to model and adjust for the probability that a unit is in the experiment.

Definition 1.1.1: Sampling propensity score

Let $Z=1$ if a unit is in the experimental sample. Then the *sampling propensity score* is defined as $e(\mathbf{X}) = Pr(Z=1|\mathbf{X})$,

where we assume $Pr(Z_1,\ldots,Z_N|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N) = \Pi_1^N e(\boldsymbol{X})^Z (1 - e(\boldsymbol{X}))^{1-Z}$.

Propensity scores have been shown to have several important properties that make them useful in practice. The next proposition highlights an important property in the sample selection case.

Proposition 1.1.1 Balancing score property

The sampling propensity score is a balancing score: $\boldsymbol{X} \perp Z \mid e(\boldsymbol{X})$.

Proof:

By extension of Rosenbaum and Rubin (1983) Theorem 1.

**

Note that a balancing score is a function of ***X*** such that the conditional distribution of ***X*** given the balancing score is the same for cases in which *Z =1* and *Z=0*, or here, the sample and population. Proposition 1.1.1 implies that two units with the same value of *e(**X**)* will on average have the same values on all covariates in ***X***. This is especially important in the generalization situation, since in the *stacked* case interpreting the propensity score *e(**X**)* as a meaningful probability will not be warranted. For example, if a PATE is of interest for Florida but the experiment was conducted in Texas, then in truth, all units in Florida had zero probability of being in the experiment. In this case, the propensity score *e(**X**)* is in no way an approximation to a meaningful probability. However, Proposition 1.1.1 implies that regardless, the propensity score can be useful as a balancing score – it provides a method for matching units in the population of interest to their most relevant comparison cases in the experiment.

Based on these assumptions, conditional treatment effects and their relationship to the PATE and SATE can be defined.

Proposition 1.1.2: Conditional treatment effects

For each value of the propensity score *e(**X**)*, let

$$\tau(e) = E[Y(1) - Y(0) \mid e(\mathbf{X}) = e]$$

be the conditional treatment effect. It can be shown that this

$= E[Y(1) - Y(0) \mid e(\boldsymbol{X}) = e, Z = 1]$

$= E[Y(1) \mid e, Z=1, W=1] - E[Y(0) \mid e, Z=1, W = 0]$

$= E[Y|W=1, e, Z=1] - E[Y|W=0, e, Z=1].$

Proof.

The first equality is definitional, the second falls from A1 and A3, and the third from A2. The fourth equality shows that when $\tau(e)$ is defined, it can be estimated from the observed data.

***

Importantly, $\tau(e)$ may not be defined for all values of $e(\boldsymbol{X})$. This occurs for two reasons. First, it may be the case that there exist population cases with a particular value of $e(\boldsymbol{X})$ but there do not exist any experimental cases with the same value. Second, since A2 only provides unconditional balance between the treatment conditions in the sample, there may exist values of $e(\boldsymbol{X})$ for which there are only treatment or control cases. However, on average, $\tau(e)$ will be defined for all values of $e(\boldsymbol{X})$. From these values, an estimator of the PATE can be developed.

Proposition 1.1.3 PATE in relation to conditional treatment effects

Assume A1 to A3 and, additionally, that for every value of $e(\boldsymbol{X})$ in population P, there exists a unit in the sample S with the same value. Let $E_{e,P}[.]$ be the expectation of $[.]$ over the distribution of $e(\boldsymbol{X})$ in the population P. Then the conditional treatment effects can be related to the PATE by,

$\tau^P = E_{e,P}[\tau(e)]$

$$= E_{e,P}[\tau(e) \mid Z=1]$$

$$= E_{e,P}[Y(1) \mid W=1, e, Z=1] - E_{e,P}[Y(0) \mid W=0, e, Z=1]$$

$$= E_{e,P}[Y \mid W=1, e, Z=1] - E_{e,P}[Y \mid W=0, e, Z=1].$$

Proof:

The first equality is definitional, and the second equality holds by assumption. The third equality says that when $\tau(e)$ is not defined for units in the experiment ($Z=1$), the expectations can be taken separately over the treatment ($W=1$) and control ($W=0$) units. This is a direct result of A2, which implies that the missing units can be treated as missing at random (MAR) (Little & Rubin, 1989). The fourth equality states that these can be estimated from the observed data.

***

Note that under the conditions found in Proposition 1.1.3, an unbiased estimator of the PATE can be defined. In practice – particularly when the sample size n and sampling fraction n/N are small – these conditions will not be met. In particular, it is likely that there will exist values of $e(\boldsymbol{X})$ in the population that do not have relevant comparison cases in the experiment. As a result, this situation is well-suited for a subclassification estimator, which we introduce in the next section.

II. Subclassification estimator of the PATE

The subclassification (or stratification) estimator was introduced by Cochran (1968) for the single variable case and extended to the treatment propensity score case by

Rosenbaum and Rubin (1984). It has been suggested for use in the sample selection case by both Hedges and O'Muircheartaigh (in press) and Stuart et al (in press).

Definition 1.2.1: General subclassification estimator of PATE

Assume that there are *n* units in the experimental sample and *N* units in the population. Let

$$T_S = \sum_{j=1}^{k} w_{p_j} T_{S_j} = \sum_{j=1}^{k} w_{p_j} \left(\bar{Y}_{T_j} - \bar{Y}_{C_j} \right)$$

where for stratum *j*, $\bar{Y}_{T_j}$ is the sample mean for units in the treatment group (*W=1*), $\bar{Y}_{C_j}$ is the sample mean for units in the control group (*W=0*), N_j is the number of population cases and $w_{pj} = N_j/\mathrm{N}$ is the proportion of population cases in the stratum, and where the *k* strata are defined by the distribution of ***e(X)*** in the population *P*.

The remainder of this section will develop conditions under which this estimator can be compared with the conventional estimator, *T*, in terms of average bias and average variance. Later these simplifications will be used to investigate the amount of reduction in average bias and increase in average sampling variance that can be expected under particular numbers of strata with the types of distributions of ***e(X)*** that can be expected in the generalization case. Following Cochran (1968), we limit our focus here to cases in which the *k* strata are defined such that $w_{pj} = 1/k$ for strata $j=1 \ldots k$.

A. Reduction in average bias

The most well known result from Cochran's (1968) study is that using a subclassification estimator with five equal-weight strata reduces the bias (as compared to a conventional estimator) by approximately 90%. In this sections, we extend these results to the sample selection bias case. Following Cochran (1968) and Rosenbaum and Rubin (1984) we make the following simplifying assumption.

(A4) *Monotonic conditional treatment effects:*

> Assume that $\tau(e) = E[Y(1) - Y(0)|e]$, the treatment effect conditional on $e(\boldsymbol{X})$, is a monotone function of $e(\boldsymbol{X})$. That is, as the probability of being in the experiment increases, the conditional treatment effect distribution either increases or decreases monotonically.

Note that when the sample is a *subset* of the population, in many cases it will be reasonable to believe that A4 will hold. For example, units that are more likely to be recruited for or select into an experiment may have larger treatment effects. However, in the *stacked* case, it is less likely that A4 will hold. In this case, the relationship between $e(\boldsymbol{X})$ and $\tau(e)$ may be highly variable and in practice, a higher degree of matching will be required. We focus here on the case in which A4 holds in order to relate average bias reduction in $\tau(e)$ to average bias reduction in terms of $e(\boldsymbol{X})$.

Proposition 1.2.1: Expected bias reduction via subclassification in generalization

Let $\tau(e) = E[Y(1) - Y(0)| e]$ be the conditional treatment effect and $\tau^P = E_{e,P}[\tau(e)]$ be the PATE. The average bias in the conventional estimator T can be written,

$$EB_I = E_{e,S}[\tau(e)|Z=1] - E_{e,P}[\tau(e)] = \tau^S - \tau^P,$$

Where the subscript e,S indicates that the expectation is over the distribution of $e(\boldsymbol{X})$ in the sample S. This is the difference between the SATE and the PATE. In comparison, the average bias in the subclassification estimator T_S can be written

$$EB_S = \Sigma \{E_{e,S}[\tau(e)|Z =1, e\in I_j] - E_{e,P}[\tau(e)| e\in I_j]\} Pr_P(e\in I_j).$$

$$= \Sigma E[\tau(e)|Z =1, e\in I_j]w_{pj} - E_{e,P}[\tau(e)] ,$$

where I_j is the set of $e(\boldsymbol{X})$ values in population P in stratum j. Assume A4. Then the proportion reduction in average bias in the $\tau(e)$ scale following subclassification on $e(\boldsymbol{X})$ equals the reduction in average bias in the $e(\boldsymbol{X})$ scale, where

$$EBR = 100(1 - EB_S/EB_I)\%$$

Proof.

Follows immediately from Rosenbaum and Rubin (1984) Theorem A.1.

**

Proposition 1.2.1 provides a method to conceptualize average bias reduction in relative terms. It says that under monotonicity, we can approximate the amount of average bias reduction by looking only at the average bias in terms of the distributions of $e(\boldsymbol{X})$ in the population and experiment. In other, non-monotonic cases, average bias will also be reduced by subclassification, but the amount of bias reduction in these cases will be more difficult to quantify, since it will depend on the functional form of $\tau(e)$.

B. Effect on sampling variance

In terms of sampling variance, Cochran shows that in observational studies, sampling variance is often reduced by use of a subclassification estimator. In order to investigate average sampling variance in the generalization case, we focus here on the special case in which the relationship is not just monotone but actually linear.

Proposition 1.2.2: Expected variance inflation via subclassification in generalization

Assume that for those units in the sample, $Y(W=0) = \mu_C + \beta_C e + \varepsilon$ and $Y(W=1) = \mu_T + \beta_T e + \varepsilon$, where $e=e(\mathbf{X})$, $E(\varepsilon|W=0)=E(\varepsilon|W=1)=0$, and $V(\varepsilon|W=0)=V(\varepsilon|W=1)=\sigma^2$, and where $E(.)$ is the expectation and $V(.)$ is the variance. That is, assume $\tau(e) = \delta + \beta e$, where $\delta = \mu_T - \mu_C$ and $\beta = \beta_T - \beta_C$. Let $\rho_{et} = Corr(e,Y|W=1)$ and $\rho_{ec} = Corr(e,Y|W=0)$ where $Corr(.)$ is the correlation. Then under A4, the average variance inflation of the subclassification estimator (relative to the conventional estimator, T) can be written

$$EVIF(T_S) = A[1 - \rho_{e*}^2 (1 - B/A)]$$

where for stratum j, w_{pj} is the proportion of the population P, w_{sj} is the proportion of the sample S, and σ_{ej}^2 is the variation in the distribution of $e(\mathbf{X})$ in S, and where σ_e^2 is the total variation of $e(\mathbf{X})$ in the sample, $A = \sum_{j=1}^{k} w_{pj}\left(\frac{w_{pj}}{w_{sj}}\right)$, $B = \sum_{j=1}^{k} w_{pj}\left(\frac{w_{pj}}{w_{sj}}\right)\left(\frac{\sigma_{ej}^2}{\sigma_e^2}\right)$, and $\rho_{e*}^2=(\rho_{et}^2+\rho_{ec}^2)/2$ is the average value of the correlation.

Proof.

See Appendix A.

***

In order to get intuition about the EVIF, focus on the case that $\rho_{e^*}^2 = 0$, indicating that there is no relationship between e and $\tau(e)$. In this case, the EVIF reduces to A, which is a rough measure of the similarity of the distributions of e in the population and experiment. Clearly $A>1$ when these distributions differ and A is very large when there exists at least one stratum j' such that $w_{sj'}$ is very small.

C. Theoretical investigation and rules of thumb

An important question is in practice how the subclassification estimator will compare to the naïve estimator. The amount of EBR and EVIF for a subclassification estimator will depend on the distributions of $e(\boldsymbol{X})$ in the population and the experimental sample. Cochran (1968) investigates this question (in an observational studies framework) for a set of distributions that largely consist of locational shift models, and these results have been extended to the treatment propensity score case as a result of Rosenbaum and Rubin (1984). In the remainder of this section we focus on the distribution of the propensity score logits, $f = f(e) = \log[e/(1-e)]$, which take values on the whole real line instead of the interval (0,1).

Empirical observation shows that in experimental generalization the distributions of f will often differ in the experiment and population. In many cases, the distribution in the population is skewed, whereas that in the experiment is symmetric. One reason for this is

that the population often contains units with very small probabilities of being in the sample. These cases will be less common in the sample, and as result its distribution will be more symmetric. Second, the distributions of many population variables are highly skewed themselves (e.g. income), whereas current methods for selecting samples tends to focus on the inclusion of average and modal units, not those at the full range of covariate values (e.g. Bloom, 2005; Cook, 1993). Additionally, it is tautological to assume that f will have the same distribution in the two groups, since it amounts to assuming that the population and experiment are *a priori* similar. Finally, the clearest method for making the distributions more similar – dropping cases – here comes at a large cost, since it requires dropping population cases. In Section III we discuss this strategy; here we focus on the case in which no population cases can be dropped.

The goal of this analysis is to compare the subclassification estimator, T_S, with the conventional estimator, T, in terms of average bias and variance; here T_S has $k=2,3,4,$ and 5 equally sized strata based on the population density of f. For comparison purposes, two cases are investigated. The first case – the "locational shift" case— is an extension of the normal locational shift model studied by Cochran; in this model both the experiment and population will be assumed to follow the normal distribution, though the two will differ in means and variances. The second case—the "generalization" case – is the case we argue is more likely to occur in generalization. Here the experiment follows a normal distribution, while the population distribution of f is skewed (a chi-squared distribution). For both sets of cases, we evaluate these measures using a variety of parameter combinations, the derivations for the average bias and average variance developed above, and a set of

functions in the statistical program **R** that evaluates the mean and variance of truncated distributions (Nadarajah and Kotz, 2006).

Table 1.1 about here

As Table 1.1 shows, for the normal locational shift case studied by Cochran, $k = 2,3,4,5$ strata correspond approximately to EBR values of 65-, 80-, 85-, and 90-percent. When the population is skewed and the sample is not, however, these EBR values are much smaller; for $k = 2,3,4,5$ strata, the EBR's are in the ranges of (20,40), (30,50), (40,60), and (50,70) percent respectively. The EBR's are the smallest in cases in which the variance in the sample is small relative to the population skew. It should be noted that additional distributional results for other parameter values not reported here were similar in all of these cases. While these bias reductions are smaller than those suggested in the observational studies case, it is important to note that they are still substantial. Any reduction in bias is better than no reduction.

Table 1.1 also reports the effect of subclassification on average sampling variance (EVIF). Note that here again the number of strata k and ρ_{e*}^2 are both varied. In general, note that as ρ_{e*}^2 increases, the EVIF's decrease; in some cases, the EVIF's diminish to less than one. For simplicity, focus only on the value $\rho_{e*}^2 = 0$. Importantly, the results presented here for the normal shift model studied by Cochran are new, since the sampling variance function here is different than in the observational studies case. When the two distributions have the same variance, the EVIF tends to be in the interval (1,2). When the

sample distribution is less variable, however, the EVIF is larger and can be greater than 5. In all cases, as the number of strata k increases, the EVIF increases, but more rapidly in the unequal variance case. In the generalization cases studied here, the EVIF's are all greater than 1 and they tend to increase rapidly as a function of k. For $k = 5$, these values can be quite large, particularly when there is much smaller variance in the experiment relative to the population skew. This result occurs largely because of the last stratum, where there is a large population weight relative to the number of experimental cases available for matching.

Overall, it is notable that whereas Cochran found (for observational studies) that the subclassification estimator often reduced both variance and bias, in the generalization case this table shows that in this case the average variance usually increases. Taken with the differences in EBR, this means that in many cases the best strategy may be to use smaller values of k and settle for a slightly smaller amount of EBR in order to reduce costs in terms of EVIF.

III. Defining a sub-population like the experiment

Up until this point we have mentioned but not thoroughly addressed the case in which there exist population units with no relevant comparison units in the sample. This situation occurs when the sample size n and the sampling ratio n/N are small, and results from the fact that the range of covariate values in $\boldsymbol{X}$ in the sample is smaller than that in the population. When this occurs, average bias reduction using a subclassification estimator is

limited, and average variance inflation can be quite large. In this section we address this problem more formally.

In survey sampling this problem is referred to as *coverage error* (Groves, Fowler, Couper, Lepkowski & Singer, 2009). In sampling, coverage error occurs when the population of interest (the target population) and the sampling frame are not identical. For example, if the population of American households is of interest and sampling is conducted using random digit dialing on land lines, then the sampling frame excludes households in the population without land lines. This means that the average for the target population and the sampling frame are not identical, inducing *coverage bias*. A solution to this problem here is to restrict generalizations to the sub-population with relevant comparison cases in the sample. To do so, we introduce the an additional assumption.

(A5) *Sampling overlap:*

The conditional probability of being in the sample is non-zero for all units in the population: $0 < Pr(Z=1 \mid \boldsymbol{X}=\boldsymbol{x})$.

Note that this assumption does not additionally require $Pr(Z=1|\boldsymbol{X}=\boldsymbol{x}) < 1$. The reason for this is that with sampling fractions $n/N << 1$, cases in which $Pr(Z=1|\boldsymbol{X}=x) = 1$ will only arise for units in the sample. While including or excluding these units can effect average bias and variance, they do not effect the ability to find matches for the population units, which is the purpose of this assumption. Finally, note that with the addition of this assumption we move from a *weakly ignorable* sample selection process to one that is *strongly ignorable*. This is clarified in the next proposition.

Proposition 1.3.1 Strongly ignorable sample selection

If A1 to A3 and A5 have been met for a covariate vector $\boldsymbol{X}$, we can say that the sample selection is *strongly ignorable* given the propensity score $e(\boldsymbol{X})$, i.e.

$$[Y(1) - Y(0)] \perp Z \mid e(\boldsymbol{X}) \text{ and } 0 < Pr(Z = 1 \mid e(\boldsymbol{X})).$$

Proof:

By extension of Rosenbaum and Rubin (1983) Theorem 3.

**

In the remainder of this section, we formalize the process of choosing a sub-population P_0 from a given population P such that sample selection is *strongly ignorable*. The next definition formalizes this.

Definition 1.3.1 Sub-population arising from strong ignorability

Let $e(\boldsymbol{X})$ take values in [0,1]. Assume $e(\boldsymbol{X})$ in the population P follows a distribution with distribution function $G_p(e)$. Let there exist truncation values a and b such that $0 \leq a < b \leq 1$ and $Pr(a < e(\boldsymbol{X}) < b \mid S) = 1$, implying A5 holds. Then the following definitions apply:

(1) The *population coverage rate*, θ, is defined as $\theta = G_p(b) - G_p(a)$,

which is the proportion of the population P that is in the sub-population P_0;

(2) The distribution function for the sub-population P_0 is $G_{p0}(e)$, where

$G_{p0}(e) = G_p(e)/\theta$; and

(3) The sub-population average treatment effect, P_0ATE, can be written

$\tau^{P0} = E_{e,p0}[\tau(e)]$,

where the subscript e,p_0 signifies that the expectation is over the distribution of $e(\boldsymbol{X})$ in the sub-population P_0.

**

One benefit of this approach is that by re-defining the population, the average bias of the original estimator T can be significantly reduced, while at the same time not impacting the variance. That is, while T is biased for τ^P, T may be less biased for a sub-population ATE, τ^{P0}. The following proposition formalizes the average bias reduction.

Proposition 1.3.2: Proportion of average bias removed by truncation

Let $\tau(e) = E[Y(1) - Y(0)| e]$ be a conditional treatment effect for a particular value of $e=e(\boldsymbol{X})$. Let $Z=1$ if a unit is in the sample and let $Q=1$ if a unit is in the sub-population P_0. Then the proportion of average bias removed by changing the estimand from τ^P to τ^{P0} is

$$EBR = 100(1 - EB_{P0}/EB_I)\%$$

where $EB_I = E_{e,S}[\tau(e)|Z=1] - E_{e,P}[\tau(e)] = \tau^S - \tau^P$, and $EB_{P0} = E_{e,S}[\tau(e)|Z=1] - E_{e,P0}[\tau(e)] = \tau^S - \tau^{P0}$. Assume A4 holds. Then the proportion reduction in average bias in the $\tau(e)$ scale following subclassification on $e(\boldsymbol{X})$ equals the reduction in average bias in the $e(\boldsymbol{X})$ scale.

Proof.

Follows immediately from Rosenbaum and Rubin (1984) Theorem A.1.

**

The method provided here improves average bias and variance by changing the population of interest. Importantly, we do not advocate dropping experimental units from the analysis. While there may certainly be pay-offs in terms of *EBR*, it can be shown that the *EVIF* will be larger for an estimator with fewer experimental cases (but better population and experimental distributional balance) than for the conventional estimator. Additionally, note that for the new population P_0, a subclassification estimator can also be applied to remove any residual bias. For simplicity, we will refer to this as the conditional subclassification estimator, since it is conditional on P_0 (in contrast to the unconditional estimator being for P).

In order to examine the amount of average bias reduction theoretically, the same distributions are used as those found in Table 1.1. Since this analysis is at the level of statistical distributions instead of observed samples, we let $b = F^{-1}(.99)$, where F is the cumulative distribution function of the distribution of $e(\boldsymbol{X})$ in the sample. In an actual observed experiment and population, the value of b would be chosen so that the two observed distributions share common support.

Table 1.2 summarizes these new results; it reports the value of θ and the percent reduction in average bias by moving from P to P_0. In cases in which the remaining average bias is non-negligible, Table 1.2 also reports the amount of additional average bias that can be removed by using a subclassification estimator. Note that this additional average bias reduction is in comparison to the average bias of T for τ^{P0}, not for τ^{P}.

Table 1.2 about here

Table 1.2 shows that there is much to be gained in terms of both average bias and variance by changing the population of interest. In some cases changing the population removes nearly all of the average bias, whereas in other cases it can reduce the bias significantly, without any costs in terms of variance (since the original estimator T is used). The θ values give a sense of what cost this change in estimand may lead to in terms of interpretation. In most cases the coverage rate $\theta > .5$, and even in cases with the largest bias problems, θ is often in (.70,.90). For those distributions with remaining bias, Table 1.2 also shows that the bias can be reduced further by applying a subclassification estimator. Importantly, in many cases the EBR values are now closer to those found in Cochran; this is not surprising, since by truncating the population distribution the results should be closer to the types of distributions studied there. For example, here we find that even in the generalization cases, using $k=5$ strata frequently leads to the 90% bias reduction reported by Cochran.

In terms of EVIF, there are also gains to be made by changing populations. First note that in cases in which a subclassification estimator is not needed, there will be no cost in terms of variance. If an additional subclassification estimator is used, when ρ_{e*}^2 is zero (the most conservative case), most of these EVIF's are greater than one. However, in many cases they are significantly smaller than those for the original population. Notably, the results given here are for a population truncation value that is somewhat conservative in

the sense that it keeps θ reasonably large. In actual application, choosing a different value (and accepting a smaller θ value) can lead to even greater improvements.

Finally, it should be noted that the main drawback to moving from P to P_0 is that this new sub-population may not be as easily conceptualized. This problem can be treated as a classification problem and approached via enumeration (i.e. creating a list of units in P_0), geographically (e.g. mapping units included in P_0 versus P/P_0), descriptively (e.g. reporting univariate statistical comparisons), or via a statistical classification framework (e.g. Johnson & Wichern, 2002). Importantly, note that this last case often leads to the definition of a new propensity score – the propensity for a unit in P to also be in P_0.

IV. Example

SimCalc is a mathematics software program aimed at middle-school students which teaches proportionality, linearity, and rates of change. In the 2008-2009 academic year, SRI conducted a cluster-randomized efficacy trial of SimCalc in 78 schools across Texas (Roschelle et al, 2010). Every effort was made to include schools representative of the state, but random sampling was not feasible or implemented. In this section, we present a re-analysis of the SimCalc experiment aimed at getting a less biased estimate of the school average treatment effect for the population of schools in Texas. The analysis here focuses only on the 78 schools in the experiment; we exclude the pilot study for illustrative purposes.

This method requires the population of interest to be well defined; in this case, data from the publically available state academic excellence indicator system (AEIS) were used.

The population was defined to be Texas schools in the 2008-2009 academic year with 7th grade classrooms that were not charter schools. Note that since the schools in the study were easily locatable in the state AEIS system, this is an example of the *subset* case. Based on the covariates available, 26 potential school-average treatment effect moderators were selected for matching, including student and teacher demographic composition, school structure, and prior year school average tests scores. Note that by matching at the school level, we assume that the students in the experimental schools that took part in the experiment were representative of the 7^{th} grade student body at that particular school.

Table 1.3 below includes a list of these covariates (columns 1-3) as well as their standardized (SMD) and unstandardized mean differences (see columns under "(1) Conventional" in the table); note that for standardization the population standard deviation is used. Notably 10 of the 26 covariates have absolute SMD's greater than 0.20, which are large. For example, in population schools, as compared to experimental schools, a larger proportion of both students and teachers are African American, while a smaller proportion are Hispanic; teachers are more experienced; a larger proportions of students are mobile; and the total number of students is smaller.

Table 1.3 about here

The outcome of interest for determining efficacy here is the school average of student gain scores on a sub-scale of a math test. This test is a proximal measure focusing directly on the issues of proportionality, linearity, and rates of change. The conventional

estimate of the treatment effect is $T = 3.28\ (.35)$, $p<.001$, indicating that students in schools with SimCalc had larger gain scores. In total, we report here on three analyses. First, a subclassification estimator is used to estimate the PATE for the full population of schools in Texas. Second, population cases not meeting the strong ignorability assumption are dropped and a new sub-population and P_0ATE are defined. Third, for this subpopulation, a subclassification estimator aimed at further reducing bias is employed for estimating the P_0ATE. In all cases changes in average bias for each covariate are assessed in relation to both absolute standardized mean differences and percent reductions in absolute mean differences; these are also presented in Table 1.3 under the column headers (2), (3), and (4). These analyses were conducted in the statistical program **R** using the **matchit** and **zelig** functions (Ho, Imai, King, & Stuart, 2007; Imai, King, & Lau, 2006).

The propensity scores were estimated using a logistic regression model with the 26 covariates. Figure 1.1 illustrates the distributions of the estimated logits in the experiment and population. As Table 1.3 shows, the distributions of the logits are 0.642 standard deviations apart, which is a large difference. For each of the three analyses, these distributions are used for determining the strata in the subclassification estimators.

Figure 1.1 about here

In estimating the PATE, a subclassification estimator with $k=3$ strata was used, where the strata were defined so that each stratum contained 1/3 of the population. Additional strata were not used since doing so resulted in strata either without any

experimental units (the *k=5* case) or with too few to estimate within-stratum standard errors (the *k=4* case). Recall that based on the theoretical results in Section II, we would expect that under monotonicity an estimator with *k=3* strata would remove approximately 40-60% of the average bias in the conventional estimator. In Table 1.3 the absolute standardized mean difference and the % reduction in average bias under the subclassification estimator for each covariate are provided under the column header "(2) Subclass (3 strata)". Average bias reduction in the logits is estimated to be 50% since the absolute SMD moves from 0.642 to 0.324. The subclassification estimate is $T_S = 3.45$ *(0.40), p<.01*. Note that the estimate is 5% larger and the standard error is 14% larger.

Figure 1.1 also shows that 22% of the population units do not have experimental units available for matching ($\theta = 0.78$); these 22% of units have zero-probability of being in the experiment. For the second analysis, we exclude these cases and define a new sub-population and treatment effect, the P_0ATE. Using the same conventional estimator, *T*, but estimating the P_0ATE instead of the PATE dramatically reduces bias; details of this can be found in Table 1.3 under the column header "(3) Conventional". Average bias in the logits is estimated to be 63% smaller when focusing on this subpopulation, and the absolute standardized mean difference moves from 0.642 to 0.234, which is nearly one third the size. Since the population has changed, not the estimator, the estimated treatment effect and standard errors do not change.

Finally, as Table 1.3 shows, there is still some average bias remaining even after shifting focus to the sub-population more like the experiment (the P_0ATE). For this sub-population, a subclassification estimator with *k=3* strata was used. Here using a larger

number of strata (*k=4,5*) resulted in some strata containing only treatment or control cases, making stratum specific treatment effect estimation impossible. Recall that the theoretical results in Section III show that the subclassification estimator can dramatically reduce bias in the sub-population case. In the logit scale, the final column header in Table 1.3 ("(4) Subclass (3 strata)") indicates that the subclassification estimator is 94% less biased than the conventional estimator for the P_0ATE and the absolute standardized mean difference is reduced from 0.234 to 0.044. Furthermore, in terms of specific covariates, only 3 of the 26 have absolute standardized mean differences greater than 0.10. Here the subclassification estimate is $T_S = 3.70\ (0.47),\ p<.01$. Note that the estimate is 13% larger, while the standard error is 32% larger.

To conclude, these analyses illustrate three different ways that subclassification on the propensity score can improve generalizations from an experiment to a population. The first analysis shows that this method can be used to find an estimator of the PATE which has smaller average bias; note that frequently some bias remains. The second analysis shows that this method can help identify a sub-population like the experimental sample. The complement of this sub-population includes schools without any sample units like them. This helps researchers determine which schools to target in future studies. Finally, the third analysis shows that once we have defined a sub-population like the experiment, further subclassification on the propensity score can produce a nearly conditionally-unbiased estimate of the P_0ATE.

V. Conclusions and guidance for applied researchers

The method and theory presented here are aimed at formalizing the process of generalization from an experiment to a population using accepted statistical methods. One of its virtues is that it allows the assumptions necessary for generalization to be clearly defined, and ensures that the experiment and population can be balanced on a large number of covariates. In addition to providing an estimator of the treatment effect that has less average bias than the standard estimator, we have shown that under certain conditions there are only small costs in terms of variance for using this method. Furthermore, it allows the identification of coverage area problems, which can help researchers clearly define sub-populations for future experimentation.

Finally, note that this method estimates the error involved in generalizing instead of insisting that generalization be a dichotomous choice (generalize versus not). With this method, standard errors take into account both the sampling and treatment assignment processes. A criticism of this approach is that it requires some degree of modeling since covariates must be chosen for matching. When non-random sample selection is used and treatment effects vary, however, using the conventional estimator is akin to assuming that the experiment really is a random sample from the population, even when evidence points to the contrary. Without random sampling, all generalizations require modeling; the virtue of this method is that it makes this model explicit instead of hidden.

CHAPTER 2

Standardized Measures of Heterogeneous Treatment Effects for Randomized Experiments

Large-scale social, educational, and medical experiments are generally designed to estimate the average treatment effect (ATE) of an intervention. When treatment effects are heterogeneous, however, the ATE is not adequate for summarizing or comparing treatments. For example, suppose that there exist two treatments A and B, where the effect size for treatment A is 0.15 and is constant for all units in the population, while the effect size for treatment B is 0.20, but varies considerably across units. For the average unit, clearly treatment B is preferable; however, the fact that the effect varies for treatment B but is constant for treatment A indicates that for some subset of the population, treatment A is actually preferable. In the heterogeneous treatment effects (HTE) case, as this example indicates, it is important to summarize the effectiveness of a treatment in a way that takes into account both the average effect and the degree of heterogeneity (Longford, 1999; Kravitz, Duan, and Braslow, 2004).

In the last twenty years, medical, educational, and social science researchers have become increasingly interested in detecting and modeling HTE in experiments. A common method for reporting HTE is via conditional or subgroup ATEs, though multivariate methods have been proposed (e.g. Kent, Rothwell, Ioanndis, Altman, and Hayward, 2010; Feller & Holmes, 2009). Recent work has shown that distributional methods, like quantile regression, often perform better than subgroup analyses, however, for detecting HTE (e.g. Bitler, Gelbach, Hoyes, 2007; Koenker, 2010). Other methods for HTE detection and modeling have included Bayesian regression trees (e.g. Imai & Strauss, 2010; Green & Kern, 2010) and instrumental variable models (e.g. Angrist, 2003). All of these approaches focus on modeling HTE in relation to unit attributes, including demographics and baseline

outcomes; these analyses are inherently dependent upon the set of covariates collected in the experiment.

A separate HTE literature focuses instead on unconditional analyses and on methods for summarizing the effect of an intervention when HTE occurs. The focus of this literature is on the detection of distributional differences between the outcomes in the treatment and control conditions. Central to this approach is the parameterization of the HTE via the measure σ_α^2, which is the variation in unit treatment effects. Estimators and test statistics for this parameter have been developed (Raudenbush & Bryk, 1987; Raudenbush, 1988; Herbert, Hayen, Macaskill, & Walter, 2011), though only for the case in which there is no relationship between the unit treatment effects and their baseline control outcomes. Some have argued that instead of focusing separately on the ATE, δ, and variability, σ_α^2, the parameter of interest should be their ratio, δ/σ_α (Longford, 1999; Kravitz, Duan, and Braslow, 2004), while others argue that the parameter truly of interest is the proportion of the population with negative or positive treatment effects (Heckman, Smith, and Clements, 1997; Gadbury and Iyer, 2000; Gadbury, Iyer, and Allison, 2001). This paper is situated in this second literature and focuses on defining summary parameters that take into account the ATE and the degree of heterogeneity and that are standardized for easy comparison across studies. Additionally, we provide estimators, large-sample distribution theory, and for some cases, small sample bias corrections. In order to evaluate the small-sample properties of these estimators, we provide simulation results for both bias and confidence interval coverage probabilities for both normal and non-normal data. We conclude with an example based on data from an educational experiment.

I. Model and Assumptions

Let each unit $i=1...2n$ in a population have two potential outcomes, $Y_i(0)$ and $Y_i(1)$, where $Y_i(0)$ is unit i's outcome under the control condition and $Y_i(1)$ is unit i's outcome under the treatment condition. Note that these can be further defined as,

$$Y_i(0) = \mu + \eta_i$$

$$Y_i(1) = \mu + \delta + \alpha_i + \eta_i$$

where μ is the average outcome under the control condition, η_i is a unit residual effect (that may be composed of both person-effects and random error), δ is the ATE, and α_i is the individual treatment effect (ITE) residual. From these, i's ITE can be written

$$\tau_i = Y_i(1) - Y_i(0) = \delta + \alpha_i.$$

Note that the η_i term cancels since it would occur under either potential outcome. Using this parameterization allows for each unit to have its own treatment effect. Furthermore, assume that: $E[\eta_i] = 0, V[\eta_i] = \sigma_c^2 < \infty$, $E[\alpha_i] = 0$, $V[\alpha_i] = \sigma_\alpha^2 < \infty$, and $Corr[\eta_i, \alpha_i] = \rho_{\alpha c} \in (-1,1)$. Then it can be shown that

$$E[Y_i(0)] = \mu, \qquad V[Y_i(0)] = \sigma_c^2 < \infty$$

$$E[Y_i(1)] = \mu + \delta \qquad V[Y_i(1)] = \sigma_t^2 = \sigma_c^2 + \sigma_\alpha^2 + 2\rho_{\alpha c}\sigma_c\sigma_\alpha < \infty$$

where δ is the ATE in the population, σ_c^2 is the variance in the control group, σ_t^2 is the total variance in the treatment group, and $\rho_{\alpha c}$ is the correlation between the control potential outcomes and the ITEs. These assumptions imply that

$$E[\tau_i] = \delta \qquad V[\tau_i] = \sigma_c^2 + \sigma_t^2 - 2\rho\sigma_c\sigma_t = \sigma_\alpha^2 < \infty.$$

which is to say that the effect of the treatment can be explained by the parameter vector (δ,

σ_α^2), where δ is the ATE and σ_α^2 is the variance of the ITEs. Note that here ρ = $Corr[Y_i(0), Y_i(1)]$ is the total correlation between the potential outcomes, which differs from $\rho_{\alpha c}$. Importantly, the above formalization does not require any assumptions regarding the distributions of the random quantities η_i or α_i or the potential outcomes $Y_i(0)$ and $Y_i(1)$.

The focus of this paper is on functions of the parameters δ and σ_α^2 which summarize the effectiveness of a treatment for a particular population. While δ can be estimated using observed data, σ_α^2 cannot. As a result of the Fundamental Problem of Causal Inference (Holland, 1986), both potential outcomes are never observed for the same unit *i*. Instead for each unit *i* the outcome $Y_i = Y_i(0)W_i + Y_i(1)(1 - W_i)$ is observed, where $W_i = 1$ if unit *i* is assigned to the treatment condition. The ATE, δ, can be estimated from the observed data using the identity

$$\delta = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1)|W_i=1] - E[Y_i(0)|W_i=0],$$

which leads to the sample estimator,

$$d = m(t) - m(c)$$

where $m(.)$ is the sample average for the respective group. However, estimating the ITE variance, σ_α^2, is more difficult since

$$\sigma_\alpha^2 = V[Y_i(1) - Y_i(0)] = V[Y_i(1)] + V[Y_i(0)] - 2\rho V[Y_i(1)]^{1/2} V[Y_i(0)]^{1/2}$$
$$= V[Y_i(1)|W_i=1] + V[Y_i(0)|W_i=0] - 2\rho V[Y_i(1)|W_i=1]^{1/2} V[Y_i(0)|W_i=0]^{1/2}$$

which leads to the estimator

$$s_\alpha^2 = s_c^2 + s_t^2 - 2\rho s_c s_t,$$

where $s_.^2 = \Sigma(Y_{.i} - m(.))^2/(n_. - 1)$ is the sample residual variance for each of the groups, and $\rho \in (-1,+1)$ is the true underlying potential outcome correlation.

Importantly, the correlation ρ cannot be identified or estimated from the observed data, making the estimation of σ_α^2 difficult. This problem has been approached in the literature in three ways. The simplest approach has been to assume that there is no relationship between a unit's baseline control outcome and its ITE; this amounts to assuming that $\sigma_\alpha^2 = \sigma_t^2 - \sigma_c^2$ (Raudenbush and Bryk, 1988; Herbert, Hayen, Macaskill and Walter, 2011). A second approach has been to bound ρ using observed data. For example, Gadbury and Iyer (2000) suggest using a single covariate that explains some of the HTE to develop tighter bounds, while Heckman, Smith, and Clements (1997) develop a method based on the correlation of the ranked treatment and control observed outcomes. The problem with this approach is two-fold. First, in many cases only a tighter lower bound, ρ_L can be found; the upper bound ρ_U is often no smaller than 0.99. Second, these bounds are themselves estimated, which induces sampling variance that must be taken into account in the estimation process. The third approach is to conduct a sensitivity analysis, providing conditional estimates of σ_α^2 for a variety of values of $\rho \in (-1,+1)$ (Gadbury, Iyer, and Allison, 2001). A problem with this approach is that the HTE variance varies considerably when ρ takes values in this whole interval, and it is hard to conceptualize in which situations a particular value of ρ may occur.

The method we propose here is an improvement on the sensitivity approach, and depends on making clear the relationship between $\rho_{\alpha c}$, the correlation between baseline control outcomes and ITEs, and ρ, the correlation between the baseline control outcomes and the treatment outcomes. These two correlations are related by

$$\rho_{\alpha c} = Corr(\eta_i, \alpha_i) = Corr[Y_i(0), Y_i(1) - Y_i(0)]$$

$$= [\rho\gamma - 1] / [1+\gamma^2 - 2\rho\gamma]^{1/2}.$$

In general, ρ_{ac} is a more intuitive parameter than ρ. When $\rho_{ac} \in (0,1)$, units with larger control outcomes receive larger ITEs; the effect of a treatment of this type is to further *disequalize*[1] the baseline control outcomes. When $\rho_{ac} \in (-1,0)$, however, units with larger control outcomes receive smaller ITEs; treatments of this type are *equalizing*, since they make the treatment outcomes more similar (as compared to the control outcomes). Importantly, in the cases in which $\gamma^2 = \sigma_t^2/\sigma_c^2 \geq 1$, the variance in the treatment group is at least as large as that in the control group, and the relationship between ρ_{ac} and ρ is monotonic for all values of γ^2. Figure 1.1 illustrates this relationship for a series of γ^2 values in the interval $(1,2)$. For the values of γ^2 shown here, treatments that are *disequalizing* take ρ values in $(.7,1)$, while those that are *mildly equalizing* take values in $(0,1)$, and those that are *strongly equalizing* take values in $(-1,.5)$. Note that these intervals overlap and are a function of both γ and ρ_{ac}.

Figure 2.1 about here

It is common practice in the analysis of experiments to assume that the variances in the control and treatment group are homogenous ($\gamma^2 = 1$). In this case, the relationship between ρ_{ac} and ρ can be written

---

[1] Bryk and Raudenbush (1988) also use the equalizing and disequalizing language. However, their analysis is not in terms of potential outcomes and is less precise, since they label all cases in which $\gamma > 1$ as equalizing and all cases in which $\gamma < 1$ as disequalizing. Here we focus only on the cases in which $\gamma \geq 1$ and further divide this case into equalizing and disequalizing based on the underlying potential outcomes correlation.

$$\rho_{\alpha c} = -[(1-\rho)/2]^{1/2},$$

which takes values in the interval *[-1, 0)*. This illustrates that even in the homogenous variances case, HTE can occur and that when it does, the treatment must be *equalizing*, since $\rho_{\alpha c} \leq 0$. If the equalization is mild, ρ will take values in *(.5,1)*, while if it is strong, it will take values in *(-1,.5)*.

In the remainder of this paper, we limit our focus to cases in which the treatment variance is at least as large as the control variance ($\gamma^2 \geq 1$). While ρ is unknown, the sensitivity approach we advocate here is to divide analyses into the *disequalizing* and *equalizing* cases. In many situations, researchers may have reason to believe that a treatment is more or less likely to be *disequalizing* than *equalizing*. Additionally, in cases in which this is difficult to discern, minimal values of HTE can be estimated by focusing on values of $\rho \approx 1$; if $\rho \neq 1$, these values will actually underestimate the amount of heterogeneity.

II. Parameterizing treatment effects under HTE

Under HTE, a treatment can be summarized by the vector of parameters, $(\delta, \sigma_\alpha^2)$, which can be estimated (for a particular value of ρ) by a vector (d, s_α^2). A difficulty with these two parameters, however, is that their interpretation depends on the scale of the underlying outcome measure. For this reason, in large-scale experiments, the measure of the ATE is generally converted into an effect size; for example, δ is converted into a standardized mean difference, $\delta_e = \delta/\sigma$, where σ^2 is the common variance of the outcomes. This parameter δ_e is then estimated using *Cohen's d* or *Hedges' g*. The benefit of the effect

size approach is that it is a scale-free quantity and can be compared and combined easily across studies (Cooper, Hedges, and Valentine, 2009).

The goal of this paper is to develop scale-free measures of HTE that can be easily interpreted and compared across studies. To this end, we provide the following two sets of parameters. We then discuss the interpretation of these parameters in the homogenous variances case, which is often made in the practice. These derivations illustrate that even in the case that $\sigma_t^2 = \sigma_c^2$, HTE can occur, and to what degree.

A. Control standardized mean difference and variation

Under HTE, the effectiveness of a treatment can be summarized by the vector (δ, σ_α^2), where δ is the ATE and σ_α^2 is the variation of the ITEs. Standardized versions of these can be written,

$$\delta_A = \delta/\sigma_c \text{ and } \sigma_A^2 = \sigma_\alpha^2/\sigma_c^2 = 1 + \gamma^2 - 2\rho\gamma,$$

where $\gamma^2 = \sigma_t^2/\sigma_c^2$ is the ratio of the variation in the treatment and control groups respectively. For both parameters the standardization is with respect to the control group variance; this is because in the HTE case, the variation in the treatment group is a function of both the baseline control variance and the effect of the treatment. Note that δ_A is the effect size parameter proposed by Glass (1977) and differs from that proposed by Hedges (1981), which assumes homogenous variances ($\sigma_t^2=\sigma_c^2$). If the goal of standardization is to make these parameters comparable across studies, then scaling in relation to the control group variance is more meaningful (Glass, 1977; McGaw and Glass, 1980). The second parameter, σ_A^2, is the standardized ITE variance. The idea is to supplement the traditional

standardized mean difference with a standardized variance term; the effect of a treatment would then be summarized by the vector (δ_A, σ_A^2).

B. Coefficient of variation (CV) and its inverse (ICV)

While the vector (δ_A, σ_A^2) is a better summary than the univariate measure δ_A, in many cases it is more desirable to focus on their ratio (Longford, 1999; Kravitz, Duan, and Braslow, 2004). There are two quantities that serve this purpose, the coefficient of variation (CV) and the inverse coefficient of variation (ICV),

$$\theta = \sigma_\alpha/\delta = \sigma_A/\delta_A \quad \text{and } \theta_{(1)} = \theta^{-1} = \delta/\sigma_\alpha = \delta_A/\sigma_A.$$

The coefficient of variation is widely used in survey sampling as a scale-free measure of population heterogeneity; several estimators and test statistics have been developed for this case (e.g. Koopmans, Owen, and Rosenblatt, 1964; Vangel, 1996). While the CV is easier to interpret, estimators of the ICV often perform better than those for the CV. The first reason for this is that the CV is undefined when $\delta_A = 0$, whereas when $\rho < 1$, the ICV is always defined. The second reason is that the Taylor series expansion for the ICV (which we will use for delta method approximations in the next section) has fewer terms than for the CV; this results in better small sample properties. As a result, there are cases in which the CV is the parameter of interest, but it is prudent to use the ICV for the calculation of confidence intervals and hypothesis testing. These cases will be discussed in Section IV.

Finally, there are times in which the ICV is the parameter directly of interest. Both Gadbury and Iyer (2000) and Gadbury, Iyer, and Allison (2001), develop maximum likelihood estimators for the ICV in the case that the vector of potential outcomes

$(Y_i(0), Y_i(1))$ is bivariate normal in distribution. In this case, they show that the proportion of the population with negative treatment effects, $P\_$, can be estimated as a function of the ICV, where $P\_ = Pr(\tau_i < 0) = \Phi(-\theta_{(1)})$ and $\Phi(.)$ is the normal CDF.

C. Homogenous variances case ($\gamma^2 = 1$)

It is common in the analysis of experimental results to assume that $\sigma^2 = \sigma_c^2 = \sigma_t^2$, which is to assume homogenous variances. In this case, $\delta_A = \delta_e$, the effect size parameter proposed by Hedges (1981). Here the standardized ITE variance becomes

$$\sigma_A^2 = 2(1 - \rho),$$

where $\sigma_A^2 \in (0, 1)$ for *mildly equalizing* treatments and $\sigma_A^2 \in (1,4)$ for *strongly equalizing* treatments. Furthermore, the CV and ICV become

$$\theta_e = [2(1 - \rho)]^{1/2}/\delta_e \text{ and } \theta_{(1)e} = \delta_e/[2(1 - \rho)]^{1/2},$$

which are both functions of δ_e, which can estimated from the observed data. For interpretation, the CV is often more intuitive, but for estimation, the ICV is here preferable since $\theta_{(1)e}$ is equivalent to a constant times δ_e. Note that *mildly equalizing* treatments will have $\theta_e \in (0,1/\delta_e)$ and $\theta_{(1)e} \in (\delta_e, \infty)$, while *strongly equalizing* treatments will take values of $\theta_e \in (1/\delta_e, 2/\delta_e)$ and $\theta_{(1)e} \in (\delta_e/2, \delta_e)$. For a set of small, medium, and large effect sizes, $\delta_e = \{0.20, 0.50, 0.80\}$, this corresponds respectively to

$\theta_e \in \{(0,5),(0,2),(0,1.25)\}$ for *mildly equalizing* treatments and

$\theta_e \in \{(5,10),(2,4),(1.25,2.50)\}$ for *strongly equalizing* treatments.

This illustrates that small amounts of HTE can lead to very large coefficients of variation when the ATE, δ_e, is small.

These two sets of standardized parameters have in common that they are standardized measures that can be easily compared and aggregated across studies. Just as rules of thumb and criteria for small, medium, and large effect sizes have been developed for the ATE (e.g. Hill, Bloom, Black and Lipsey, 2008), the same could be developed for these standardized measures of HTE (e.g. using a common value of ρ).

III. Estimators of the standardized HTE parameters

Assume that for units $i = 1....n_c$ in the control condition, Y_{ci} are i.i.d. random variables with $\mu_c = E[Y_{ci}] < \infty$ and $\sigma_c^2 = V[Y_{ci}] < \infty$, and $\sigma_c^4 = V[(Y_{ci} - \mu_c)^2] < \infty$. Assume that for units $i=1...n_t$ in the treatment condition, Y_{ti} are i.i.d. random variables with $\mu_t = E[Y_{ti}] < \infty$, $\sigma_t^2 = V[Y_{ti}] < \infty$, and $\sigma_t^4 = V[(Y_{ti} - \mu_t)^2] < \infty$. Furthermore assume that the control and treatment condition vectors are independent, as occurs under random assignment. Let

$$m(c) = (1/n_c)\Sigma\, Y_{ci} \text{ and } m(t) = (1/n_t)\Sigma\, Y_{ti}$$

be the sample means in the control and treatment groups and

$$s_c^2 = 1/(n_c - 1)\Sigma(Y_{ci} - m(c))^2 \text{ and } s_t^2 = 1/(n_t - 1)\Sigma(Y_{ti} - m(t))^2$$

be the associated sample variances. The Central Limit Theorem implies that

$$n_c^{1/2}[m(c) - \mu_c] \rightarrow N(0, \sigma_c^2),$$

$$n_t^{1/2}[m(t) - \mu_t] \rightarrow N(0, \sigma_t^2),$$

$$n_c^{1/2}[s_c^2 - \sigma_c^2] \rightarrow N(0, \sigma_c^4(2+\kappa_c)), \text{ and}$$

$$n_t^{1/2}[s_t^2 - \sigma_t^2] \rightarrow N(0, \sigma_t^4(2 + \kappa_t)),$$

all in distribution. Here $\kappa_\cdot = \mu_\cdot^4/\sigma_\cdot^4 - 3$ is the excess kurtosis of the respective distribution,

where $\mu_{\cdot}^{4} = E(Y_{i\cdot} - \mu_{\cdot})^4$ is the fourth central moment. In the bivariate normal case, $\kappa_c = \kappa_t = 0$.

Since the asymptotic variances of these estimators are themselves functions of the parameters, they can be estimated consistently by their sample quantities. The sample excess kurtosis estimator, however, is negatively biased in non-normal distributions (Bonett, 2006a,b; Royston, 1992). Instead, the estimator

$$K_{\cdot} = n_{\cdot}(A_{\cdot}/B_{\cdot}) - 3$$

is preferable, where $A_{\cdot} = \Sigma(y_{i\cdot} - m'(.))^4$, and $B_{\cdot}=\{\Sigma(y_{i\cdot}-m(.))^2\}^2$, where $m'(.)$ is the trimmed sample mean obtained with a trim proportion of $1/[4(n_{\cdot} - 4)]^{1/2}$ and $m(.)$ is the usual sample mean (Bonett, 2006a,b). This estimator has lower MSE and is less biased than the usual sample estimator.

A. Control standardized mean difference and variance

Both the standardized mean difference, δ_A, and the standardized ITE variance σ_A^2 can be estimated by their sample quantities,

$$d_A = [m(t) - m(c)]/s_c \quad \text{and} \quad s_A^2 = 1 + g^2 - 2\rho g,$$

where $g^2 = s_t^2/s_c^2$. Since these are continuous and differentiable functions of $m(t)$, $m(c)$, s_c^2, and s_t^2, the first order delta method can be used to find the asymptotic distributions and variances of the parameters,

$$V[d_A] = (1/n_c)\{1+\gamma^2(n_c/n_t) + \delta_A^2(2+\kappa_c)/4\}, \text{ and}$$

$$V[s_A^2] =(1/n_c)\{(2(\gamma - \rho)^2 [(2+\kappa_c) +(2+\kappa_t)(n_c/n_t)]\}.$$

In small samples, these estimators are biased. Using a second-order delta method

approximation, reduced bias estimators can be found. Let

$$d_{Au} = k_1 d_A,$$

where $k_1 = 1 - 3[2+\kappa_c]/[8n_c + 3(2+\kappa_c)]$. Note that if the Y_{ic} are normally distributed, this reduces to $k_1 = 1 - 3/(4n_c + 3)$. This estimator is less biased than d_A and has smaller variance, $V[d_{Au}] = k_1^2 V[d_A] \leq V[d_A]$, since $k_1 \leq 1$.

In order to derive a small-sample correction for s_A^2, note that s_A^2 is a linear function of g^2 and g; since the expectation operator is additive, bias in s_A^2 is reduced when bias in g_2 and g is reduced. Let

$$g_u^2 = k_2 g^2$$

$$g_u = k_3 g$$

where $k_2 = 1 - (2+\kappa_c)/(n_c+2+\kappa_c)$ and $k_3 = 1 - [3(2+\kappa_c)/8n_c - (2+\kappa_t)/8n_t]$. In the bivariate normal case, these reduce to $k_2 = 1 - 2/(n_c+2)$ and $k_3 = 1 - [3 - (n_c/n_t)]/4n_c$ respectively. Therefore a reduced bias estimator of s_A^2 can be written,

$$s_{Au}^2 = 1 + k_2 g^2 - 2\rho k_3 g$$

which has asymptotic variance

$$V[s_{Au}^2] = (1/n_c)\{\gamma^2(k_2\gamma - k_3\rho)^2[2 + \kappa_c + (2+\kappa_t)(n_c/n_t)]\}.$$

Note that for $\rho = 0$, $V[s_{Au}^2] < V[s_A^2]$, but for all other values of ρ, the relative size of $V[s_{Au}^2]$ compared to $V[s_A^2]$ depends on the values of ρ and γ.

B. Coefficient of variation (CV) and its inverse (ICV)

The coefficient of variation, θ and its inverse, $\theta_{(1)}$ can be estimated respectively by

$T = [s_c^2 + s_t^2 - 2\rho s_c s_t]^{1/2} / |m(t) - m(c)|$ and

$$T_{(1)} = |m(t) - m(c)|/[s_c^2 + s_t^2 - 2\rho s_c s_t]^{1/2},$$

where |.| denotes the absolute value, which ensures that $T > 0$ and $T_{(1)} > 0$. Since these are continuous and differentiable functions of $m(t)$, $m(c)$, s_c^2, and s_t^2, again the first order delta method can be used to find the asymptotic distribution and variance for the parameters,

$$V[T] = (1/n_c)\{(1/\delta_A^2)[\theta^2(1+\gamma^2(n_c/n_t))+Q]\} \text{ and}$$

$$V[T_{(1)}] = (1/n_c)\{(1/\sigma_A^2)[(1+\gamma^2(n_c/n_t))+\theta_{(1)}^2 Q]\}$$

where $Q = [(1 - \rho\gamma)^2(2+\kappa_c) + (\gamma - \rho)^2\gamma^2(2+\kappa_t)(n_c/n_t)]/4\sigma_A^2$. Note that for both estimators, the denominator of the sampling variance is undefined when $\sigma_A^2 = 0$. As ratio estimators, both T and $T_{(1)}$ are biased and non-normal in small samples. However, since these are functions of four random variables, the second-order delta-method approximations result in reduced bias estimators that are themselves complex functions of the underlying parameters, γ and δ. As a result, these parameters will be more useful in larger samples, as will be investigated in Section IV.

C. Homogenous variances case ($\gamma^2 = 1$)

If homogenous variances are assumed, $\delta_{Ae} = [\mu_t - \mu_c]/\sigma$, where $\sigma^2 = \sigma_c^2 = \sigma_t^2$ is the common variance. This is the effect size parameter proposed by Hedges (1981) and can be more efficiently estimated using $d_e = [m(t) - m(c)]/s$, where $s^2 = [(n_c - 1)s_c^2 + (n_t - 1)s_t^2]/[n_c+n_t - 2]$ is the pooled estimator of the variance. When Y_{ic} and Y_{it} are normally distributed, the *Hedges' g* estimator is unbiased, where

$$g = k_4 d_e,$$

and $k_4 = 1 - 3/[4(n_c+n_t) - 9]$. If the data is possibly non-normal, other effect size

estimators may be more appropriate (e.g. Hedges and Olkin, 1984; Grissom and Kim, 2001; Kromrey et al, 2005). Since HTE can occur even under homogenous variances, the CV and ICV are useful statistics. Since the CV is a function of the inverse of g, we focus here only on the ICV case; any criteria for the CV (e.g. $\theta_e < 2$) could be easily converted to the ICV (e.g. $\theta_{(1)e} > 1/2$). The ICV can be estimated by

$$T_{(1)e} = |g|[2(1-\rho)]^{-1/2}.$$

This relationship implies that in the normal case, $T_{(1)e}$ is unbiased for $\theta_{(1)e}$, and

$$V[T_{(1)e}] = 2(1-\rho)V[g].$$

Note that for $\rho \in (.5,1)$, the *mildly equalizing* treatment case, $V[T_{(1)e}] < V[g]$, whereas for $\rho \in (-1,.5)$, the *strongly equalizing* treatment case, $V[T_{(1)e}] > V[g]$.

IV. Small sample properties and simulation results

A. Simulation set-up

In order to evaluate properties of point and interval estimators (95% confidence intervals) for the parameters δ_A, σ_A^2, θ, and $\theta_{(1)}$, a series of simulations was conducted. Table 2.1 lists the combination of parameters used in the simulations. These parameters were chosen to represent a broad range of ATEs, small and moderate variance ratios, a variety of sample sizes, and values of the unknown correlation (including a couple very close to and including 1). Note that in all cases, for simplicity, we assume equal sample sizes in the treatment and control groups (n_t=n_c=n). These simulations were repeated for three combinations of the normal and log-normal distributions. The log-normal distribution was chosen since it is right skewed with non-zero excess kurtosis (κ).

Table 2.1 about here

In the case that that the $Y_{it} \sim LN(a_t,b_t)$, values of the parameters a_t and b_t were chosen so that $a_t = 2log(\delta + \mu_c) - (1/2)log(\gamma^2 + (\delta + \mu_c)^2)$ and $b_t = log(\gamma^2 + (\delta + \mu_c)^2) - 2log(\delta + \mu_c)$, where $\mu_c = E(Y_{ic})$. In the case that $Y_{ic} \sim LN(a_c,b_c)$, the above relationships hold when $\delta = 0$ and $\sigma^2 = 1$. In both the *N(c)LN(t)* and *LN(c)LN(t)* cases, $\mu_c = 3$ was used in order increase the spread of the Y_{it} distribution. Note that in this case, the excess kurtosis can be calculated as $\kappa = b^4 + 2b^3 + 3b^2 - 9$.

The simulations were conducted in the statistics program **R.** For each parameter combination, 10,000 simulations were conducted, and for each simulation both a point estimate and confidence interval were created. For a general parameter ω and estimate W, confidence intervals were created using the normal distribution with

$$\omega \in (W - z_{\alpha/2}\, v[W]^{1/2},\ W + z_{\alpha/2}\, v[W]^{1/2})$$

where $v[W]$ is the estimated variance found by substituting the estimated values of the parameters into the variance approximation $V[W]$. Note that in all of the cases studied here, $\alpha = .05$ and $z_{\alpha/2} = 1.96$. Additionally, the degree of estimation bias across simulations was calculated as the bias ration (BR),

$$BR(W) = [E(W) - \omega]/\omega$$

where $E(W)$ is the average value of the estimate across all 10,000 simulations. We focus here on the standardized bias ratio instead of the absolute bias in order to make the results more easily comparable across parameter values. When $BR(W) = 0$, the estimate is

unbiased, while when $BR(W) = k$, the average estimate is $k*100\%$ larger than the true value.

B. Simulation results

The simulation study we conducted covered a wide variety of parameter value combinations and resulted in a large amount of data. We do not present all of this data here, and instead focus on overall trends. In particular, we focus on interpreting these trends in relation to ρ and n, since these two parameters are controlled by the analyst, whereas γ and δ are unknown. We present the overall results in a series of figures and divide the discussion into those for the mean difference, those for the ITE variance, and those for their ratio.

1. Estimators of δ_A: d_A and d_{Au}

The unbiased estimator d_{Au} outperforms the biased estimator, d_A, in terms of both coverage and bias. Figure 2.2 summarizes the results of these simulations via a series of boxplots for both bias and coverage. In general, coverage for d_{Au} is larger, particular for $n \leq 100$. For the *N(c)N(t)* and *N(c)LN(t)* cases, coverage for both estimators approaches 95% and is within simulation error bounds for $n \geq 25$. However, for the *LN(c)LN(t)* case, the median coverage value approaches 93% as n grows large and the range of coverage values over the parameters does not grow narrower, though for all parameter values coverage is in the 90 – 96% range. Analyses not shown here, however, indicate that coverage is better for smaller values of δ_A; for example, for $\delta_A < 0.5$, coverage is above 92% for all values of n.

In terms of bias, the estimator d_{Au} clearly outperforms d_A for all three distributional pairs. For the *N(c)N(t)* case, the median bias approaches zero very quickly as *n* grows larger. Importantly by for $n \geq 25$, the bias is never larger than 20% of the value of δ_A. For the *LN(c)LN(t)* case, however, the bias is larger when $n=10$, though for d_{Au} the bias here is between 0 and 60% of the size of δ_A. For $n \geq 25$, this bias is never greater than 40%, however. Overall, this suggests that the estimator d_{Au} outperforms d_A; it performs the worst when there is non-zero excess kurtosis, though this is only problematic for samples with $n < 25$.

Figure 2.2 about here

---

2. Estimators of σ_A^2: s_A^2 and s_{Au}^2

The overall results for coverage and bias for both s_A^2 and s_{Au}^2 are shown in Figure 2.3. In the results depicted here, the same variance estimator, $v[s_A^2]$, is used for both s_A^2 and s_{Au}^2; this is because the variance estimator $v[s_{Au}^2]$ resulted in confidence intervals that were too short. In terms of coverage, s_{Au}^2 has slightly lower values, in terms of both median and range. The median coverage rises over 90% for s_{Au}^2 at $n=50$, $n=250$, and $n=250$, respectively for the three sets of distributions. This means that the confidence intervals created using $v[s_A^2]$ are typically liberal, particularly in samples smaller than $n=250$. Additionally, analyses not shown here revealed that for values of $\rho \leq .50$, the same trend in terms of median coverage probabilities by *n* holds, but with the range of coverage

probabilities is much smaller. The largest problems occur for $\rho \geq 0.9$, particularly for small n.

In terms of bias, the estimator s_{Au}^2 clearly outperforms s_A^2. The difference between these is most pronounced for $n=10$, though in this case, for some parameter value combinations the bias in s_A^2 can still be quite large. For $n=25$, however, the bias for s_{Au}^2 is more reasonable, and is largest for the *LN(c)LN(t)* case. For both estimators, when $n \geq 100$, bias is very small. The bias problems, however, largely disappear when we focus only on values of $\rho \leq 0.9$. In this case, for all three distributional pairs, the range of bias ratios for all values of n are less than 1.5, and for $n \geq 25$, less than 1. Furthermore, the median bias ratio is less than $.31$ for s_A^2 and $.15$ for s_{Au}^2 when $n=10$, and less than 0.11 for s_A^2 and 0.04 for s_{Au}^2 for $n=25$. Bias is largest when $\rho > 0.9$, particularly when the value of γ is close to 1 (which is, of course, unknown). Altogether this suggests that using s_{Au}^2 and $\rho \leq 0.9$ is optimal in terms of minimizing estimation bias.

Figure 2.3 about here

---

3. Ratio estimators (θ and $\theta_{(1)}$): T and $T_{(1)}$

Results for coverage and bias for the coefficient of variation and its inverse are illustrated in Figure 2.4. We compare these to each other here since in both cases, the interest is in comparing the standard deviation of the ITEs to the ATE. In many cases, the decision to interpret this ratio in terms of the CV or ICV will have to do with which one

has better sampling properties. For example, note that confidence intervals for θ can be converted to confidence intervals for $\theta_{(1)}$ (and vice versa), using the fact that $\theta \in (T_L, T_U)$ implies that $1/\theta = \theta_{(1)} \in (T_{L(1)} = 1/T_U, T_{U(1)} = 1/T_L)$, where in the case that $T_L \leq 0$, $T_{U(1)} = \infty$ and the interval estimate for $\theta_{(1)}$ becomes a one- sided lower interval estimate.

In terms of coverage, as Figure 2.4 shows, when looking across all parameter values T clearly outperforms its inverse, $T_{(1)}$. However, when we focus on values of $\rho \geq 0.9$ (results not shown in the figure), the trend reverses. For $\rho \geq 0.9$, the median coverage is greater than 93% for T and greater than 95% for $T_{(1)}$. The range of coverage values is much smaller, however, for $T_{(1)}$; when $n = 10$, coverage ranges from 83 – 100% for $T_{(1)}$, but from 56 – 100% for T. Coverage is above 90% for all parameter value combinations for $T_{(1)}$ for $n \geq 25$. These results hold for all three distributional pairs studied here. Overall this means that if a value of $\rho \geq 0.9$ is used, it is preferable to do hypothesis testing and confidence interval construction in relation to the ICV instead of the CV. When smaller values of ρ are of interest, the CV, however, is preferred.

In terms of bias, overall it appears that the opposite trend is true here – that the ICV estimator $T_{(1)}$ is less biased than T. This is not surprising, however, since as noted previously the Taylor series expansion for the ICV has fewer terms than that for the CV, which results in a better linear approximation with less bias. For both parameters, the inter-quartile range and median bias grow smaller as n increases; however, for both, there remain some outliers with large bias ratios even for n as large as 250. When $\rho \leq 0.9$, the maximum bias ratio for T is reduced to about 25 when $n=10$, 15 for $n=25$, 50 and 10 for $n \geq 100$. Using $\rho = 0.99$ leads to much higher maximal values. In contrast, for $T_{(1)}$, restricting

to $\rho \leq 0.99$ leads to maximal bias ratios less than 15 for $n=10$, 10 for $n=25,50$, and less than 5 for $n \geq 100$. Further investigations show that the bias is a complex function of the parameters; the relationship is not nearly as straightforward as that between ρ and d_A or s_A^2.

Taking coverage and bias into account, these simulations suggest that if the CV is the parameter of interest, the value of $\rho = 0.9$ has the best coverage and bias properties for all distributions and parameter values studied here. Additionally, this analysis suggests that when $\rho = 0.9$ is used, interval estimates for θ should be based on the inverted confidence intervals for $\theta_{(1)}$ shown at the beginning of this section. For values of $\rho < 0.9$, confidence intervals based on the CV perform better and the bias ratios are typically smaller. Finally, note that if bias is a concern, it may be better to focus interest on the ICV.

Figure 2.4 about here

V. Example

The Tennessee STAR class size experiment has been widely studied and is an example of a large-scale educational experiment. Since each of the schools in the experiment had at least one small classroom (the treatment) and one regular sized classroom (the control), school specific treatment effects can be calculated and aggregated using a multi-level model. These studies have shown that the average school level small-classroom effect is positive, but that not for some schools, the school-specific average treatment effect is negative (Nye, Hedges and Konstantopoulos, 2000). Recently, Konstantopoulos (2008) investigated if within-school treatment effects might be

heterogeneous. In particular, he focused on comparing the quantiles of the treatment effects for high and low achievers and found that higher-achieving students benefited more from being in a small classroom than lower-achieving students, and that these differences were most pronounced in early grades.

For each school in the Tennessee STAR dataset, we calculated the average reading and math test scores and the variation in these test scores separately for students in small classrooms (the treatment group) and those in regular or regular plus aide classrooms (the control group). For each school, we then calculated school specific estimates of d_{Au}, s_{Au}^2, T, and $T_{(1)}$, their sampling variances, and 95% confidence intervals. When the lower-bound of the confidence intervals for s_{Au}^2, T, or $T_{(1)}$ were negative, we truncated these to zero. We also calculated confidence intervals for T by inverting the confidence intervals for $T_{(1)}$ – labeled "*CI (T)inv*" in Table 2.2 – as suggested in Section IV. We focus here on the potential outcome correlation $\rho = 0.9$, since this is close enough to one as to provide only a minimal value of HTE and since simulations suggest that this value leads to estimators with good properties. Note that $\rho = 0.9$ corresponds to a treatment that is likely to be *disequalizing* or *mildly equalizing*. Finally, following the results found in the simulations, we use the variance estimators based on the biased parameters ($v[d_A]$, $v[s_A^2]$).

In the interest of space, we report here on only 10 of the schools in the experiment. These are not a random sample, but have been chosen for illustrative purposes. For each of these schools, Table 2.2 includes estimates of the four parameters and 95% confidence intervals for both reading and math. Note that while we suggest that the inverted confidence intervals for T perform better, we include the confidence intervals based

directly on T as well for comparison. As the table shows, the inverted confidence intervals tend to have larger lower-bounds but are much wider; in many cases, the upper bound is infinity. This reveals that in many cases, the benefit of the analytic approach developed here is that it puts a lower-bound on the amount of HTE in an experiment, both in terms of ρ and in terms of interval estimation.

Table 2.2 about here

Table 2.2 reveals a few interesting cases. Seven of the schools for reading and math have confidence intervals for δ_A that do not include zero; in only one of those cases is the effect significant and negative. Additionally, seven of the schools for reading and eight for math have confidence intervals for σ_A^2 that do not include zero. The average estimate of σ_A^2 for reading is 0.25 and for math is slightly larger, 0.29. Interestingly, while the ATE for School 4 is significant and negative, $\delta_A \in$ (-0.936,-.051), the HTE variance is non-zero, $\sigma_A^2 \in$ (.145, .193). This indicates that while the treatment effect for the average student is in fact negative, for some subset of students, the effect of the treatment may actually be positive. This can be summarized more clearly through the coefficient of variation which is estimated to be $T=0.875$, with an (inverted) confidence interval of (.472, 13.5).

Note, however, that the confidence intervals for σ_A^2, θ, and $\theta_{(1)}$ should be interpreted carefully here since the sample sizes are generally less than 100 in both the treatment and control groups. Since both d_A and s_{Au}^2 exhibit only small amounts of bias for samples of these sizes, to draw larger conclusions these results could be meta-analyzed; note that

meta-analyzing T or $T_{(1)}$ could be problematic since the estimators are often biased in even moderate samples. Finally, since the variance estimators are functions of the parameters and are also biased in small samples, a meta-analysis method robust to distributional and weight misspecification would here be preferable (Hedges, Tipton, and Johnson, 2010).

VI. Conclusion

The focus of this paper has been on the development of standardized coefficients for the average treatment effect and the individual treatment effect variability. A problem with estimating this variability is that it depends on an unidentified parameter – the correlation between the potential outcomes. The approach we advocate here is a sensitivity analysis that takes into account whether a treatment is more or less likely to be *equalizing* or *disequalizing* and which allows for minimal values to be calculate (by choosing a large correlation). We have focused here on evaluating the bias of the point estimates and the confidence intervals under a variety of parameter values and for small and moderate sample sizes. Future work in this area could address the estimation of additional moments (e.g. the skew) of the distribution of individual treatment effects, issues with meta-analyzing these quantities, and guidelines for what values of the coefficient of variation are desirable.

CHAPTER 3

Robust Variance Estimation in Meta-regression with Binary Dependent Effects

Traditional meta-analytic theory requires that each study contribute only one outcome to a research synthesis. While methods for combining dependent effect sizes exist, these require information on the dependence structure of estimates, which is often not reported in published studies. Recently, Hedges, Tipton, and Johnson [HTJ] (2010) proposed a robust method for estimating meta-regression coefficients and their standard errors when there are dependent effects.

HTJ provides a theorem that shows that their robust variance estimator is asymptotically unbiased, as well as results from a simulation study which investigates the accuracy of the variance estimator when the number of studies is small. The focus of both their simulations and example, however, is the standardized mean difference effect size. In many studies the outcomes of interest are instead a function of underlying dichotomous data -- the log odds ratio, the log risk ratio, or the risk difference.

After reviewing estimators and the main results of HTJ, this paper summarizes the results of several simulation studies which investigate the empirical coverage probabilities for 95% confidence intervals for the average effect (intercept) and slope cases. Additionally, an example is presented which reanalyzes data from Landenberger and Lipsey (2005) in a robust standard error framework.

I. Effect sizes in dichotomously generated data

This paper focuses on three effect size measures: the risk difference, log risk ratio, and log odds ratio. Here we define relevant estimators of both the parameters and their

sampling variances. Note that while we will assume these measures are correlated, this correlation does not effect either the estimators or their sampling variances shown here.

Assume that there are j=1… m studies each with i=1…k_j correlated outcomes and that for each outcome i in study j,

$m_{Tij} \sim bin(n_{Tij}, \pi_{Tij})$, and $m_{Cij} \sim bin(n_{Cij},\pi_{Cij})$.

That is, assume that for outcome i in study j and treatment group T or C, m_{ij} successes out of n_{ij} trials are observed with the probability of success π_{ij}. The outcomes of studies of this type are often reported in a 2x2 table, which on one axis consists of the two groups, treatment and control, and on the other axis consists of the two categories, success and failure. In summarizing a table of this type, the statistics of interest are generally

$p_{Cij} = m_{Cij}/ n_{Cij}$

$p_{Tij} = m_{Tij}/ n_{Tij}$

which have expected values π_{ij} and sampling variances $\pi_{ij}(1 – \pi_{ij})/n_{ij}$. For the remainder of this section, note that we suppress the (i,j) notation for convenience.

A. Risk Difference

The risk difference is easily defined as $D = \pi_T – \pi_C$ and takes values in [-1,+1]. This can be estimated using the maximum likelihood estimator $d = p_T – p_C$ which is unbiased even in small samples. Its exact variance is $V(d) = \pi_T(1 – \pi_T)/n_T + \pi_C/(1 – \pi_C)/n_C$. This variance is estimated by substituting p for π, and when p is 0 or 1, by adding a half count to the underlying 2x2 table.

B. Risk Ratio

The ratio of risks, $R=\pi_T/\pi_C$, is generally estimated (and tested) using its logarithm,

$$L_R = log\ R = log\ \pi_T - log\ \pi_C,$$

which takes values in $(-\infty,\infty)$. The maximum likelihood estimator, $l_R = log\ p_T - log\ p_C$, is undefined when $p_C = 0$; therefore we use the estimator provided by Pettigrew, Gart, and Thomas (1986),

$$l_R = log\ (p_T + 1/2n_T) - log(p_C + 1/2n_C).$$

Note that this estimator is biased when the underlying within-study sample size is small. This estimator has sampling variance

$$V(l_R) = (1 - \pi_T) / (n_T\ \pi_T)\ + (1 - \pi_C)/ (n_C\ \pi_C),$$

which can again be estimated by substituting p for π, and when p is 0 or 1, by adding a half count to the cells in the underlying 2x2 table.

C. Odds Ratio

The least intuitive, but most mathematically desirable, parameter is the ratio of odds, $O = [\pi_T/(1 - \pi_T)]/ [\pi_C/(1 - \pi_C)]$. Again, estimation and testing is usually conducted on the log scale, and

$$l_O = log\ O = log\ [\pi_T/(1 - \pi_T)] - log\ [\pi_C/(1 - \pi_C)],$$

which also takes values in $(-\infty,\infty)$. The maximum likelihood estimator (which substitutes each π with its relevant p), however, is undefined when $p_C = 0$ or $p_T = 1$. We therefore use the estimator (Pettigrew, Gart, & Thomas, 1986),

$$l_O = log(p_T + 1/2n_T) - log(1 - p_T + 1/2n_T) - log(p_C + 1/2n_C) + log(1 - p_C + 1/2n_C),$$

which is biased in small samples. This estimator has asymptotic variance

$$V(l_O) = 1/[n_T \pi_T (1 - \pi_T)] + 1/[n_C \pi_C (1 - \pi_C)],$$

which we again estimate by substituting p for π, and when p is 0 or 1, by adding a half count to the cells in the underlying 2x2 table.

II. Summary of the results in HTJ

When there are multiple effect size measures per study, traditional methods have involved modeling directly the underlying dependence structure in each study (Gleser & Olkin, 1994), using only one effect size per study (e.g. either randomly choose one estimate or using their average), or using all measures but basing results on the (incorrect) assumption that the estimators are independent. However, HTJ proposed a new method based on robust variance estimation, and these results apply for all effect sizes estimators. The main theorem of the paper states the following.

Theorem 3.1: Robust variance estimation (from HTJ)

Let $\mathbf{X}_j$ be the design matrix for meta regression, $\mathbf{W}_j$ be a general weight matrix, and $\mathbf{T}_j$ be the vector of effect sizes for the jth study. We can relate these via the regression

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{T} = (\mathbf{T}_1,\ldots,\mathbf{T}_m)'$ is a vector of m vectors, each with k_j effect size measures, and $\mathbf{X} = (\mathbf{X}_1,\ldots,\mathbf{X}_m)'$ is a design matrix of m stacked matrices, each of dimension k_j x p, and $\boldsymbol{\beta}$ is a p x 1 vector of coefficients to be estimated. Finally, note that $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_m)'$ is the

vector of stacked residual vectors, each of dimension k_j, and that we make no distributional assumptions.

We can estimate the set of regression coefficients by

$$\mathbf{b} = \left(\sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \mathbf{X}_j \right)^{-1} \left(\sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \mathbf{T}_j \right), \tag{3.1}$$

and a robust variance estimator for these coefficients can be obtained by substituting the matrix of cross products of within-study residuals in the j^{th} study for $\mathbf{\Sigma}_j$, that is

$$\mathbf{V}^R = \left(\sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \mathbf{X}_j \right)^{-1} \left(\sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \mathbf{e}_j \mathbf{e}_j' \mathbf{W}_j \mathbf{X}_j \right) \left(\sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \mathbf{X}_j \right)^{-1}, \tag{3.2}$$

where $\mathbf{e}_j = \mathbf{T}_j - \mathbf{X}_j\mathbf{b}$ is the (k_j x 1) residual vector in the j^{th} cluster. Although $\mathbf{e}_j\mathbf{e}_j'$ is a rather poor estimate of $\mathbf{\Sigma}_j$, it is good enough that (3.2) converges in probability to the correct value as $m \to \infty$.

Proof: See HTJ Appendix A.

****

Additionally, HTJ show that using a variance correction factor of $m/(m-p)$, where m is the number of studies and p is the number of regression coefficients estimated gives confidence intervals with empirical coverage probabilities closer to nominal. Using this method, a 95% confidence interval for the k^{th} regression coefficient β_k can be estimated using

$$b_k - t_{m-p,\alpha}\,[V_{Rkk}(m/(m-p))]^{1/2} \le \beta_k \le b_k + t_{m-p,\alpha}\,[V_{Rkk}(m/(m-p))]^{1/2} \tag{3.3}$$

where b_k is the estimate of β_k found using (3.1), $t_{m-p,\alpha}$ is the level-α t-value with $m-p$ degrees of freedom, and V_{Rkk} is the kth diagonal value of the V_R robust variance matrix.

While *Theorem 3.1* holds for any weight matrix **W**, HTJ suggest that in the correlated outcomes case an easy method of weighting is to use inverse average variance weights. Assuming that each outcome $i=1...k_j$ in study j has estimated variance v_{ij}, then fixed effects weights can be written

$$w_{ij} = 1/(k_j \bar{v}_j) \qquad (3.4)$$

where $\bar{v}_j = (1/k_j)\Sigma v_{ij}$ is the average variance estimate in study j. The benefit of using these weights is that the total weight for study j does not depend on the number of estimated outcomes reported within the study.

Finally, note that in the simple case of estimating a mean effect for a set of dependent effect sizes, this estimator reduces to

$$b_1 = \frac{\sum_{j=1}^{m}\sum_{i=1}^{k_j} w_{ij}T_{ij}}{\sum_{j=1}^{m}\sum_{i=1}^{k_j} w_{ij}}$$

where w_{ij} is the weight given to estimate T_{ij}, which is the outcome i in study j. In the case that every outcome in the same study receives the same weight (as above), the robust estimator of the sampling variance of b_1 can be shown to be

$$v^R = \frac{\sum_{j=1}^{m} w_j^2 \left(\bar{T}_j - b_1\right)^2}{\left(\sum_{j=1}^{m} w_j\right)^2}$$

where $\bar{T}_j$ is the un-weighted average effect size and in study j, b_1 is estimated as above, and $w_j = \Sigma w_{ij}$ is the total weight for the same study.

III. Simulation study

A. Simulation set-up

A simulation study was conducted to investigate the small sample properties of the variance estimator in (3.2) when the underlying data consists of functions of dichotomous outcomes. The focus here is the empirical coverage probabilities for 95% confidence intervals which are constructed using equation (3.3). Note that these intervals include the small sample correction term $m/(m-p)$ suggested by HTJ, where m is the number of studies and p is the number of predictors. We focus here on two cases: the estimation of an average effect size (intercept) and of a single slope. The simulations were conducted in **R**, using the built-in **rbinom** function for generating random binomials. For the intercept and slope estimation cases and for each of the three effect sizes and parameter sets, we report empirical coverage probabilities based on 10,000 simulations. *Appendix B* contains details on the exact data generation method used.

In the intercept case, the data generation process involved two steps. First, for each study $j = 1 \ldots m$ values of the underlying probabilities in the treatment and control groups, π_{Tj} and π_{Cj}, were generated from a beta distribution. We assumed that the outcome specific probabilities $\pi_{Tij} = \pi_{Tj}$ and $\pi_{Cij} = \pi_{Cj}$ for all outcomes $i=1 \ldots k_j$ in study j. We parameterized the beta distribution in a way that allowed the π_{Tj} and π_{Cj} pairs to have average probabilities π_T and π_C and to vary in relation to the measure of between study variation I^2 (Higgins & Thompson, 2002). Importantly, we did not invoke the normal distribution in generating the between-study variability in effect sizes. By using the beta distribution,

which is a natural prior to the binomial distribution, we are able to simulate a true hierarchical binomial data generating process.

In the second step, for each study j, k_j correlated binary outcomes were generated for each of n_j units in the treatment and control groups. We assume throughout that the treatment and control group outcomes are independent and that $n_{Tij} = n_{Cij} = n_j$, as is the case in balanced experiments with multiple outcome measures. From these, k_j probability pairs (p_{Cij}, p_{Tij}) were estimated for each study j. For a particular p_{ij} value in the treatment or control groups, the data generation method used has the following properties: $E(p_{ij}) = \pi_j$, $V(p_{ij}) = \pi_j(1 - \pi_j)/n_j$, and $Corr(p_{ij},p_{kj}) = \rho$, where $E(.)$, $V(.)$, and $Corr(.)$ are the expectation, variance, and correlation functions. Note that ρ is the within study correlation between effect size estimates and takes values in $[0,1]$.

In the slope estimation case an additional step was required. After generating the study specific probabilities π_{Cj} and π_{Tj}, for the treatment group for each outcome $i = 1 \ldots k_j$ probabilities π_{Tij} were generated using the appropriate regression relationship based on the effect size of interest. For example, in the risk difference case, this relationship was $\pi_{Tij} = \pi_{Tj} + \beta x_{ij}$. For simplicity, we assumed $\beta = 1$. Note that we also generated the underlying x_{ij} values according to distributions that allowed each π_{Tij} value to remain in $(0,1)$.

Finally, in each study after generating k_j values of p_{Tij} and p_{Cij}, we summarize the effect of the treatment by calculating k_j estimated effect sizes T_{ij}. Note that T_{ij} could be the risk difference (d), the log risk ratio (l_R), or the log odds ratio (l_O). This means that the data is generated using the same method regardless of the effect size used. We use this

approach since it allows for greater comparability of the coverage probabilities across effect sizes.

Tables 3.1 and 3.2 below summarize the parameter values used in this simulation study. These values were chosen since they are common in the literature (Emerson, Hoaglin, & Mosteller, 1993; 1996; Hartung & Knapp, 2001; Knapp & Hartung, 2003). We include both cases in which the within study sample sizes are the same across studies (*n=20, 50, 100*) and in which they vary study to study (*n= nV1, nV2*). The first case, *nV1,* is one in which there are mostly small studies and a few larger studies; the second case, *nV2,* is one in which these sample sizes vary uniformly. Note that we chose $n = 20$ to be the smallest value since smaller sample sizes induce large estimation biases, which are not the focus of this study. We also include cases in which the number of within-study correlated outcomes k is constant across studies (*k=1,2,5,10*) and cases in which this varies study to study (*k=kV1, kV2*). The first case, *kV1* is the case in which there are a combination of a variety of k_j values; conversely, *kV2* is the case in which most studies have only one outcome but 20% of studies have many outcomes. We chose these values since cases of these types are likely to occur in practice. Additionally, we focus on three values of the between-study parameter variability term, I^2 *(= 0, 1/3, 1/2)* and on three sets of underlying probabilities, $(\pi_C, \pi_T) = (.10, .10), (.50,.50),$ and *(.30,.40).* These first two sets of values were the focus of Hartung and Knapp (2001) and address both moderately small and moderate probabilities. We add the case *(.30,.40),* which corresponds to a risk difference of *0.10*, a risk ratio of *1.33*, and a odds-ratio of *1.56* to cover a case in which the null-hypothesis of a zero effect would not be true.

Tables 3.1 and 3.2 about here

For all of the simulations, we use fixed effects weights such that each measure within the same study receives the same weight, as found in (3.4). We use these weights even in the case that the treatment effect varies across studies, since the asymptotic results found in the main theorem apply for *any* weight matrix, optimal or not. Finally, for each combination of parameter values, we summarize the results of the simulation by calculating the empirical coverage probability for the theoretical 95% confidence interval.

B. Simulation results

Summarizing the results of six simulation studies, each covering a broad range of parameter values, can be cumbersome. The solution used here is to provide condensed results in the tables and to supplement these with commentary. While we divide the tables by effect size, we divide the commentary into sections on intercepts and slopes separately since there appear to be trends common to all three effect sizes within these groupings. Note that in all tables, we average across values of the underlying correlation ρ, since in our simulations we found there to be no strong relationship.

Table 3.3 reports results for the mean and slope estimation cases for the risk difference (RD), while Tables 3.4 and 3.5 report the same for the log risk ratio (LRR) and log odds ratio (LOR) respectively. Reading the tables from left to right, the tables report the empirical coverage probabilities for the estimation of the mean effect (intercept) and

slope. Within each case, results are reported separately for the three underlying probability pairs. We also report these results by the number of within-study outcomes; for *k = 2, 5, 10* instead of reporting each value separately, the tables show the minimum and maximum values. Finally, reading the tables from top to bottom, the results are divided by the within-study sample size (*n*), within these by the amount of between-study variability in effect sizes (I^2), and finally, by the number of studies (*m*).

In order to make reading the tables easier, two guides are provided. First, note that lines are drawn on the tables in such a way that each box contains three values; from top to bottom these boxes give the coverage probabilities as the number of studies increases, from *m = 10* to *m=40*. Theoretically it is expected that as *m* increases these coverage probabilities approach nominal. Second, we have included arrows on the table to indicate cases in which coverage is outside the interval (94.57%, 95.42%), which is a simulation error interval based on 10,000 simulations. Down-arrows indicate confidence intervals that are too short, while up-arrows indicate confidence intervals that are too long.

Tables 3.3, 3.4, and 3.5 about here

---

1. Intercepts

As the left-half of the three tables illustrates, confidence interval coverage in the mean effect size (intercept) case is generally close to nominal. When the underlying sample sizes are identical in all studies (the equal *n* case), for moderate and large *n*, coverage is close to nominal for all three effect sizes. When *n* and the underlying

probabilities, (π_C, π_T), are smaller, coverage tends to be either larger than nominal (the RD case), or smaller than nominal for the LRR and LOR cases. In the simulations shown here, coverage remains above 94% in all of these cases. However, in simulations not reported here, we found that for even smaller probabilities and values of *n*, coverage was often far below nominal. These small coverage problems were entirely the result of small sample bias in the estimators and their weights; using non-estimated weights (e.g. inverse n-weights; Emerson, Hoaglin & Mosteller, 1993, 1996) greatly improved the coverage in these cases. We do not focus on these cases here, however, since these issues of bias effect all meta-analytic confidence interval estimators. In general we find that in the equal sample size (*n*) case, coverage is close to nominal for all effect sizes investigated here.

We also investigated two cases in which the study specific sample sizes, n_j, varied from study to study. In the first case, *nV1*, where 60% of the studies had sample sizes of *n=20* and the remaining studies had larger sample sizes of *n=100*, coverage is found to be less than nominal for every case investigated here and for all three effect sizes. For the values studied here, coverage was generally in the 87 – 94% range. A general trend is that holding the number of studies (*m*) constant, as the between-study heterogeneity (I^2) increases, coverage slightly diminishes. However, as the number of studies increases, coverage approaches nominal values. While for *m = 10*, coverage is in the 88 – 93% range, for *m = 40*, coverage is in the 93 – 94% range consistently. These trends are similar for all three effect sizes.

In the second varying case, *nV2*, where the sample sizes vary more uniformly (from 20 to 100), while coverage is less than nominal it is in the 93 to 94% range for all

parameter values studied and for all three effect sizes. For m = 20 and 40, in fact, coverage is 94% or better in all cases, and in general as the number of studies increases, coverage improves. Finally, note that the tables do not report coverage probabilities separately for the *k = 2,5,10*, *kV1*, or *kV2* cases. Instead the minimum and maximum coverage probabilities are shown; these values are close in most cases, indicating that the coverage probabilities do not vary in much in relation to k.

Overall, for the estimation of the mean effect case these simulations suggest that the robust variance estimator leads to close to nominal confidence interval coverage. However, coverage is generally less than nominal when m is small, which means that the estimated confidence intervals are too short.

2. Slopes

Whereas coverage in the intercept estimation case is often close to nominal, in the intercept estimation case coverage is generally less than nominal and sometimes closer to 90%. In general, coverage is best for the log odds ratio and log risk ratio; it is generally good for the risk difference, though there are a few problematic cases. The main trends here are in relation to k and n.

a. Simulation results related to increasing k

The simulation results in HTJ for the standardized mean difference indicated that as the number of within study estimates increased, coverage also increased. This trend was particularly noticeable in the slope estimation case (HTJ *Table II*). The general trend

reported there was that coverage for *m=10* is about 90% with *k* =1, but about 93% with *k* = *2* and 95% with *k=5*. By the time *m = 40*, this trend becomes about 93% coverage for *k=1*, 94% coverage for *k = 2* and 95% coverage for *k = 5*. This trend is similar across values of ρ, I^2, and *n*.

In the simulations conducted here for the binary outcomes case, a similar trend is found. The comparability of this trend is most obvious for the *n* = 100 case, which has the least estimation bias. In Tables 3.3 – 3.5, we report separately coverage for *k=1* and the minimum and maximum coverage for *k=2,5,* and *10*. While not reported in the tables, the minimum coverage generally occurs for *k=2* and the maximum for *k=10*. In the cases in which *k* varies from study to study (*kV1* and *kV2*), coverage is generally in between these extremes. For example, when 80% of the studies have *k=1* and 20% have *k=10* (pattern *kV1*) coverage is between that of the constant *k=1* and *k=10* cases. This is the same trend found in previous simulations (HTJ *Table III*).

An important question is why coverage should increase with *k*, particularly when the result of *Theorem 3.1* implies that coverage should primarily increase as the number of studies *m* increases. There are two reasons this should occur. In the general case, not particular to binary outcomes, assuming each outcome *i* within the same study *j* with $k_j > 1$ effect sizes receives the same weight $w_j = w_{j*}/k_j$ and where $w_{j*} = \Sigma w_{ij}$ is the total weight for study *j*, the variance estimator of the slope can be written as,

$$V_R(b_1) = \frac{\sum_{j=1}^{m} w_{j*}^2 \left(\frac{1}{k_j} \sum_{i=1}^{k_j} \left(e_{ij} - \bar{e}_j\right)\left(X_{ij} - \bar{X}_j\right) \right)^2}{\left\{ \sum_{j=1}^{m} w_{j*} \left(\frac{1}{k_j} \sum_{i=1}^{k_j} \left(X_{ij} - \bar{X}_j\right)^2 \right) \right\}} \tag{3.5}$$

when the covariate X_{ij} varies within studies and is centered about its study average $\bar{X}_j = (1/k_j)\Sigma X_{ij}$. Recall that the estimated residuals $e_{ij} = Y_{ij} - [b_0 + b_1 X_{ij}]$ where b_0 and b_1 are estimated using (3.1). This variance estimator involves averaging over not just the number of studies m but also the number of within study residual estimates k_j. Clearly as k_j increases (3.5) should improve; this is the reason that in general we should expect the variance estimator to improve not just with m but also with k_j.

A second reason for coverage improvement with increasing k is specific to the binary outcomes case and applies particularly when n is small. Recall that in all of these simulations, within each study j each outcome i receives weight $w_{ij} = 1/(k_j \bar{v}_j)$. An additional reason for the improved coverage in these cases, particularly for the risk difference, is that the when n is small the estimated effect size parameters and their estimated variances are correlated. This is because both are functions of the estimated probabilities p_{Cij} and p_{Tij}. By using inverse-average-variance weights (instead of inverse-variance weights, as occurs in the $k=1$ case), this correlation is reduced, which reduces bias and improves confidence interval coverage. This trend is explained more formally in *Appendix C*.

b. Trends related to n

A general problem that affects all three effect sizes here is within-study estimation bias. Both the estimators of the log odds ratio and log risk ratios are biased in small samples (Pettigrew, Gart, & Thomas, 1986). While the estimator of the risk difference is not biased, the correlation between the estimate and its estimated weights can induce large biases when *n* is small (Emerson, Hoaglin, & Mosteller, 1993). As a result, coverage tends to vary in relation to *n* and matters most when the underlying probabilities (π_C, π_T) are small or rare (Bradburn, Deeks, Berlin, & Localio, 2007). Additionally, in the cases in which *n* varies from study to study (*nV1* and *nV2*), coverage is typically worse than in the equal *n* cases; this is particularly true for *nV1* where 60% of the studies have n_j =20 while the other 40% have n_j = 100. In order to separate the *k* and *n* issues, in the remainder of this section we focus on the *k=1* case. Note that the trends for *n* are less pronounced when *k >1*.

For the LOR, in the equal *n* case the trend is subtle. In general the coverage for the LOR is the most stable of the three effect sizes studied here. For example, for *(.10,.10)* when $I^2 = 0.5$ and *n = 20* coverage increases from 88% to 91% as *m* increases from 10 to 40, whereas when *n=100* coverage increases from 90% to 93%. In the case *nV2* (where n_j varies from *20* to *100* uniformly), coverage follows a similar trend to the equal *n* case in which *n =20*. Here as *m* increases from *10* to *40*, coverage ranges from about 89% to about 93%. Most problematic is the *nV1* case, where coverage ranges instead from about 85% to about 90% as *m* increases. Notably this trend is similar for all values of (π_C, π_T) studied.

Like the LOR, the LRR performs most like the HTJ results when the baseline probabilities are not small, *(.50,.50)* and *(.30,.40)*. The general trends here are that while coverage increases with *m*, it appears to increase at a slower rate. For *n=100* in these two cases, coverage increases from about 91% to about 93% as *m* increases. When *n* is smaller, coverage hovers around 91% for *n = 50* and 89 – 90% for *n=20* and does not change dramatically as *m* increases. This trend replicates itself with the varying *n* cases; for *nV1* coverage tends to range from about 85 – 87% to about 90% as *m* increases, while for *nV2* coverage tends to be between 90% and 92% with increasing *m*. For the LRR, coverage is much worse, however, for the rarer events case *(.10,.10)*. At its best, coverage here is about 90% for all values of *m* and I^2, when *n = 100*. For smaller values of *n* and in the *nV1* case, coverage drops considerably and decreases as *m* increases; this is a direct result of estimation bias.

Finally, the risk difference case is the most problematic. In the *(.50,.50)* and *(.30,.40)* cases, the trends are approximately similar to HTJ, particularly when *n=100*. For other values of *n*, coverage typically increases with *m* but often decreases as I^2 increases; for example, when *n = 20*, coverage here ranges from about 90% to 93% when $I^2 = 0$ but remains around 89% when $I^2 = 1/2$. In the small to moderate case shown here *(.10,.10)* and in other cases we investigated but do not report here, coverage can be quite low for small *n* and typically decreases as *m* increases and as I^2 increases. The trend of decreasing coverage in relation to increasing I^2 here has to do with a floor effect on plausible values for the underlying covariate. For example, when I^2 is large, it is possible to have very low study-specific probabilities (e.g. *0.02*), which severely restrict the possible values of the

covariate x_{ij}. As n increases, though, coverage even for this moderate to rare case, *(.10,.10),* is close to the more general HTJ trends.

Overall the simulation results here indicate that for all three effect sizes, when n is large the coverage trends are very similar to that found in the standardized mean difference case studied in HTJ. However, when n is small or varies considerably from study to study, coverage can be greatly reduced and varies by effect size. This variability in coverage has less to do with the variance estimator however, and more to do with estimation bias. The results shown here suggest that in these cases the log odds ratio is the most stable.

IV. Example

Landenberger and Lipsey (2005) report the results of a meta-analysis of $m = 58$ studies of the effects of cognitive-behavior therapy on the recidivism of adult and juvenile offenders. The outcome of interest was the log odds-ratio, which was analyzed using a random effects meta-analysis model. However, 22 of the 58 studies originally reported multiple outcomes; these studies had an average of $\bar{k} = 3$ outcomes (the overall average $\bar{k}$ = 1.75), which ranged from 2 to 7. Since the correlation structure of these estimates was unknown, in the original meta-analysis only one outcome per study was used. Tables 3.6 and 3.7 present a comparison of their results with an analysis based on the full data set with robust standard errors. The robust standard error models use equal within-study inverse-average-variance weights and estimate the between-study variation parameter τ^2 using the estimator provided in HTJ. As suggested, we vary ρ in the estimator of τ^2 and provide results here for both of the extreme values ($\rho = 0, 1$).

Tables 3.6 and 3.7 about here

Table 3.6 shows that when estimating the mean effect size (intercept) the results are robust to model specifications. That is, all estimates of the mean odds ratio are in the interval (*1.52, 1.57*); note that the estimates based on multiple outcomes with the studies tend to have larger log odds ratios. The t-test values are larger for the robust standard error models in general, but the p-values are all very small. Finally, the effect size variation is slightly larger in the dependent effects models. Depending upon the model used, the estimate of τ^2 is in the interval (*0.114, 0.133*) and does not vary much with ρ.

Table 3.7 shows the estimates for a model with several covariates. Overall, the results found in Landenberger and Lipsey (2005) are similar to those found with the RSE approach. The variables 'recidivism risk rating' and 'composite implementation factor' are the only two with $p<.05$ in the original study. The coefficient estimates are slightly smaller for both with the full, correlated data set and the p-values are larger. Models 2 and 3 also allow a new variable to be estimated; since the variable 'arrest recidivism' varies across outcomes within studies, it can be partitioned into a between-studies effect and a (centered) within studies effect. Neither of these effects are statistically significant however. Finally, note that there remains some effect size heterogeneity, with τ^2 estimated to be in (*.06,.09*) depending upon the model. Note that models in which $\rho = 1$ have larger estimated τ^2 parameters, though these differences are minimal.

V. Recommendations for using RSE for meta-analysis with binary data

The results of this paper show that in general robust variance estimation performs well. In the mean effect size (intercept) case, coverage is generally very close to nominal except when *n* is small or varies widely. Even in these cases, coverage increases from about 90% to about 94% as *m* increases from *10* to *40*. In the slope estimation case, coverage improves as *k*, *n*, and *m* increase. Here the *k=1* HTJ trend of 90%, 92%, and 93% for *m = 10, 20, 40* is found to occur here in a large number of cases. Additionally, when *k* > 1, coverage is often closer to 93% to 95% for all values of *m* and *n*.

In estimating the slope, the most problematic cases occur in the rare events and small *n* cases for the risk difference and in the varying *n* cases for the log odds ratio and log risk ratios. In most cases coverage does not fall below 90%. In cases in which small sample bias or rare events are likely to occur, these simulations suggest that using the log odds ratio addresses most of these concerns.

The fact that the trend shown here is similar to that found in HTJ suggests that there is an analytic correction for the *k=1* case that could improve coverage when the number of studies is small. Future work will address the development of a correction of this type.

Table 1.1 EBR and EVIF by numbers of equal population weighted strata

			Original EB	EBR				EVIF (ρ = 0)				EVIF (ρ = .10)				EVIF (ρ = .50)			
Cases			Mean																
	Sample	Population	Diff.	|SMD|	k= 2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
Generalization	N(8,2)	Chisq(10)	2	0.577	42.2%	54.4%	65.0%	70.6%	1.33	2.56	6.08	13.12	1.24	2.35	5.53	11.92	0.87	1.50	3.37	7.11
	N(9,2)		1	0.289	21.7%	31.3%	45.7%	52.8%	1.02	1.48	2.43	3.96	0.95	1.36	2.22	3.61	0.69	0.90	1.39	2.20
	N(9,3)		1	0.263	21.8%	32.0%	47.3%	54.8%	1.01	1.08	1.19	1.29	0.94	1.00	1.09	1.17	0.69	0.66	0.69	0.72
	N(9,1)		1	0.309	21.7%	27.6%	41.5%	48.6%	1.08	10.15	328.64	8991.73	1.01	9.26	297.64	8128.26	0.73	5.67	173.64	4674.38
	N(4,2)	Chisq(6)	2	0.707	42.3%	56.0%	66.3%	71.9%	1.33	1.86	2.90	4.27	1.24	1.71	2.64	3.89	0.88	1.10	1.63	2.35
	N(5,2)		1	0.354	22.1%	33.5%	47.4%	54.8%	1.02	1.18	1.44	1.76	0.95	1.08	1.32	1.61	0.69	0.72	0.83	0.99
	N(5,3)		1	0.309	22.2%	33.6%	47.9%	55.5%	1.01	1.02	1.05	1.08	0.94	0.94	0.96	0.98	0.69	0.62	0.60	0.60
	N(5,1)		1	0.392	22.1%	31.0%	43.4%	50.3%	1.08	3.86	29.08	216.18	1.01	3.54	26.41	195.78	0.73	2.23	15.71	114.18
	N(2,2)	Chisq(4)	2	0.816	42.6%	56.8%	66.9%	72.5%	1.34	1.61	2.05	2.52	1.25	1.48	1.87	2.29	0.88	0.95	1.16	1.39
	N(3,2)		1	0.408	22.7%	35.0%	48.4%	55.9%	1.02	1.07	1.17	1.28	0.96	0.99	1.08	1.17	0.69	0.65	0.68	0.72
	N(3,3)		1	0.343	22.7%	34.8%	48.4%	56.0%	1.08	2.33	8.07	29.49	1.01	2.13	7.35	26.76	0.73	1.37	4.45	15.83
	N(3,1)		1	0.471	22.6%	33.6%	45.6%	52.6%	1.01	1.03	1.06	1.09	0.94	0.95	0.97	0.99	0.69	0.61	0.60	0.59
Locational shift	N(2,2)	N(4,1)	2	1.265	61.9%	76.5%	83.8%	87.7%	1.87	2.19	2.35	2.43	1.74	2.00	2.13	2.20	1.19	1.23	1.26	1.28
	N(3,1)		1	1.000	61.9%	77.1%	85.3%	89.1%	1.87	2.10	2.28	2.36	1.74	1.92	2.08	2.15	1.19	1.21	1.27	1.28
	N(3.5,1)		0.5	0.500	63.2%	76.9%	85.9%	89.6%	1.17	1.20	1.24	1.25	1.09	1.11	1.13	1.14	0.78	0.72	0.70	0.69
	N(3.75,1)		0.25	0.25	63.5%	74.9%	86.0%	89.7%	1.04	1.04	1.06	1.06	0.97	0.96	0.96	0.96	0.71	0.63	0.60	0.58
	N(3,.75)	N(4,1)	2	1.131	60.7%	76.6%	85.0%	89.0%	3.02	4.43	6.04	7.26	2.78	4.05	5.49	6.59	1.83	2.51	3.32	3.93
	N(3,.5)		1	1.260	57.9%	74.5%	83.4%	87.8%	11.24	47.98	157.26	354.93	10.27	43.56	142.50	321.32	6.36	25.89	83.46	186.91
	N(3.5,.5)		0.5	0.632	61.9%	75.4%	84.6%	88.6%	1.87	3.74	7.36	12.03	1.74	3.42	6.70	10.93	1.19	2.14	4.06	6.52
	N(3.75,.5)		0.25	0.316	63.2%	73.8%	84.9%	88.8%	1.17	1.72	2.61	3.60	1.09	1.58	2.38	3.28	0.78	1.02	1.48	2.00

Note: $\rho = [(\rho_t^2 + \rho_c^2)/2]^{1/2}$, where $\rho_t = Corr(Y_t, logit[e(x)])$ and $\rho_c = Corr(Y_c, logit[e(x)])$ are the correlations between the logits of the propensity score distributions and Y_c and Y_t are the outcomes in the treatment and control groups

Table 1.2 The effect of moving to P_0 (from P) on EBR and the effect of additional subclassification on EBR and EVIF

			Original EB		Reduced EB			statification				EVIF (ρ= 0)				EVIF (ρ = .10)				EVIF (ρ = .50)			
Cases	Sample	Population	Mean Diff.	|SMD|	Mean Diff.	EBR	θ	k= 2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
Generalization	N(8,2)	Chisq(10)	2	0.577	0.003	99.9%	0.756	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	N(9,2)		1	0.289	-0.653	34.7%	0.811	63.2%	81.6%	81.9%	85.1%	1.07	1.28	1.42	1.54	1.00	1.18	1.30	1.40	0.72	0.77	0.82	0.86
	N(9,3)		1	0.263	-0.022	97.8%	0.900	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	N(9,1)		1	0.309	-1.525	-52.5%	0.667	54.9%	70.9%	76.8%	81.5%	3.28	23.67	97.59	306.46	3.02	21.52	88.47	277.47	1.98	12.90	51.98	161.51
	N(4,2)	Chisq(6)	2	0.707	0.672	66.4%	0.806	56.0%	70.0%	82.5%	87.5%	1.06	1.09	1.13	1.15	0.99	1.01	1.04	1.05	0.72	0.66	0.65	0.64
	N(5,2)		1	0.354	-0.048	95.2%	0.860	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	N(5,3)		1	0.309	0.426	57.4%	0.938	14.8%	26.5%	48.6%	58.7%	1.00	1.00	1.01	1.02	0.94	0.92	0.92	0.93	0.68	0.60	0.58	0.56
	N(5,1)		1	0.392	-0.780	22.0%	0.708	61.9%	78.6%	82.5%	86.4%	1.46	2.80	4.28	5.92	1.35	2.57	3.90	5.38	0.95	1.63	2.39	3.25
	N(2,2)	Chisq(4)	2	0.816	1.059	47.1%	0.845	52.7%	67.8%	78.5%	83.7%	1.13	1.17	1.22	1.24	1.06	1.08	1.12	1.13	0.76	0.70	0.69	0.68
	N(3,2)		1	0.408	0.287	71.3%	0.895	7.0%	19.9%	44.4%	55.7%	1.00	1.00	1.01	1.03	0.94	0.92	0.93	0.93	0.68	0.61	0.58	0.57
	N(3,3)		1	0.343	0.647	35.3%	0.959	22.5%	35.7%	52.4%	61.0%	1.11	1.34	1.47	1.56	1.04	1.23	1.35	1.42	0.75	0.81	0.85	0.87
	N(3,1)		1	0.471	-0.328	67.2%	0.745	78.1%	95.9%	94.5%	96.1%	1.00	1.03	1.06	1.09	0.94	0.95	0.97	0.99	0.68	0.61	0.59	0.59
Locational shift	N(2,2)	N(4,1)	2	1.265	1.988	0.6%	0.996	62.1%	76.7%	84.1%	87.9%	1.87	2.18	2.35	2.43	1.73	1.99	2.13	2.20	1.18	1.23	1.26	1.28
	N(3,1)		1	1.000	0.818	18.2%	0.908	67.3%	82.7%	90.5%	93.9%	1.64	1.74	1.81	1.83	1.52	1.59	1.65	1.66	1.05	1.01	1.01	0.99
	N(3.5,1)		0.5	0.500	0.422	15.6%	0.966	68.6%	82.5%	91.6%	95.0%	1.14	1.16	1.18	1.19	1.07	1.07	1.08	1.08	0.77	0.69	0.67	0.65
	N(3.75,1)		0.25	0.25	0.203	18.8%	0.981	70.9%	82.3%	94.2%	97.5%	1.03	1.04	1.04	1.04	0.97	0.95	0.95	0.95	0.70	0.62	0.59	0.57
	N(3,.75)	N(4,1)	1	1.131	0.608	39.2%	0.772	72.4%	87.9%	95.5%	98.3%	1.76	1.92	2.02	2.04	1.63	1.76	1.84	1.86	1.12	1.12	1.13	1.11
	N(3,.5)		1	1.260	0.303	69.7%	0.565	87.3%	98.1%	91.3%	90.0%	1.57	1.79	1.92	1.95	1.46	1.64	1.75	1.77	1.02	1.05	1.08	1.07
	N(3.5,.5)		0.5	0.632	0.071	85.8%	0.746	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	N(3.75,.5)		0.25	0.316	0.071	71.7%	0.819	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Note: (1) When the Reduced EB is very small, no additional subclassification estimator is needed; therefore for these values "NA" is reported; (2) The bias reduction shown here is in relation to the Reduced EB, not the Original EB.

Table 1.3 Example comparison of aveage bias in estimators

			PATE (1713 schools in population)						P_0ATE (1336 schools in sub-population)			
			(1) Conventional				(2) Subclass (3 strata)		(3) Conventional		(4) Subclass (3 strata)	
Variable	Level	Description	Experiment	Population	Mean Diff.	|SMD|	|SMD|	% EBR	|SMD|	% EBR	|SMD|	% EBR
CPSTTENA	Teacher	Teacher tenure (mean)	6.801	7.104	-0.303	0.116	0.062	46.33	0.081	30.01	0.023	73.87
CPSTEXPA	Teacher	Teacher experience (mean)	10.947	11.611	-0.664	0.221	0.014	93.59	0.205	7.45	0.024	89.79
CPSTKIDR	Teacher	Teacher-student ratio	13.266	12.674	0.592	0.196	0.060	69.44	0.126	35.80	0.044	69.69
CPSTBLFP	Teacher	Teachers that are African American (%)	2.557	8.652	-6.095	0.373	0.260	30.32	0.056	84.92	0.004	97.72
CPSTHIFP	Teacher	Teachers that are Hispanic (%)	21.571	14.416	7.156	0.302	0.070	76.74	0.238	21.18	0.014	93.57
CPSTTOFC	Teacher	Teachers in the school (total)	42.984	39.734	3.249	0.135	0.165	-21.69	0.102	24.55	0.155	-46.45
CPST00FP	Teacher	Teachers in first year of teaching (%)	8.739	8.303	0.436	0.059	0.042	28.98	0.097	-63.69	0.023	78.13
CPST01FP	Teacher	Teachers with 1-5 years experience (%)	28.744	27.979	0.765	0.062	0.051	17.97	0.086	-38.26	0.099	-5.95
CPST20FP	Teacher	Teachers with > 20 years experience (%)	17.695	20.364	-2.669	0.245	0.016	93.64	0.192	21.76	0.067	69.53
CPETDISP	Student	Students in disciplinary alternative education programs (%)	3.418	3.088	0.330	0.106	0.114	-7.33	0.138	-29.99	0.106	27.16
CPERRA7R	Student	7th grade retention (rate)	1.307	1.857	-0.550	0.096	0.146	-51.79	0.009	90.67	0.044	-87.51
CPEMALLP	Student	Students that are mobile (%)	14.804	19.425	-4.621	0.307	0.297	3.27	0.047	84.74	0.162	-27.93
CPETG07P	Student	Students in school that are in 7th grade (%)	34.995	31.037	3.957	0.292	0.008	97.28	0.204	30.22	0.023	89.56
CPETG07C	Student	Students in 7th grade (total)	224.247	188.892	35.355	0.226	0.105	53.43	0.186	17.65	0.075	61.47
CPETBLAP	Student	Students that are African American (%)	5.114	12.089	-6.975	0.431	0.368	14.49	0.102	76.24	0.095	55.18
CPETHISP	Student	Students that are Hispanic (%)	47.195	39.960	7.235	0.244	0.036	85.15	0.215	11.63	0.037	82.03
CPETLEPP	Student	Students that are LEP (%)	9.440	7.453	1.987	0.198	0.032	83.98	0.191	3.78	0.031	83.44
CPETECOP	Student	Students that are economically disadvantaged (%)	52.084	53.712	-1.629	0.067	0.117	-74.25	0.059	12.75	0.082	-33.58
CPETRSKP	Student	Students that are at risk (%)	40.607	43.595	-2.988	0.163	0.180	-10.06	0.034	79.02	0.030	25.47
CA007TR07R	Student	Students proficient in 7th grade reading (%)	86.000	81.718	4.282	0.201	0.174	13.53	0.047	76.81	0.080	17.1
CA007TM07R	Student	Students proficient in 7th grade math (%)	75.562	72.668	2.894	0.132	0.221	-66.78	0.096	27.43	0.030	79.93
CA311TM07R	Student	Students proficient in grades 3-11 math (%)	75.014	73.536	1.478	0.086	0.241	-181.8	0.131	-52.93	0.071	62.85
CA311TA07R	Student	Students proficient in grades 3-11 all (%)	63.836	63.262	0.573	0.035	0.202	-485.24	0.163	-372.91	0.002	99.17
CA311CM07R	Student	Students with commended performance, grades 3-11, math (%)	20.315	19.581	0.734	0.064	0.072	-11.91	0.100	-56.50	0.066	37.35
CA311CA07R	Student	Students with commended performance, grades 3-11, reading (%)	8.877	8.703	0.174	0.026	0.040	-50.92	0.109	-316.03	0.088	20.13
RURAL	School	County of school is rural	0.315	0.330	-0.015	0.032	0.221	-601.61	0.062	-96.86	0.075	-22.55
LOGIT (PS)	School	logits of propensity scores (mean)	2.625	4.116	-1.491	0.642	0.324	49.51	0.234	63.60	0.044	94.24
		Average			1.719	0.170	0.127	-29.048	0.118	-11.175	0.059	41.891
		|SMD| > 0.10				7	15		16		3	
		|SMD| > 0.20				6	8		4		0	
		|SMD| > 0.30				4	1		0		0	

Table 2.1 Parameters and distribution values used in simulations

Parameter		Values						
Corr[$Y_i(0), Y_i(1)$]	ρ	*-1,*	*-.5,*	*0,*	*.5,*	*0.9,*	*0.99,*	*1*
Std. Mean diff.	δ_A	*0,*	*.05,*	*.15,*	*.20,*	*...,*	*0.95*	*1*
Variance ratio	γ^2	*1,*	*1.1^2,*	*1.2^2,*	*1.3^2,*	*...,*	*1.9^2*	*2^2*
Sample size ($n_t=n_c$)	*n*	*25,*	*50,*	*100,*	*250,*	*500*		
Distribution of Y_{ic}	*N(c) =N(0,1),*			*LN(c) =Lognormal(a_c, b_c)*				
Distribution of Y_{it}	*N(t) =N(δ_A, γ^2),*			*LN(t) =Lognormal(a_t,b_t)*				
Combinations:	*N(c)N(t),*	*N(c)LN(t),*		*LN(c)LN(t)*				

Table 2.2 Example parameter and interval estimates

						ρ=0.90						
Subject	SchoolID	n_C	n_t	d_{Au}	95% CI (d_{Au})	s_{Au}^2	95% CI (s_{Au}^2)	T	95% CI (T)	CI $(T)_{inv}$	$T_{(1)}$	95% CI ($T_{(1)}$)
Reading	1	41	27	1.094	(.472,1.72)	0.283	(0.00,.623)	0.522	(.072,.973)	(.280,3.82)	1.914	(.262,3.57)
	2	60	32	0.502	(.049,.956)	0.186	(.089,.284)	0.908	(.032,1.78)	(.463,25.6)	1.102	(.039,2.16)
	3	185	49	0.480	(.174,.785)	0.181	(.162,.199)	0.910	(.310,1.51)	(.549,2.67)	1.099	(.375,1.82)
	4	58	28	-0.494	(-.936,-.051)	0.169	(.145,.193)	0.875	(.056,1.69)	(.472,13.5)	1.142	(.074,2.12)
	5	34	12	0.927	(-.108,1.96)	0.651	(0.00,1.88)	0.947	(0.00,2.18)	(.412, ∞)	1.056	(0.00,2.43)
	6	82	34	1.399	(.907,1.89)	0.223	(.063,.384)	0.355	(.134,.576)	(.219,.943)	2.818	(1.06,4.57)
	7	59	16	1.185	(.603,1.77)	0.159	(.105,.213)	0.368	(.150,.586)	(.231,.901)	2.717	(1.11,4.32)
	8	63	27	0.729	(.193,1.27)	0.249	(.006,.493)	0.728	(.023,1.43)	(.369,23.3)	1.374	(.043,2.71)
	9	86	44	-0.052	(-.399,.295)	0.176	(.162,.191)	8.301	(0.00,63.0)	(1.09, ∞)	0.120	(0.00,.915)
	10	75	32	0.366	(-.097,.829)	0.224	(.054,.394)	1.363	(0.00,3.22)	(.578, ∞)	0.734	(0.00,1.73)
Math	1	41	27	1.489	(.887,2.09)	0.182	(.061,.302)	0.307	(.130,.484)	(.195,.725)	3.254	(1.38,5.13)
	2	60	32	0.708	(.275,1.14)	0.171	(.146,.196)	0.611	(.201,1.02)	(.366,1.86)	1.637	(.539,2.73)
	3	185	49	0.324	(-.095,.743)	0.416	(.105,.728)	2.041	(0.00,4.84)	(.870, ∞)	0.490	(0.00,1.16)
	4	58	28	-0.367	(-.797,.064)	0.169	(.161,.177)	1.174	(0.00,2.60)	(.532, ∞)	0.852	(0.00,1.88)
	5	34	12	1.209	(.393,2.02)	0.205	(0.00,.549)	0.431	(0.00,.932)	(.199, ∞)	2.320	(0.00,5.02)
	6	82	34	1.211	(.679,1.74)	0.340	(.013,.666)	0.503	(.136,.870)	(.291,1.86)	1.989	(.537,3.44)
	7	59	16	2.163	(1.14,3.19)	0.983	(0.00,2.37)	0.480	(.148,.812)	(.284,1.55)	2.083	(.644,3.52)
	8	63	27	0.988	(.495,1.48)	0.180	(.092,.268)	0.456	(.175,.737)	(.282,1.19)	2.194	(.841,3.55)
	9	86	44	-0.047	(-.399,.305)	0.177	(.152,.203)	9.220	(0.00,77.4)	(1.10, ∞)	0.108	(0.00,.911)
	10	75	32	0.693	(.274,1.11)	0.174	(.136,.212)	0.630	(.219,1.04)	(.382,1.82)	1.587	(.551,2.62)

Notes: (1) CI (T)inv are confidence intervals created by inverting those for T(1). (2) For all parameters except dA, when a confidence interval lower bound is negative, we truncate the value to 0.

Table 3.1 Parameter values used in simulations

Parameter	Description	Values
m	number of studies	10, 20, 40
n	within study sample size	20, 50, 100, nV1, nV2
k	number of effect sizes within each study	1, 2, 5, 10, kV1, kV2
ρ	within study correlation between ES	0, .5, .8
I^2	between-study variability in π_j values	0, 1/3, 1/2
(π_T, π_C)	average study probabilities	(.10,.10), (.50,.50), (.30,.40)

Note: For n, k, and ρ for values given by a number, the value is assumed constant across all studies. For the set of values kV1, kV2, nV1, and nV2 these within-study values vary.

Table 3.2 Patterns for varying within-study parameter values

Parameter	Values (for m = 10)
nV1	20 20 20 20 20 20 100 100 100 100
nV2	20 20 40 40 60 60 80 80 100 100
kV1	1 1 1 1 2 2 5 5 10 10
kV2	1 1 1 1 1 1 1 1 10 10

Note: for m =20, 40, each of these sets of values is repeated 2 or 4 times respectively.

Table 3.3 Empirical coverage probabilities for 95% confidence intervals for the Risk Difference

			Intercept									Slope								
			(π_C,π_T) = (.10,.10)			(.50,.50)			(.30,.40)			(.10,.10)			(.50,.50)			(.30,.40)		
n	I^2	m	k=1	min	max	1	min	max	1	min	max	1	min	max	1	min	max	1	min	max
20	0	10	0.9403	0.9417	0.9456	0.9479	0.9479	0.9496	0.9462	0.9459	0.9501	0.8516	0.9150	0.9459	0.8980	0.9428	0.9578	0.8996	0.9401	0.9589
		20	0.9551	0.9501	0.9511	0.9494	0.9467	0.9480	0.9472	0.9476	0.9490	0.7619	0.8904	0.9250	0.9173	0.9431	0.9530	0.9140	0.9405	0.9518
		40	0.9688	0.9494	0.9531	0.9485	0.9469	0.9494	0.9497	0.9470	0.9502	0.6104	0.8467	0.9025	0.9263	0.9398	0.9487	0.9230	0.9430	0.9493
	0.333333	10	0.9523	0.9512	0.9531	0.9461	0.9460	0.9471	0.9468	0.9464	0.9497	0.7992	0.8793	0.9236	0.8990	0.9425	0.9585	0.8867	0.9357	0.9535
		20	0.9712	0.9556	0.9574	0.9476	0.9467	0.9501	0.9489	0.9474	0.9507	0.6864	0.8444	0.9113	0.9134	0.9389	0.9537	0.8992	0.9356	0.9498
		40	0.9820	0.9569	0.9645	0.9489	0.9488	0.9499	0.9483	0.9486	0.9512	0.5088	0.7785	0.8895	0.9170	0.9332	0.9502	0.9077	0.9361	0.9489
	0.5	10	0.9660	0.9586	0.9609	0.9435	0.9444	0.9458	0.9471	0.9459	0.9472	0.7626	0.8520	0.9040	0.8968	0.9430	0.9583	0.8809	0.9388	0.9556
		20	0.9814	0.9649	0.9692	0.9458	0.9468	0.9484	0.9442	0.9466	0.9493	0.6358	0.8079	0.8970	0.9083	0.9405	0.9512	0.8885	0.9326	0.9468
		40	0.9901	0.9630	0.9751	0.9497	0.9491	0.9501	0.9496	0.9472	0.9488	0.4604	0.7409	0.8833	0.9092	0.9327	0.9486	0.8885	0.9331	0.9486
50	0	10	0.9453	0.9474	0.9489	0.9502	0.9499	0.9522	0.9485	0.9500	0.9514	0.8995	0.9387	0.9559	0.9045	0.9447	0.9597	0.9085	0.9419	0.9587
		20	0.9458	0.9482	0.9503	0.9486	0.9486	0.9492	0.9492	0.9485	0.9502	0.9014	0.9360	0.9448	0.9245	0.9410	0.9528	0.9260	0.9430	0.9515
		40	0.9486	0.9479	0.9496	0.9478	0.9486	0.9505	0.9496	0.9476	0.9505	0.8839	0.9266	0.9365	0.9330	0.9410	0.9508	0.9374	0.9448	0.9491
	0.333333	10	0.9448	0.9474	0.9522	0.9483	0.9478	0.9497	0.9483	0.9482	0.9502	0.8727	0.9254	0.9467	0.8998	0.9417	0.9607	0.8976	0.9403	0.9576
		20	0.9454	0.9462	0.9526	0.9483	0.9485	0.9511	0.9503	0.9469	0.9509	0.8534	0.9171	0.9373	0.9164	0.9456	0.9525	0.9141	0.9419	0.9518
		40	0.9489	0.9477	0.9498	0.9496	0.9507	0.9526	0.9479	0.9477	0.9502	0.8060	0.9016	0.9298	0.9224	0.9401	0.9519	0.9281	0.9434	0.9513
	0.5	10	0.9480	0.9488	0.9500	0.9466	0.9462	0.9501	0.9469	0.9478	0.9507	0.8470	0.9119	0.9374	0.8993	0.9458	0.9597	0.8924	0.9415	0.9554
		20	0.9498	0.9484	0.9510	0.9487	0.9478	0.9489	0.9491	0.9487	0.9511	0.8158	0.8992	0.9329	0.9180	0.9451	0.9547	0.9099	0.9406	0.9492
		40	0.9521	0.9507	0.9523	0.9472	0.9482	0.9502	0.9486	0.9476	0.9508	0.7425	0.8766	0.9240	0.9219	0.9415	0.9513	0.9215	0.9393	0.9489
100	0	10	0.9500	0.9475	0.9500	0.9493	0.9467	0.9505	0.9508	0.9484	0.9509	0.9098	0.9409	0.9569	0.9042	0.9440	0.9587	0.9102	0.9433	0.9589
		20	0.9478	0.9499	0.9517	0.9500	0.9483	0.9504	0.9485	0.9474	0.9514	0.9230	0.9395	0.9500	0.9237	0.9441	0.9510	0.9285	0.9408	0.9533
		40	0.9525	0.9483	0.9499	0.9494	0.9486	0.9506	0.9499	0.9482	0.9509	0.9231	0.9410	0.9460	0.9339	0.9455	0.9496	0.9394	0.9450	0.9517
	0.333333	10	0.9493	0.9491	0.9509	0.9485	0.9496	0.9517	0.9495	0.9492	0.9500	0.8956	0.9361	0.9532	0.9017	0.9441	0.9617	0.8991	0.9409	0.9577
		20	0.9478	0.9477	0.9526	0.9512	0.9464	0.9500	0.9461	0.9489	0.9499	0.9002	0.9329	0.9460	0.9205	0.9445	0.9530	0.9209	0.9420	0.9500
		40	0.9516	0.9493	0.9513	0.9501	0.9470	0.9510	0.9499	0.9493	0.9504	0.8915	0.9270	0.9431	0.9288	0.9441	0.9516	0.9338	0.9451	0.9497
	0.5	10	0.9492	0.9485	0.9506	0.9501	0.9499	0.9513	0.9487	0.9494	0.9509	0.8806	0.9264	0.9495	0.9028	0.9479	0.9603	0.8977	0.9416	0.9566
		20	0.9473	0.9499	0.9514	0.9502	0.9498	0.9512	0.9509	0.9482	0.9498	0.8748	0.9219	0.9415	0.9220	0.9467	0.9548	0.9144	0.9417	0.9505
		40	0.9493	0.9491	0.9503	0.9487	0.9501	0.9516	0.9505	0.9468	0.9496	0.8559	0.9143	0.9381	0.9269	0.9411	0.9523	0.9302	0.9447	0.9479
nV1	0	10	0.9319	0.9207	0.9229	0.9232	0.9206	0.9231	0.9226	0.9190	0.9238	0.8475	0.9119	0.9323	0.8729	0.9178	0.9358	0.8765	0.9235	0.9369
		20	0.9448	0.9337	0.9354	0.9362	0.9332	0.9364	0.9356	0.9332	0.9352	0.8544	0.9139	0.9326	0.9008	0.9235	0.9404	0.9008	0.9275	0.9389
		40	0.9458	0.9419	0.9425	0.9429	0.9395	0.9413	0.9414	0.9402	0.9425	0.8475	0.9170	0.9289	0.9182	0.9333	0.9442	0.9205	0.9350	0.9426
	0.333333	10	0.9142	0.8937	0.9052	0.9027	0.8938	0.8971	0.9028	0.8942	0.9006	0.8086	0.8856	0.9125	0.8479	0.9048	0.9236	0.8453	0.8972	0.9264
		20	0.9350	0.9196	0.9244	0.9238	0.9191	0.9199	0.9222	0.9195	0.9222	0.8120	0.8920	0.9157	0.8760	0.9152	0.9308	0.8761	0.9122	0.9290
		40	0.9425	0.9343	0.9383	0.9387	0.9311	0.9363	0.9370	0.9326	0.9396	0.7860	0.8910	0.9221	0.9002	0.9274	0.9355	0.9030	0.9219	0.9371
	0.5	10	0.9119	0.8914	0.8992	0.8931	0.8833	0.8931	0.8985	0.8856	0.8899	0.7894	0.8749	0.9008	0.8340	0.9017	0.9224	0.8302	0.8920	0.9175
		20	0.9341	0.9190	0.9245	0.9209	0.9152	0.9203	0.9234	0.9157	0.9197	0.7831	0.8799	0.9069	0.8714	0.9151	0.9306	0.8643	0.9109	0.9265
		40	0.9435	0.9346	0.9388	0.9365	0.9330	0.9344	0.9363	0.9331	0.9355	0.7580	0.8734	0.9117	0.8996	0.9302	0.9353	0.8953	0.9248	0.9374
nV2	0	10	0.9448	0.9427	0.9452	0.9446	0.9420	0.9428	0.9427	0.9435	0.9462	0.8855	0.9296	0.9519	0.8973	0.9347	0.9551	0.8987	0.9343	0.9524
		20	0.9498	0.9451	0.9472	0.9472	0.9463	0.9472	0.9459	0.9470	0.9484	0.8888	0.9301	0.9445	0.9188	0.9407	0.9504	0.9220	0.9382	0.9500
		40	0.9476	0.9474	0.9484	0.9463	0.9475	0.9496	0.9493	0.9485	0.9495	0.8734	0.9261	0.9365	0.9280	0.9443	0.9472	0.9353	0.9411	0.9501
	0.333333	10	0.9428	0.9390	0.9411	0.9394	0.9359	0.9404	0.9402	0.9368	0.9412	0.8591	0.9194	0.9378	0.8879	0.9366	0.9554	0.8852	0.9346	0.9509
		20	0.9473	0.9431	0.9454	0.9451	0.9424	0.9442	0.9439	0.9425	0.9450	0.8480	0.9132	0.9326	0.9116	0.9373	0.9489	0.9056	0.9347	0.9468
		40	0.9479	0.9470	0.9499	0.9477	0.9434	0.9485	0.9475	0.9461	0.9466	0.8072	0.8995	0.9266	0.9250	0.9360	0.9457	0.9214	0.9388	0.9453
	0.5	10	0.9395	0.9372	0.9409	0.9359	0.9365	0.9374	0.9386	0.9361	0.9407	0.8341	0.9060	0.9342	0.8832	0.9356	0.9561	0.8744	0.9330	0.9478
		20	0.9490	0.9427	0.9470	0.9423	0.9433	0.9472	0.9429	0.9415	0.9429	0.8066	0.8989	0.9247	0.9068	0.9374	0.9500	0.8993	0.9339	0.9437
		40	0.9492	0.9461	0.9485	0.9465	0.9456	0.9462	0.9434	0.9443	0.9456	0.7483	0.8836	0.9236	0.9168	0.9358	0.9472	0.9169	0.9373	0.9460

Table 3.4 Empirical coverage probabilities for 95% confidence intervals for the Log Risk Ratio

			Intercept									Slope								
			(π_C,π_T) = (.10,.10)			(.50,.50)			(.30,.40)			(.10,.10)			(.50,.50)			(.30,.40)		
n	I^2	m	k=1	min	max	1	min	max	1	min	max	1	min	max	1	min	max	1	min	max
20	0	10	0.9503	0.9485	0.9507	0.9498	0.9496	0.9502	0.9485	0.9505	0.9520	0.8938	0.8906	0.9177	0.9005	0.9316	0.9431	0.9083	0.9357	0.9479
		20	0.9494	0.9479	0.9495	0.9527	0.9500	0.9534	0.9507	0.9503	0.9523	0.8854	0.8008	0.8868	0.8986	0.9127	0.9145	0.9057	0.9187	0.9257
		40	0.9513	0.9488	0.9510	0.9489	0.9488	0.9514	0.9500	0.9504	0.9511	0.8315	0.6267	0.8166	0.8638	0.8700	0.8721	0.8746	0.8845	0.8917
	0.333	10	0.9415	0.9447	0.9465	0.9507	0.9494	0.9508	0.9476	0.9471	0.9502	0.8920	0.8884	0.9214	0.8984	0.9338	0.9443	0.9086	0.9349	0.9494
		20	0.9469	0.9450	0.9467	0.9523	0.9488	0.9523	0.9477	0.9479	0.9502	0.8923	0.8042	0.8933	0.8980	0.9141	0.9206	0.9086	0.9220	0.9308
		40	0.9466	0.9465	0.9496	0.9490	0.9486	0.9507	0.9509	0.9504	0.9516	0.8540	0.6507	0.8385	0.8682	0.8753	0.8806	0.8781	0.8944	0.8973
	0.5	10	0.9364	0.9383	0.9407	0.9501	0.9476	0.9502	0.9482	0.9466	0.9488	0.8848	0.8851	0.9178	0.8975	0.9328	0.9456	0.9038	0.9365	0.9491
		20	0.9431	0.9423	0.9431	0.9497	0.9490	0.9505	0.9498	0.9477	0.9504	0.8886	0.8059	0.8957	0.9020	0.9146	0.9232	0.9080	0.9205	0.9301
		40	0.9448	0.9451	0.9476	0.9502	0.9492	0.9510	0.9483	0.9497	0.9505	0.8636	0.6632	0.8459	0.8776	0.8798	0.8908	0.8810	0.8964	0.9000
50	0	10	0.9527	0.9504	0.9513	0.9509	0.9503	0.9516	0.9486	0.9485	0.9518	0.9030	0.9333	0.9426	0.9110	0.9412	0.9539	0.9144	0.9442	0.9548
		20	0.9509	0.9482	0.9510	0.9496	0.9484	0.9526	0.9484	0.9485	0.9514	0.8996	0.9099	0.9126	0.9163	0.9312	0.9395	0.9209	0.9356	0.9435
		40	0.9497	0.9487	0.9517	0.9521	0.9493	0.9520	0.9499	0.9497	0.9527	0.8673	0.8622	0.8723	0.9057	0.9187	0.9249	0.9131	0.9254	0.9331
	0.333	10	0.9492	0.9476	0.9498	0.9514	0.9504	0.9510	0.9505	0.9476	0.9508	0.8999	0.9346	0.9420	0.9105	0.9414	0.9550	0.9152	0.9462	0.9570
		20	0.9505	0.9495	0.9507	0.9493	0.9483	0.9501	0.9491	0.9506	0.9511	0.9042	0.9125	0.9199	0.9183	0.9333	0.9433	0.9231	0.9374	0.9472
		40	0.9513	0.9487	0.9525	0.9507	0.9505	0.9515	0.9510	0.9466	0.9519	0.8802	0.8715	0.8876	0.9109	0.9215	0.9301	0.9148	0.9270	0.9344
	0.5	10	0.9459	0.9459	0.9496	0.9489	0.9498	0.9503	0.9483	0.9473	0.9498	0.9001	0.9338	0.9449	0.9133	0.9452	0.9547	0.9097	0.9469	0.9568
		20	0.9483	0.9466	0.9501	0.9500	0.9497	0.9511	0.9494	0.9490	0.9501	0.9055	0.9120	0.9226	0.9183	0.9357	0.9437	0.9243	0.9378	0.9450
		40	0.9474	0.9483	0.9502	0.9492	0.9500	0.9513	0.9508	0.9489	0.9503	0.8897	0.8748	0.8918	0.9146	0.9227	0.9298	0.9191	0.9262	0.9344
100	0	10	0.9493	0.9492	0.9533	0.9497	0.9486	0.9512	0.9499	0.9517	0.9525	0.9121	0.9392	0.9539	0.9179	0.9420	0.9559	0.9152	0.9444	0.9580
		20	0.9527	0.9484	0.9515	0.9504	0.9473	0.9507	0.9499	0.9491	0.9505	0.9150	0.9314	0.9372	0.9276	0.9416	0.9484	0.9270	0.9427	0.9495
		40	0.9480	0.9496	0.9508	0.9504	0.9497	0.9505	0.9514	0.9488	0.9519	0.9011	0.9083	0.9157	0.9281	0.9331	0.9383	0.9288	0.9364	0.9420
	0.333	10	0.9500	0.9499	0.9512	0.9493	0.9489	0.9506	0.9500	0.9482	0.9493	0.9122	0.9407	0.9530	0.9141	0.9446	0.9559	0.9163	0.9431	0.9559
		20	0.9516	0.9501	0.9503	0.9516	0.9479	0.9522	0.9497	0.9495	0.9518	0.9200	0.9315	0.9402	0.9268	0.9405	0.9499	0.9272	0.9435	0.9503
		40	0.9501	0.9493	0.9506	0.9505	0.9487	0.9512	0.9503	0.9511	0.9513	0.9081	0.9141	0.9200	0.9273	0.9318	0.9407	0.9287	0.9357	0.9445
	0.5	10	0.9512	0.9489	0.9508	0.9515	0.9484	0.9512	0.9506	0.9473	0.9510	0.9109	0.9401	0.9522	0.9144	0.9469	0.9576	0.9141	0.9449	0.9585
		20	0.9517	0.9492	0.9523	0.9512	0.9492	0.9517	0.9494	0.9489	0.9504	0.9171	0.9308	0.9387	0.9280	0.9427	0.9488	0.9303	0.9432	0.9506
		40	0.9498	0.9492	0.9496	0.9534	0.9486	0.9497	0.9511	0.9493	0.9522	0.9079	0.9158	0.9220	0.9246	0.9353	0.9428	0.9279	0.9397	0.9433
nV1	0	10	0.9103	0.9056	0.9116	0.9191	0.9148	0.9173	0.9190	0.9159	0.9178	0.8606	0.8989	0.9078	0.8737	0.9106	0.9254	0.8734	0.9084	0.9253
		20	0.9273	0.9251	0.9278	0.9311	0.9285	0.9319	0.9327	0.9283	0.9338	0.8829	0.9020	0.9064	0.8936	0.9136	0.9272	0.8943	0.9152	0.9294
		40	0.9374	0.9369	0.9396	0.9426	0.9389	0.9409	0.9421	0.9393	0.9403	0.8840	0.8796	0.8944	0.9028	0.9135	0.9236	0.9053	0.9167	0.9294
	0.333	10	0.8950	0.8823	0.8869	0.8990	0.8904	0.8932	0.8994	0.8880	0.8959	0.8435	0.8902	0.9028	0.8573	0.8986	0.9152	0.8511	0.8984	0.9185
		20	0.9209	0.9118	0.9186	0.9214	0.9181	0.9221	0.9214	0.9161	0.9204	0.8703	0.8988	0.9032	0.8838	0.9090	0.9205	0.8818	0.9084	0.9212
		40	0.9381	0.9320	0.9366	0.9324	0.9335	0.9351	0.9339	0.9321	0.9371	0.8809	0.8867	0.8998	0.8975	0.9148	0.9231	0.9004	0.9163	0.9272
	0.5	10	0.8875	0.8749	0.8809	0.8909	0.8837	0.8882	0.8917	0.8818	0.8833	0.8384	0.8870	0.9017	0.8472	0.8961	0.9149	0.8459	0.8933	0.9173
		20	0.9142	0.9125	0.9136	0.9212	0.9159	0.9203	0.9178	0.9160	0.9181	0.8673	0.8980	0.9072	0.8795	0.9073	0.9201	0.8747	0.9038	0.9226
		40	0.9330	0.9270	0.9333	0.9327	0.9317	0.9325	0.9372	0.9306	0.9336	0.8857	0.8932	0.9028	0.9018	0.9167	0.9237	0.9013	0.9178	0.9286
nV2	0	10	0.9411	0.9391	0.9423	0.9422	0.9435	0.9451	0.9420	0.9416	0.9437	0.8938	0.9266	0.9372	0.9032	0.9329	0.9471	0.9054	0.9352	0.9489
		20	0.9424	0.9459	0.9468	0.9464	0.9457	0.9473	0.9448	0.9453	0.9471	0.9025	0.9138	0.9194	0.9146	0.9291	0.9392	0.9179	0.9306	0.9401
		40	0.9461	0.9466	0.9478	0.9474	0.9465	0.9481	0.9460	0.9471	0.9490	0.8809	0.8794	0.8878	0.9111	0.9204	0.9285	0.9171	0.9244	0.9327
	0.333	10	0.9398	0.9352	0.9376	0.9417	0.9374	0.9383	0.9395	0.9362	0.9398	0.8914	0.9236	0.9368	0.8989	0.9317	0.9477	0.9007	0.9332	0.9459
		20	0.9442	0.9419	0.9432	0.9431	0.9435	0.9451	0.9443	0.9408	0.9423	0.8980	0.9137	0.9202	0.9078	0.9297	0.9403	0.9170	0.9328	0.9397
		40	0.9437	0.9457	0.9470	0.9488	0.9456	0.9474	0.9470	0.9462	0.9489	0.8851	0.8791	0.8963	0.9137	0.9241	0.9309	0.9140	0.9272	0.9367
	0.5	10	0.9368	0.9308	0.9342	0.9384	0.9364	0.9382	0.9387	0.9335	0.9347	0.8924	0.9227	0.9354	0.8990	0.9336	0.9435	0.8982	0.9335	0.9470
		20	0.9401	0.9392	0.9420	0.9453	0.9406	0.9442	0.9409	0.9410	0.9425	0.9013	0.9121	0.9206	0.9123	0.9259	0.9363	0.9153	0.9340	0.9438
		40	0.9458	0.9443	0.9447	0.9448	0.9460	0.9476	0.9485	0.9465	0.9470	0.8950	0.8896	0.9007	0.9164	0.9225	0.9298	0.9164	0.9286	0.9352

Table 3.5 Empirical coverage probabilities for 95% confidence intervals for the Log Odds Ratio

			Intercept									Slope								
			(π_C,π_T) = (.10,.10)			(.50,.50)			(.30,.40)			(.10,.10)			(.50,.50)			(.30,.40)		
n	I^2	m	k=1	min	max	1	min	max	1	min	max	1	min	max	1	min	max	1	min	max
20	0	10	0.9523	0.9502	0.9510	0.9548	0.9515	0.9535	0.9529	0.9507	0.9516	0.8968	0.9330	0.9509	0.9036	0.9349	0.9582	0.9000	0.9370	0.9579
		20	0.9495	0.9513	0.9532	0.9513	0.9490	0.9528	0.9518	0.9507	0.9523	0.9127	0.9304	0.9417	0.9222	0.9419	0.9524	0.9190	0.9416	0.9524
		40	0.9512	0.9491	0.9529	0.9512	0.9493	0.9521	0.9504	0.9504	0.9521	0.9151	0.9197	0.9273	0.9340	0.9436	0.9513	0.9333	0.9445	0.9509
	0.333	10	0.9462	0.9466	0.9480	0.9524	0.9500	0.9527	0.9519	0.9503	0.9516	0.8882	0.9291	0.9454	0.9022	0.9404	0.9605	0.9049	0.9363	0.9573
		20	0.9502	0.9465	0.9509	0.9521	0.9512	0.9516	0.9505	0.9511	0.9526	0.9103	0.9284	0.9346	0.9245	0.9415	0.9546	0.9196	0.9422	0.9508
		40	0.9490	0.9493	0.9506	0.9492	0.9498	0.9516	0.9516	0.9507	0.9527	0.9142	0.9107	0.9245	0.9326	0.9442	0.9523	0.9307	0.9435	0.9519
	0.5	10	0.9434	0.9434	0.9458	0.9531	0.9515	0.9551	0.9518	0.9518	0.9531	0.8886	0.9267	0.9429	0.9083	0.9369	0.9607	0.9039	0.9387	0.9567
		20	0.9495	0.9457	0.9471	0.9515	0.9503	0.9534	0.9506	0.9515	0.9519	0.9040	0.9221	0.9314	0.9253	0.9415	0.9543	0.9229	0.9428	0.9517
		40	0.9464	0.9491	0.9503	0.9494	0.9495	0.9526	0.9529	0.9488	0.9504	0.9114	0.9070	0.9248	0.9345	0.9423	0.9535	0.9327	0.9435	0.9496
50	0	10	0.9521	0.9501	0.9531	0.9501	0.9490	0.9498	0.9497	0.9477	0.9515	0.9009	0.9375	0.9566	0.9026	0.9348	0.9580	0.9016	0.9387	0.9585
		20	0.9495	0.9505	0.9535	0.9512	0.9482	0.9509	0.9509	0.9502	0.9522	0.9183	0.9376	0.9511	0.9221	0.9402	0.9520	0.9219	0.9430	0.9549
		40	0.9516	0.9496	0.9515	0.9510	0.9485	0.9513	0.9503	0.9493	0.9513	0.9279	0.9355	0.9437	0.9351	0.9448	0.9493	0.9327	0.9460	0.9526
	0.333	10	0.9522	0.9491	0.9510	0.9502	0.9481	0.9521	0.9503	0.9487	0.9515	0.8996	0.9358	0.9577	0.9017	0.9373	0.9597	0.9028	0.9405	0.9572
		20	0.9505	0.9500	0.9527	0.9503	0.9500	0.9503	0.9479	0.9500	0.9511	0.9119	0.9378	0.9487	0.9216	0.9386	0.9523	0.9206	0.9430	0.9526
		40	0.9516	0.9501	0.9509	0.9511	0.9487	0.9506	0.9514	0.9502	0.9517	0.9223	0.9319	0.9449	0.9355	0.9450	0.9523	0.9340	0.9443	0.9508
	0.5	10	0.9495	0.9467	0.9489	0.9516	0.9501	0.9520	0.9513	0.9494	0.9504	0.8967	0.9322	0.9540	0.9005	0.9367	0.9600	0.9042	0.9414	0.9605
		20	0.9494	0.9491	0.9519	0.9507	0.9487	0.9537	0.9480	0.9500	0.9511	0.9147	0.9345	0.9484	0.9230	0.9416	0.9530	0.9206	0.9436	0.9524
		40	0.9479	0.9492	0.9512	0.9501	0.9496	0.9520	0.9516	0.9498	0.9513	0.9265	0.9349	0.9428	0.9334	0.9462	0.9532	0.9334	0.9440	0.9515
100	0	10	0.9537	0.9486	0.9533	0.9484	0.9481	0.9520	0.9499	0.9469	0.9511	0.9030	0.9385	0.9564	0.9021	0.9365	0.9572	0.8976	0.9357	0.9563
		20	0.9510	0.9500	0.9521	0.9498	0.9485	0.9512	0.9512	0.9500	0.9504	0.9226	0.9396	0.9506	0.9205	0.9395	0.9542	0.9227	0.9391	0.9524
		40	0.9487	0.9503	0.9529	0.9493	0.9471	0.9507	0.9506	0.9490	0.9514	0.9297	0.9390	0.9476	0.9349	0.9440	0.9480	0.9352	0.9429	0.9514
	0.333	10	0.9513	0.9493	0.9518	0.9510	0.9492	0.9501	0.9505	0.9487	0.9510	0.9028	0.9363	0.9551	0.9011	0.9351	0.9606	0.9023	0.9360	0.9580
		20	0.9480	0.9500	0.9531	0.9507	0.9492	0.9517	0.9503	0.9502	0.9511	0.9218	0.9389	0.9509	0.9214	0.9413	0.9536	0.9224	0.9381	0.9520
		40	0.9539	0.9507	0.9528	0.9496	0.9509	0.9520	0.9495	0.9493	0.9504	0.9289	0.9408	0.9474	0.9313	0.9443	0.9530	0.9370	0.9402	0.9512
	0.5	10	0.9521	0.9495	0.9524	0.9513	0.9481	0.9505	0.9507	0.9506	0.9516	0.9037	0.9351	0.9578	0.9040	0.9368	0.9588	0.9031	0.9369	0.9605
		20	0.9499	0.9487	0.9507	0.9492	0.9500	0.9516	0.9499	0.9486	0.9505	0.9206	0.9373	0.9521	0.9232	0.9383	0.9536	0.9245	0.9400	0.9519
		40	0.9518	0.9494	0.9519	0.9504	0.9499	0.9519	0.9502	0.9495	0.9508	0.9309	0.9403	0.9491	0.9360	0.9410	0.9522	0.9341	0.9425	0.9506
nV1	0	10	0.9094	0.9062	0.9102	0.9179	0.9182	0.9188	0.9153	0.9152	0.9176	0.8586	0.8981	0.9214	0.8662	0.9098	0.9313	0.8658	0.9052	0.9298
		20	0.9331	0.9262	0.9297	0.9315	0.9291	0.9341	0.9305	0.9291	0.9325	0.8830	0.9067	0.9269	0.8944	0.9189	0.9329	0.8872	0.9126	0.9356
		40	0.9366	0.9373	0.9386	0.9395	0.9404	0.9424	0.9413	0.9410	0.9423	0.9030	0.9190	0.9318	0.9158	0.9299	0.9400	0.9084	0.9274	0.9393
	0.333	10	0.8917	0.8795	0.8893	0.8992	0.8867	0.8969	0.8960	0.8878	0.8951	0.8349	0.8821	0.9100	0.8415	0.8890	0.9212	0.8448	0.8889	0.9196
		20	0.9205	0.9159	0.9180	0.9223	0.9179	0.9209	0.9236	0.9169	0.9224	0.8682	0.8984	0.9227	0.8732	0.9049	0.9286	0.8736	0.9070	0.9281
		40	0.9336	0.9321	0.9336	0.9366	0.9338	0.9375	0.9353	0.9329	0.9354	0.8986	0.9173	0.9280	0.9014	0.9240	0.9365	0.9042	0.9231	0.9356
	0.5	10	0.8848	0.8747	0.8806	0.8918	0.8854	0.8869	0.8931	0.8819	0.8878	0.8283	0.8778	0.9051	0.8332	0.8825	0.9154	0.8293	0.8876	0.9184
		20	0.9163	0.9127	0.9148	0.9218	0.9169	0.9183	0.9160	0.9129	0.9187	0.8683	0.8954	0.9193	0.8667	0.9026	0.9274	0.8668	0.9051	0.9281
		40	0.9347	0.9310	0.9332	0.9361	0.9313	0.9329	0.9349	0.9326	0.9345	0.8908	0.9155	0.9312	0.9008	0.9199	0.9350	0.8981	0.9211	0.9378
nV2	0	10	0.9440	0.9407	0.9442	0.9423	0.9427	0.9448	0.9432	0.9415	0.9439	0.8929	0.9241	0.9501	0.8897	0.9289	0.9534	0.8906	0.9282	0.9524
		20	0.9462	0.9446	0.9470	0.9458	0.9444	0.9449	0.9449	0.9460	0.9494	0.9093	0.9299	0.9468	0.9174	0.9348	0.9483	0.9140	0.9338	0.9478
		40	0.9471	0.9460	0.9476	0.9486	0.9463	0.9485	0.9464	0.9466	0.9483	0.9207	0.9347	0.9440	0.9331	0.9414	0.9515	0.9296	0.9384	0.9478
	0.333	10	0.9397	0.9372	0.9380	0.9393	0.9382	0.9398	0.9410	0.9365	0.9373	0.8861	0.9239	0.9486	0.8833	0.9277	0.9537	0.8883	0.9255	0.9500
		20	0.9440	0.9419	0.9442	0.9441	0.9428	0.9441	0.9441	0.9432	0.9458	0.9028	0.9294	0.9421	0.9125	0.9343	0.9478	0.9119	0.9331	0.9481
		40	0.9481	0.9470	0.9488	0.9459	0.9429	0.9465	0.9484	0.9451	0.9469	0.9173	0.9348	0.9425	0.9300	0.9388	0.9483	0.9277	0.9387	0.9499
	0.5	10	0.9351	0.9347	0.9355	0.9376	0.9350	0.9400	0.9389	0.9385	0.9406	0.8800	0.9250	0.9465	0.8848	0.9244	0.9509	0.8883	0.9257	0.9522
		20	0.9415	0.9407	0.9422	0.9429	0.9429	0.9439	0.9425	0.9424	0.9434	0.9016	0.9255	0.9434	0.9086	0.9298	0.9468	0.9102	0.9306	0.9477
		40	0.9466	0.9448	0.9470	0.9472	0.9463	0.9471	0.9464	0.9462	0.9471	0.9170	0.9313	0.9415	0.9222	0.9380	0.9509	0.9262	0.9384	0.9490

Table 3.6 Example model comparisons, estimate of mean effect 100

Model	Outcomes	weights	Est	SE	t-test	Pr(|t|>)	τ^2
Model 0 (LL)	58	M-IVW	0.425	(.062)	6.842*	<0.001	0.114
Model 1	58	R-IVW	0.429	(.061)	7.025	<0.001	0.123
Model 2 (ρ=0)	102	R-IVW	0.449	(.061)	7.337	<.001	0.133
Model 3 (ρ=1)	102	R-IVW	0.449	(.061)	7.338	<.001	0.133

** For model standard errors, a z-test is used instead.*

Table 3.7 Example comparison of estimates using different weights and correlation values

Variable	Model 0 (M-IVW)*				Model 1 (R-IVW)				Model 2 (R-IVW,ρ=0)				Model 3 (R-IVW,ρ=1)			
	Est	SE	z-test	Pr(|z|>)	Est	SE	t-test	Pr(|t|>)	Est	SE	t-test	Pr(|t|>)	Est	SE	t-test	Pr(|t|>)
Intercept	NR	NR	NR	NR	0.23	(.62)	0.36	0.72	0.15	(.6)	0.25	0.80	0.13	(.59)	0.22	0.83
Design problem	0.11	(.11)	1.02	0.31	0.12	(.11)	1.12	0.27	0.14	(.12)	1.14	0.26	0.14	(.12)	1.17	0.25
Attrition proportion	-0.13	(.62)	-0.21	0.83	-0.12	(.59)	-0.21	0.83	0.01	(.62)	0.02	0.98	0.01	(.62)	0.01	0.99
Intent to Treat	-0.13	(.11)	-1.21	0.23	-0.12	(.09)	-1.31	0.20	-0.12	(.08)	-1.48	0.15	-0.12	(.08)	-1.47	0.15
Arrest Recidivism (m)	0.13	(.13)	1.04	0.30	0.14	(.15)	0.99	0.33	0.15	(.14)	1.14	0.26	0.16	(.14)	1.16	0.25
Arrest Recidivism (w)	--	--	--		--	---	---	---	0.02	(.13)	0.17	0.86	0.02	(.13)	0.17	0.86
Recidivism risk rating	0.19	(.1)	1.99	0.05	0.19	(.09)	2.12	0.04	0.15	(.09)	1.64	0.11	0.15	(.09)	1.65	0.11
Sessions per week	0.05	(.04)	1.21	0.23	0.05	(.04)	1.13	0.27	0.05	(.04)	1.13	0.26	0.05	(.05)	1.15	0.26
Length in weeks (logged)	0.04	(.11)	0.36	0.72	0.04	(.12)	0.36	0.72	0.06	(.13)	0.48	0.63	0.06	(.13)	0.50	0.62
Sessions x length	0.03	(.04)	0.73	0.47	0.03	(.05)	0.60	0.55	0.00	(.05)	0.09	0.93	0.00	(.05)	0.08	0.94
Composite implementation factor	0.26	(.09)	2.93	0.00	0.25	(.11)	2.21	0.03	0.23	(.11)	2.13	0.04	0.23	(.11)	2.14	0.04
CBT emphasis	-0.10	(.11)	-0.90	0.37	-0.08	(.15)	-0.52	0.61	-0.04	(.13)	-0.33	0.74	-0.04	(.13)	-0.30	0.76
Reasoning and Rehabilitation	-0.01	(.1)	-0.10	0.92	0.00	(.18)	-0.03	0.98	0.00	(.18)	0.01	0.99	0.00	(.18)	0.02	0.98
Moral Reconation Therapy	0.16	(.16)	0.99	0.32	0.15	(.23)	0.63	0.53	0.20	(.22)	0.91	0.37	0.20	(.22)	0.90	0.37
Aggression Replacement Therapy	-0.09	(.26)	-0.35	0.73	0.05	(.41)	0.12	0.90	0.00	(.45)	0.00	1.00	0.01	(.45)	0.02	0.98
Interpersonal Problem Solving	-0.31	(.38)	-0.82	0.41	-0.29	(.34)	-0.85	0.40	-0.17	(.4)	-0.43	0.67	-0.17	(.4)	-0.42	0.68
Thinking for Change	0.00	NR	0.02	0.98	0.01	(.23)	0.05	0.96	0.10	(.27)	0.38	0.71	0.10	(.27)	0.38	0.70
Substance absue focus	-0.19	(.2)	-0.93	0.35	-0.15	(.19)	-0.80	0.43	-0.07	(.18)	-0.37	0.71	-0.07	(.18)	-0.36	0.72
Outcomes	58				58				102				102			
τ^2	0.062				0.060				0.078				0.083			

Note: (*) indicates Table 5 from LL

Note: *Arrest Recidivisim (w)* is the within study effect and is centered around the study mean, *Arrest Recidivism (m)*

Note: p-values less than 0.10 are bolded.

FIGURE 1.1 Example distributions of propensity score logits in population and experiment

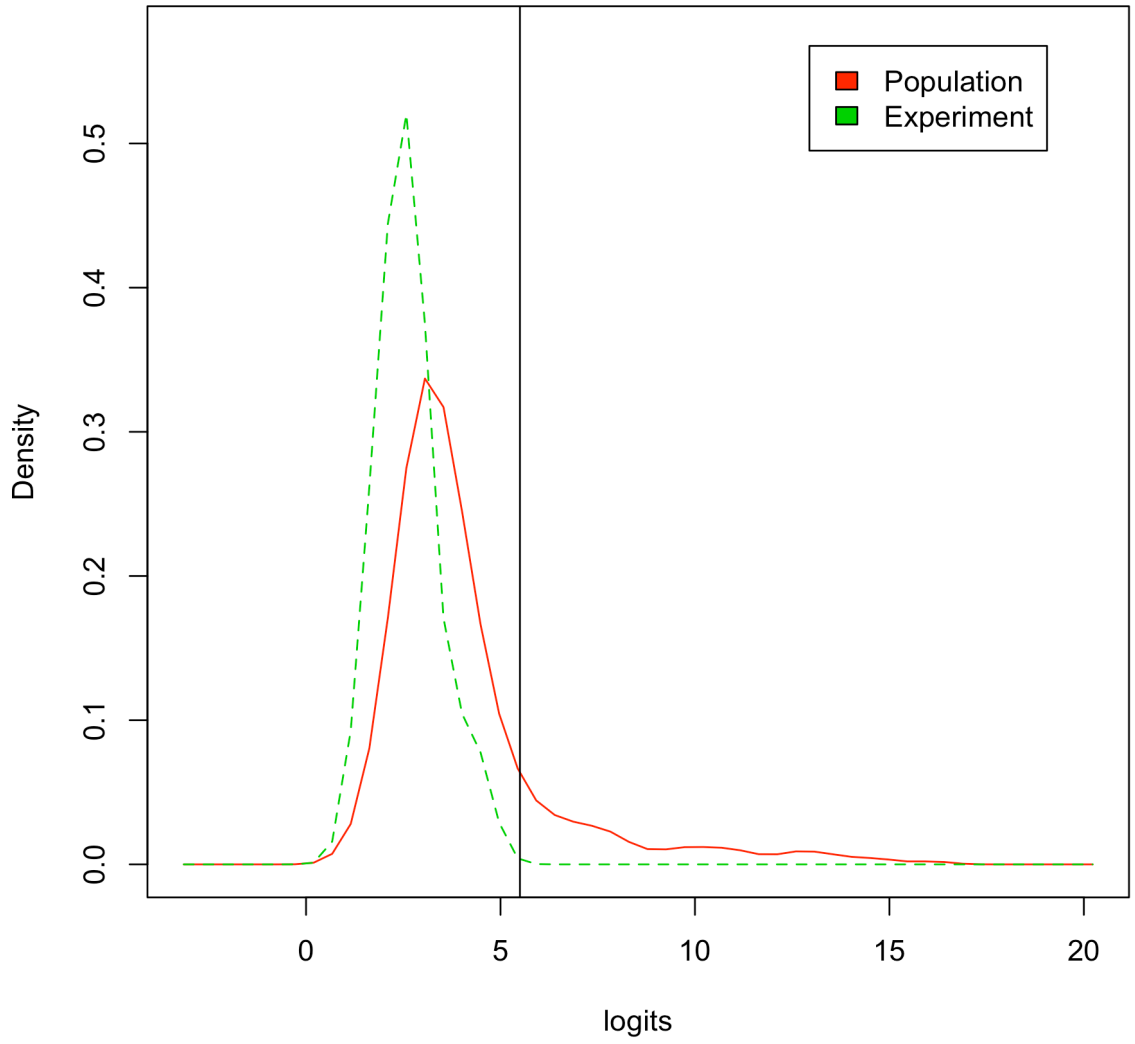


Figure 2.1 Relationship between $\rho_{c\alpha}$ and ρ

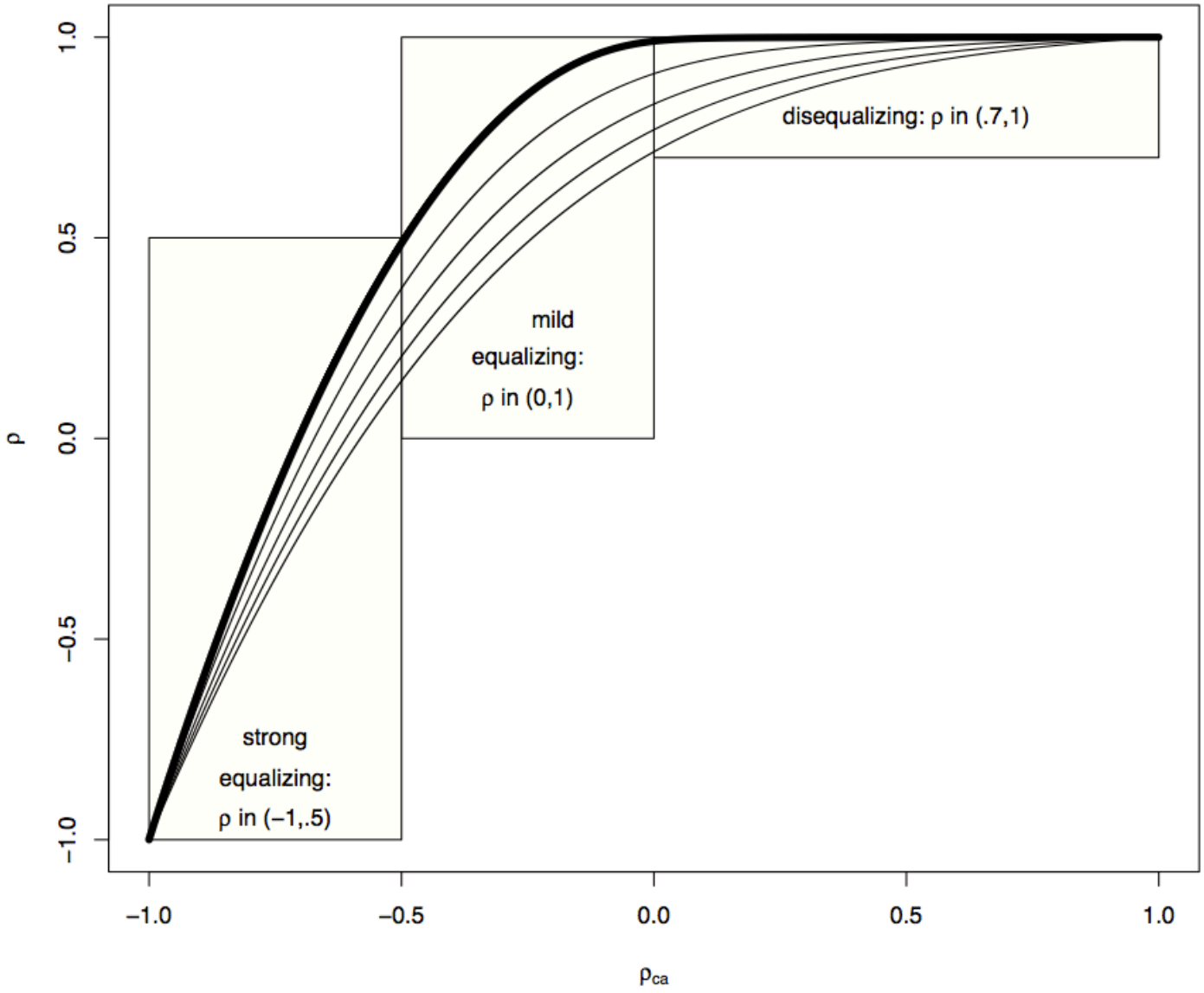
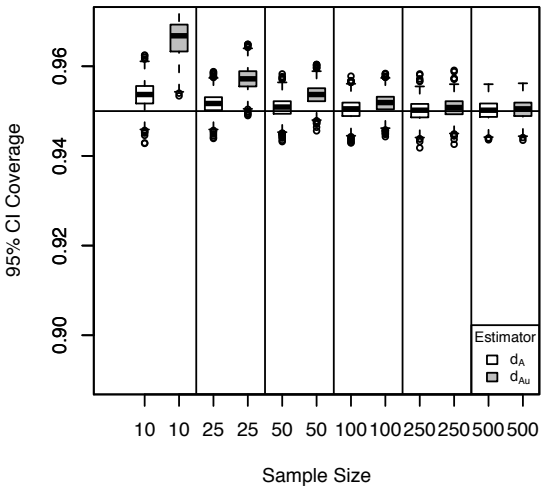
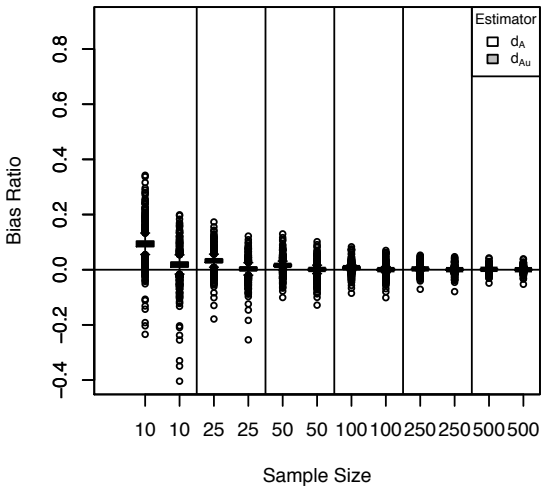


Figure 2.2 Coverage and bias for d_A & d_{Au}

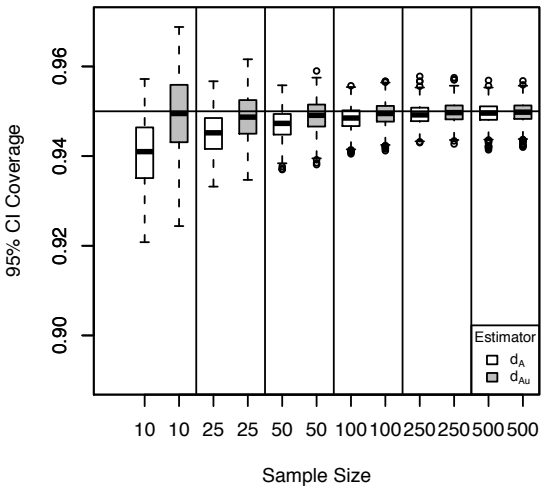
N(c)N(t): Coverage by sample size for d_A & d_{Au}



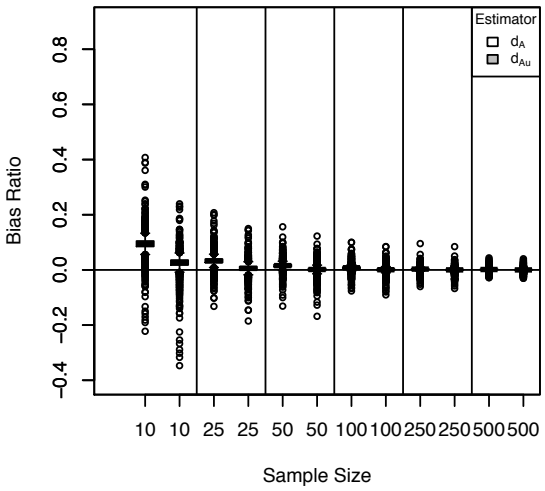
N(c)N(t): Bias ratios for d_A & d_{Au}



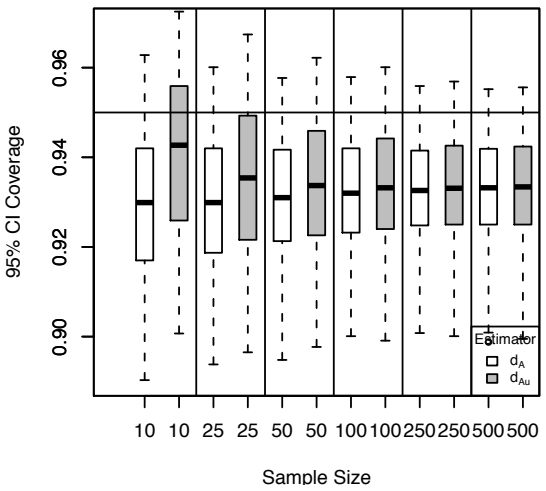
N(c)LN(t): Coverage by sample size for d_A & d_{Au}



N(c)LN(t): Bias ratios for d_A & d_{Au}



LN(c)LN(t): Coverage by sample size for d_A & d_{Au}



LN(c)LN(t): Bias ratios for d_A & d_{Au}

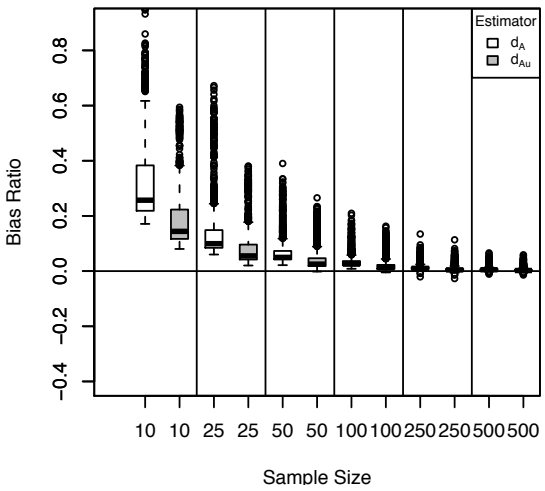
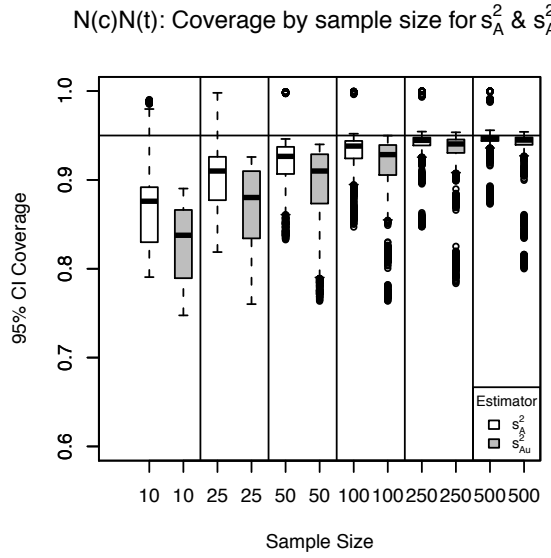
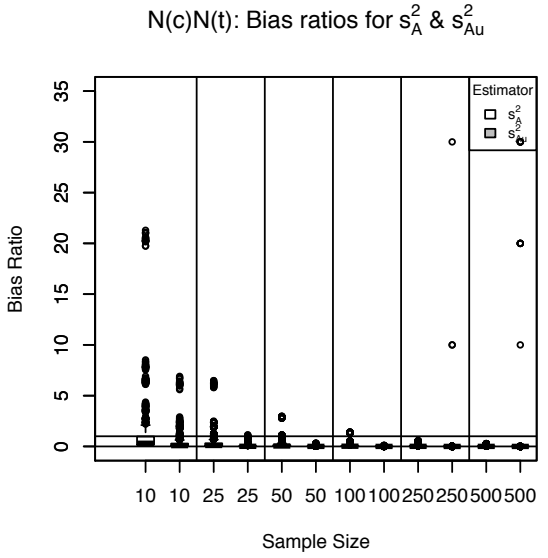
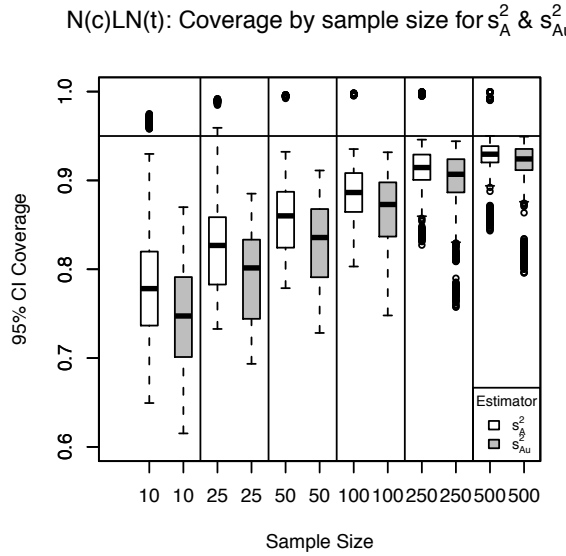
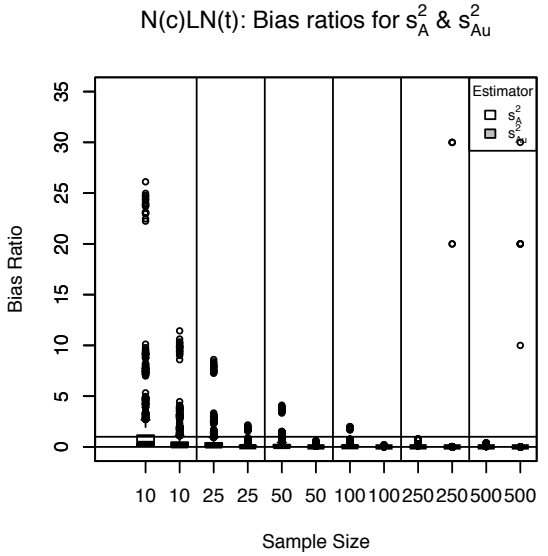


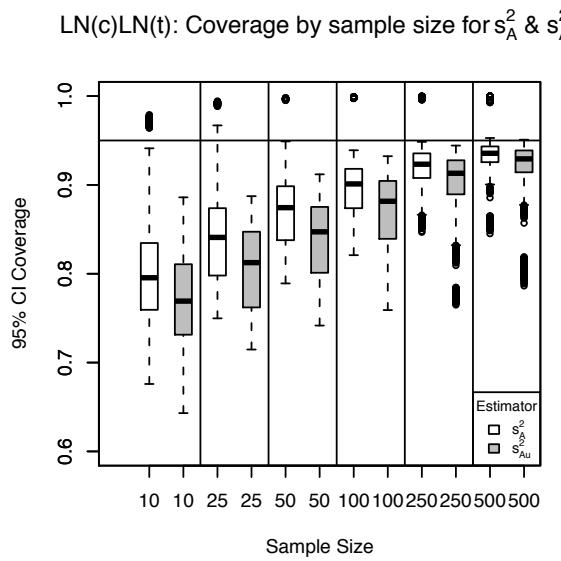
Figure 2.3 Coverage and bias for s_A^2 & s_{Au}^2











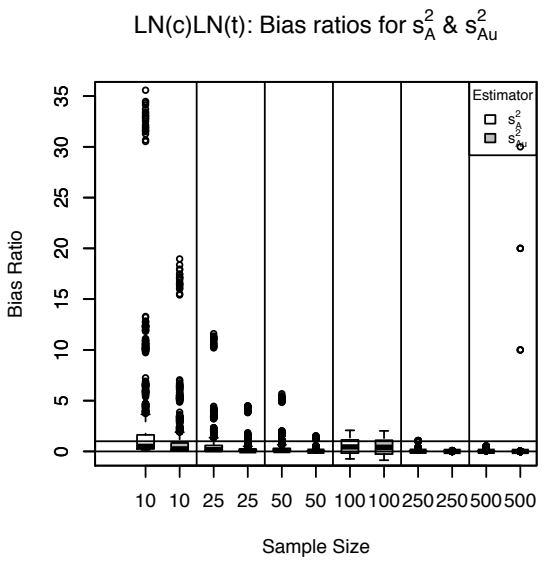
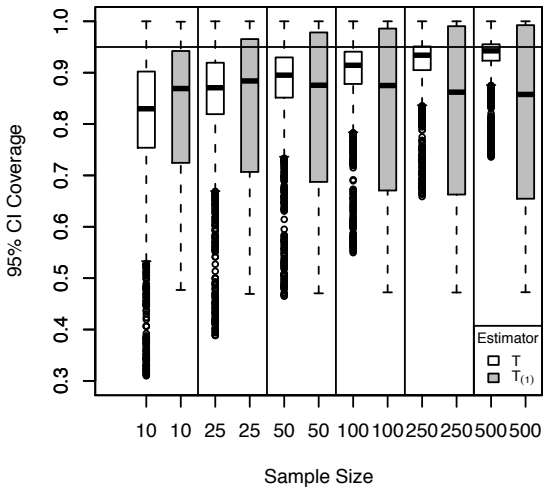
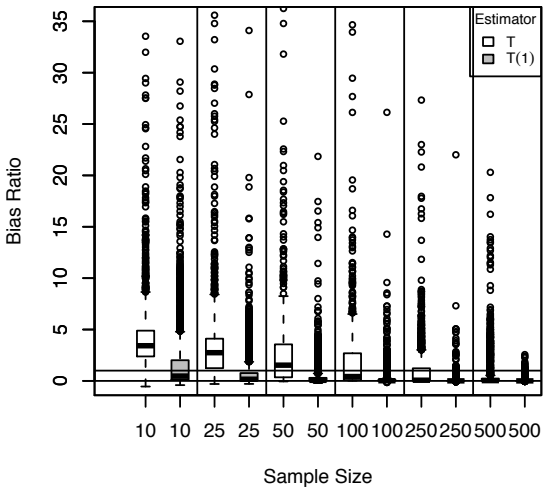


Figure 2.4 Coverage and bias for T & $T_{(1)}$

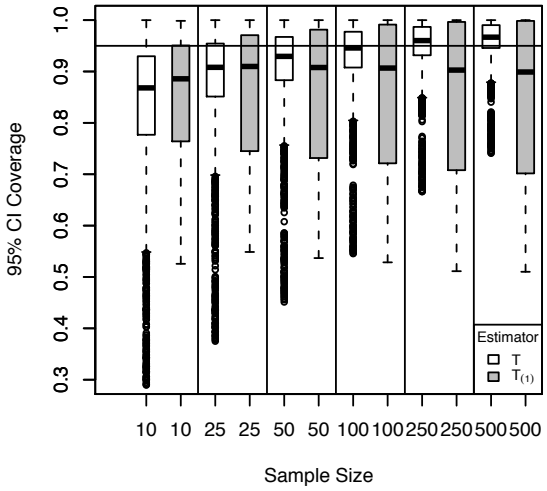
N(c)N(t): Coverage by sample size for T & $T_{(1)}$



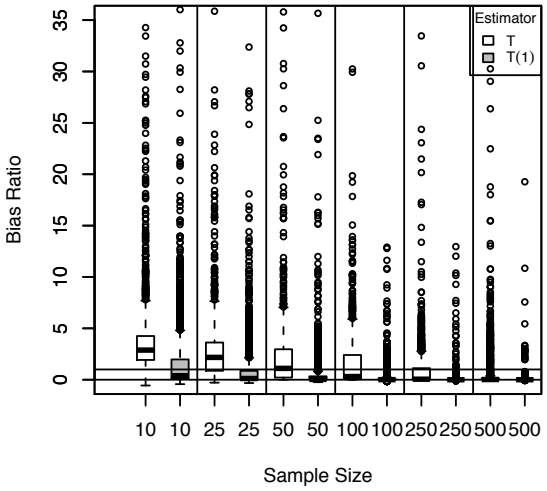
N(c)N(t): Bias ratios for T & $T_{(1)}$



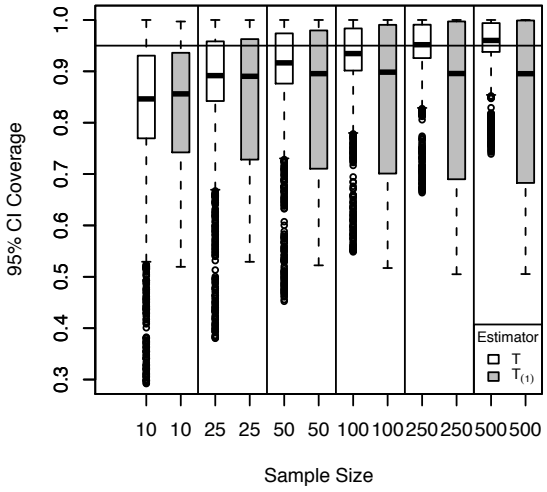
N(c)LN(t): Coverage by sample size for T & $T_{(1)}$



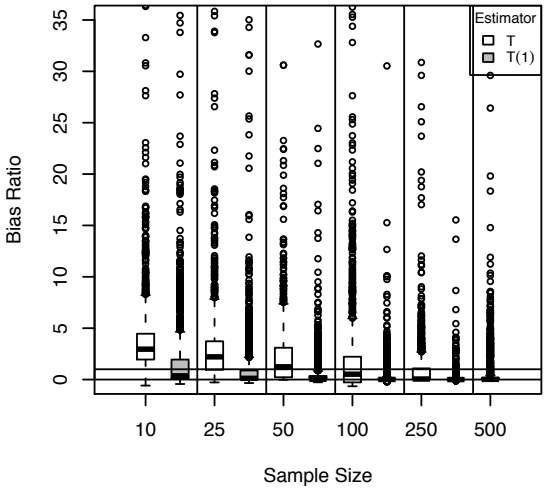
N(c)LN(t): Bias ratios for T & $T_{(1)}$



LN(c)LN(t): Coverage by sample size for T & $T_{(1)}$



LN(c)LN(t): Bias ratios for T & $T_{(1)}$



References

Angrist, J.D. (2003) Treatment effect heterogeneity in theory and practice. IZA Discussion Paper No 851. www.iza.org.

Bradburn, M.J., Deeks, J.J., Berlin, J.A., & Localio, A.R. (2007) Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*. 26(1): 53-77.

Bitler, M.P., Gelbach, J.B., & Hoynes, H.W. (2007) Can subgroup-specific mean treatment effects explain heterogeneity in welfare reform effects? Evidence from Connecticut's Jobs First experiment. Available at http://www.frbsf.org/economics/conferences/0806/bitler.pdf.

Bloom, H.S.(2005) Learning More from Social Experiments: Evolving Analytic Approaches. Russell Sage Foundation.

Bonett, D. G. (2006a) Confidence interval for a ratio of variances in bivariate non-normal distributions. Journal of Statistical Computation and Simulation 76:637–644.

Bonett, D. G. (2006b) Robust confidence interval for a ratio of standard deviations. Applied Psychological Measurement 30:432–439.

Bryk, A.S. & Raudenbush, S.W. (1988) Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. Psychological Bulletin. 104(3) p396-404.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295-313.

Cook, T.D. (1993) A quasi-sampling theory of the generalization of causal relationships. In *Understanding Causes and Generalizing About Them: New Directions for Program Evaluation*, ed. L Sechrest, AG Scott, 57:39-82. San Francisco: Jossey-Bass.

Cook, T.D., Shadish, W. R., & Wong, V.C. (2008) Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27: 724–750. doi: 10.1002/pam.20375

Cooper, H.M., Hedges, L.V., & Valentine, J.C. (2009) The handbook of research synthesis and meta-analysis. Russell Sage Foundation.

Cornfield, J., & Tukey, J. W. (1956) Average values of mean squares in factorials. *The Annals of Mathematical Statistics*, *27*(4), 907-949.

Cronbach, L.J. (1982) Designing Evaluations of Educational and Social Programs. San Francisco: Jossey-Bass.

Emerson, J.D., Hoaglin D.C., & Mosteller, F. (1993) A modified random-effect procedure for combining risk difference in sets of 2x2 tables from clinical trials. *Journal of the Italian Statistical Society*, 3:69-290.

Emerson, J.D., Hoaglin D.C., & Mosteller, F. (1996) Simple robust procedures for combining risk differences in sets of 2x2 tables. *Statistics in Medicine* 15:1465-1488.

Farrell, P.J. & Rogers-Stewart, K. (2008). Methods for generating longitudinally correlated binary data. *International Statistical Review*, 76(1): 28-38.

Feller, A. & Holmes, C. (2009) Beyond toplines: Heterogeneous treatment effects in randomized experiments. Available at http://www.stat.columbia.edu/~gelman/stuff_for_blog/feller.pdf.

Gadbury, G.L. & Iyer, H.K. (2000) Unit-treatment interaction and its practical consequences. *Biometrics.* 56(3): 882-885.

Gadbury, G.L. Iyer, H.K. & Allison, D.B. (2001) Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics.* 11(4): 313-333.

Glass, G.V. (1977) Integrating findings: The meta-analysis of research. *Review of Research in Education,* 5: 351-379.

Gleser, L.J., & Olkin, I. (1994) Stochastically dependent effect sizes. In H. Cooper & L.V. Hedges (Eds.), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Green, D.P. & Kern, H.L. (2010) Modeling heterogeneous treatment effects in large-scale experiments using Bayesian additive regression trees. Available at http://www.polisci.uiowa.edu/polmeth/papers/GreenandKern2010.pdf.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68:155–165.

Grissom, R.J. & Kim, J.J. (2001) Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*. 6(2):135-146.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., & Singer, E. (2009) *Survey methodology*. John Wiley and Sons.

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, *20*, 3875-3889.

Heckman, J.J., Smith, J., & Clements, N. (1997) Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*. 64(4): pp487-535.

Hedges, L.V. (1981) Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, 6(2): 107-128.

Hedges, L.V. & Olkin, I. (1984) Nonparametric estimators of effect size in meta-analysis. *Psychological Bulletin*, 96(3):573-580.

Hedges, L.V., Tipton, E., & Johnson, M. (2010a) Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods* 1(1): 39-65.

Hedges, L.V., Tipton, E., & Johnson, M.C. (2010b) Erratum: Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, (1)2: 164-165.

Hedges, L.V. & O'Muircheartaigh, C.A. (in press) Improving generalization from designed experiments. *Journal of the American Statistical Association*.

Herbert, R.D, Hayen, A., Macaskill, P., & Walter, S.D. (2011) Interval estimation for the difference of two independent variances. *Communication in Statistics—Simulation*

and Computation. 40: 744-758.

Higgins, J.P.T. & Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* (21): 1539-1558.

Hill, C.J., Bloom, H.S., Black, A.R., & Lipsey, M.W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3): 172-177.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2007) Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*. Software, http://gking.harvard.edu/matchit/.

Holland, P.W.(1986) Statistics and causal inference. Journal of the American Statistical Association. 81(396):945-960.

Imai, K., King, G. & Lau, O. (2006) Zelig: Everyone's statistical software. http://gking.harvard.edu/ zelig

Imai, K., King, G. & Stuart, E.A. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, A*. 171(2):481-502.

Imai, K. & Strauss, A. (2011) Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19:1-19. doi:10.1093/pan/mpq035.

Imbens, G. W. (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4-29. doi:

10.1162/003465304323023651.

Johnson, R.A. & Wichern, D.W. (2002) Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall.

Kang, S., & Jung, S. (2001) Generating correlated binary variables with complete specification of the joint distribution. *Biometrical Journal*. 43(3): 263-269.

Kent, D.M., Rothwell, P.M., Ioannidis, J.P.A., Altman, D.G. & Hayward, R.A. (2010) Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials Journal,* 11:85.

Knapp, G. & Hartung, J. (2003) Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* (22): 2693-2710.

Koenker, R. (2010) Rank tests for heterogeneous treatment effects with covariates. Working paper. *Accessed at* http://www.econ.uiuc.edu/~roger/research/ranks/qte.pdf.

Konstantopoulos, S. (2008) Do small classes reduce the achievement gap between low and high achievers? Evidence from Project STAR. *The Elementary School Journal*, 108(4): 275- 291.

Koopmans, L.H., Owen, D.B. & Rosenblatt, J.I. (1964) Confidence intervals for the coefficient of variation for the normal and log normal distributions. *Biometrika.* 51(1/2): 25-32.

Kromrey, J.D, Hogarty, K.Y., Ferron, J.M., Hines, C.V. & Hess, M.R. (2005) Robustness in meta-analysis: An empirical comparison of point and interval estimates of

standardized mean differences and Cliff's delta. Paper presented at the Joint Statistical Meetings, August 7-11, 2005. Minneapolis.

Landenberger, N.A., & Lipsey, M.W. (2005) The positive effects of cognitive-behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology* (1): 451 - 476.

Little, R. J. A. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 139-157.

Little, R.J.A. & Rubin, D.B. (1989) The analysis of social science data with missing values. Sociological Methods & Research. 18: 292-324.

Longford, N.T. (1999) Selection bias and treatment heterogeneity in clinical trials. *Statistics in Medicine*. 18, p1467-1474.

Lunn, A.D., & Davies, S.J. (1998) A note on generating correlated binary variables. *Biometrika,* 85(2): 487-490.

McGaw, B. & Glass, G.V. (1980) Choice of the metric for effect size in meta-analysis. *American Educational Research Journal*. 17(3): 325-337.

Nadarajah, S., & Kotz, S. (2006) R programs for computing truncated distributions. *Journal Of Statistical Software*, *16*(2), 1-8.

Neyman, J. (1923 [1990]) On the application of probability theory to agricultural experiments: Essay on principles, Section 9. *Statistical Science* 5 (4): 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed.

Nye, B., Hedges, L.V., & Konstantopoulos, S. (2000) The effects of small classes on academic achievement: The results of the Tennessee Class Size Experiment. *American Educational Research Journal*, 37(1): 123-151.

Oman, S.D. & Zucker, D.M. (2001) Modeling and generating correlated binary variables. *Biometrika* 88(1): 87-290.

Park, C.G., Park, T., & Shin, D.W. (1996) A simple method for generating correlated binary variables. *The American Statistician*, 50(4): 306-310.

Pettigrew, H.M., Gart, J.J. & Thomas, D.G. (1986) The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika* 73(2): 425-435.

Raudenbush, S.W. & Bryk, A.S. (1987) Examining correlates of diversity. *Journal of Educational and Behavioral Statistics*, 12(3): 241.

Raudenbush, S.W. (1988) Estimating change in dispersion. *Journal of Educational and Behavioral Statistics*, 13(2) p148-171.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55. doi: 10.1093/biomet/70.1.41.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524. .

Roschelle, J., Tatar, D., Shechtman, N., Hegedus, S., Hopkins, B., Knudsen, J., & Stroter, A. (2007). *Scaling Up SimCalc Project: Can a technology enhanced curriculum*

improve student learning of important mathematics? Menlo Park, CA: SRI International.

Royston, P. (1992) Which measures of skewness and kurtosis are best? *Statistics in Medicine*, 11: 333–343.

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 66(5): 688-701.

Rubin, D.B. (1978) Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*. 6(1): 34-58.

Rubin, D.B. (1990) Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference.* 25(3): 279-292.

Rubin, D.B. (2006) Matched Sampling for Causal Effects. Cambridge, England: Cambridge University Press.

Schneider, B. & McDonald, S.K. (2007) Scale-up in Education: Ideas in Principle. Volume I. Rowman & Littlefield Publishers, Inc.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton-Mifflin.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological methods*, *15*(1), 3-17. doi: 10.1037/a0015916.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (in press). The use of propensity scores to assess the generalizability of results from randomized trials. *The Journal of the Royal Statistical Society, Series A. PMC Journal.*

Vangel, M.G. (1996). Confidence intervals for a normal coefficient of variation. *The American Statistician.* 50(1):21-26.

Appendices

Appendix A: Proof of Proposition 1.2.2

From the assumptions given note that we can write,

$$\sigma^2 = \sigma_t^2 = \beta_T^2\sigma_e^2 + \sigma_{\varepsilon t}^2 = \beta_T^2\sigma_e^2 + (1 - \rho_{et}^2)\sigma^2 \text{ , and}$$

$$\sigma^2 = \sigma_c^2 = \beta_C^2\sigma_e^2 + \sigma_{\varepsilon c}^2 = \beta_C^2\sigma_e^2 + (1 - \rho_{ec}^2)\sigma^2,$$

where for each case $\beta^2 = \rho^2\sigma^2/\sigma_e^2$, σ^2 is the total variation in the outcomes, β is the linear coefficient relating e and the outcomes, σ_e^2 is the variation in $e(\boldsymbol{X})$ in the experiment, and ρ is the correlation between the outcomes Y and the variable e. In the experiment it can be shown that,

$$V(T) = V\left(\bar{Y}_T - \bar{Y}_C\right) = 4\sigma^2 / n \text{ ,}$$

whereas the variation of the subclassification estimator can be written,

$$V(T_S) = \sum_{j=1}^{k} w_{pj}^2 V\left(\bar{Y}_{T_j} - \bar{Y}_{C_j}\right) = \sum_{j=1}^{k} w_{pj}^2 \left(\frac{\sigma_{Tj}^2}{n_{Tj}} + \frac{\sigma_{Cj}^2}{n_{Cj}}\right).$$

By A2, $E(\sigma_{eCj}^2) = E(\sigma_{eTj}^2) = \sigma_{ej}^2$ and $E(n_{Cj}) = E(n_{Tj}) = n_j/2$, and using a first-order Taylor expansion,

$$E[V(T_S)] \approx 2\sum_{j=1}^{k} \frac{w_{pj}}{n_j}\left\{\left(\beta_C^2 + \beta_T^2\right)\sigma_{ej}^2 + [2 - \rho_{ec}^2 - \rho_{et}^2]\sigma^2\right\}.$$

Next let $E(n_j) = n\, w_{sj}$, where n_j is the observed number of sample units in stratum j and w_{sj} is the proportion expected when $e(\boldsymbol{X})$ follows a particular distribution. Then using a first-order Taylor expansion,

$$EE[V(T_S)] \approx \frac{2}{n}\sum_{j=1}^{k} \frac{w_{pj}^2}{w_{sj}}\left(\left(\beta_C^2 + \beta_T^2\right)\sigma_{ej}^2 + [2 - \rho_{ec}^2 - \rho_{et}^2]\sigma^2\right) = \frac{4\sigma^2}{n}[EVIF]$$

.

Appendix B: Method for generating correlated binomials

B.1 Generating the parameters π_{Cj} and π_{Tj}

Let (π_C,π_T) be the vector of average population proportions for the control and treatment group and (π_{Cj},π_{Tj}) be their corresponding population proportions in study *j*. In the intercept only model, assume for simplicity, that $(\pi_{Cj},\pi_{Tj}) = (\pi_{Cij},\pi_{Tij})$ for every measure *i* in study *j*; that is, assume that while there may be between study variability in parameter values, there is no within study parameter variability. Note that this means that both the baseline risk and the risk in the treatment groups are allowed to vary; for simplicity, we assume that the correlation between these study specific proportions is zero, i.e. $Corr(\pi_{Cj},\pi_{Tj})=0$.

In order to account for between study variability in the underlying parameters, the beta distribution is used, which is both a natural prior for the binomial distribution and assures us that the π_j values will remain in $(0,1)$. We parameterize this model as follows. Let $\pi_j \sim beta(\alpha,\beta)$, where $\alpha = \pi(1/\phi - 1)$ and $\beta = (1-\pi)(1/\phi - 1)$; using this parameterization allows $E(\pi_j) = \pi$ and $V(\pi_j) = \phi\pi(1-\pi)$. Note that this parameterization accounts for between study variability in the parameters π_j through the parameter ϕ. The same ϕ values are then used in both the treatment and control conditions. In order to choose ϕ values that are realistic, we focus on values of I^2 in relation to the π_j values themselves (instead of in relation to the effect size measures). We do this to guarantee comparability between parameters for the three outcome measures. For the parameters π_j, we have

$$I_j^2 = \frac{\phi\pi(1-\pi)}{\phi\pi(1-\pi) + \frac{1}{n}\pi_j(1-\pi_j)}$$

which has the approximate average value $I^2 = E(I_j^2) = n\phi/[\phi(n-1)+1]$. When $I^2 = 0$, $\phi=0$, and when $I^2>0$, $\phi = I^2/[I^2(1-n) + n]$.

We focus here on I^2 values of approximately *0, 1/3*, and *1/2*, which are not unusual in the literature. This leads to ϕ values of *0*, *1/(2n+1),* and *1/(n+1)*. For example, in cases in which all studies and groups have the same *n*, for *n* = 20,50,100 this gives ϕ values of .025, .01, and .005 for I^2 = 1/3 and .05, .02, and .01 for I^2 = 1/2. Note that for I^2 = 0, we use a small value of ϕ *(=.000001*) for programming reasons.

Finally, in the slopes case, we generate the treatment π_{ij} values using the following relationships:

Risk Difference*:* $\pi_{ij} = \pi_j + \beta x_{ij}$

Risk Ratio: $\pi_{ij} = exp(log\pi_j + \beta x_{ij})$

Log Odds Ratio: $\pi_{ij} = 1/\{1 + exp[log(1-\pi_j) - log\pi_j - \beta x_{ij}]\}$

For all simulations we let $\beta = 1$ for simplicity and generate the x_{ij} variables so as to guarantee that each π_{ij} is in *(0,1)*. This is easiest in the log-odds ratio case, where we generate $x_{ij} \sim N(0,.2^2)$ for each study and measure. In the log risk ratio case, we instead generate $x_{ij} \sim unif(-1,0)$; for a fixed π_j in the treatment group, this guarantees the $\pi_{ij} | x_{ij} \in (\pi_j e^{-1}, \pi_j)$. Finally, in the risk difference case, we generate the x_{ij} within each study such that $E(x_{ij}) = 0$. That is, we generate $x_{ij} \sim unif(\ min\ x, max\ x)$ where $min\ x = max(\pi_j - 1, -\pi_j)$ and $max\ x = min(\pi_j, 1-\pi_j)$. The important feature of all of these data generation

methods is that there is no covariance between the average values of x_{ij} in each study and their study specific π_j.

Additionally, it should be noted that in the slopes case, when the π_{ij}'s differ within studies, the value of ρ is bounded (Oman & Zucker, 2001). Therefore the generation of the correlated binomials is such that $Corr(p_{ij},p_{kj}) = \rho\,\bar{\rho}_{ik}$ where $\bar{\rho}_{ik}$ is the maximum possible pair-wise correlation for the *(i,k)* pair.

B.2 Generating the sample values, p_{Cij} and p_{Tij}

While the generation of multivariate normal random variables for simulations is standard practice, generating correlated binomial random variables that are additionally nested is not as common, though a few methods exist (Farrell & Rogers-Stewart, 2008; Kang & Jung, 2001; Lunn & Davies, 1998; Oman & Zucker, 2001; Park, Park & Shin, 1996). In order to generate the within study correlated binomial random variables, we use the method provided by Lunn and Davies (1998) and Oman and Zucker (2001). This method limits us to the case that $\rho > 0$; this is a reasonable assumption for the correlated measurement case we are interested in.

The method is as follows. For each treatment group and study *j*, outcome *i* and observation *k*, let $X_{ijk} \sim N(0,1)$, $X_{i0k} \sim N(0,1)$, and $U_{ijk} \sim bern(\rho^{1/2})$. We assume all of the random variables are mutually independent. Then let each outcome

$$Z_{ijk} = X_{ijk}\,(1 - U_{ijk}) + X_{i0k}\,U_{ijk}$$

and for a given set of marginal probabilities, where $c_{ij} = \Phi^{-1}(\pi_{ij})$ and $\Phi(.)$ is the normal CDF. Let

$$Y_{ijk} = \mathbf{1}_{\{Zijk \geq cij\}}$$

be the outcome of interest. The resulting summary statistics $p_{ij} = \frac{1}{n_j}\sum_{k=1}^{n_j} Y_{ijk}$ can be shown to be such that

$$E(p_j) = \pi_{ij}$$

$$V(p_j) = (1/n_j)\,\pi_{ij}(1 - \pi_{ij})$$

$$Corr(p_{ij}, p_{kj}) = \rho\,\bar{\rho}_{ik}$$

where $\bar{\rho}_{ik} = max\ \rho_{i,k|j}$ as defined previously. This method allows us to consider the cases in which the marginal probabilities of the *k* within study measures are all identical and the case in which they differ (while allowing their treatment effects to remain constant). Note that in the case that the marginal probabilities are equivalent, $\bar{\rho}_{ik} = 1$.

Appendix C: Average weights and binary outcomes

Theorem 3.2: Optimality of inverse mean variance weights

Let there be a single study, *j*, which we suppress in the following notation. Assume an estimated outcome *i* is of the form $f_i = f(p_{ci}, p_{ti}) = h(p_{ci}) - h(p_{ti})$; note that for the risk difference *h* is the identify function, for the log risk ratio *h* is the logarithm, and for the log odds ratio *h* is the log odds. Let $Corr(f_i, w_i)$ be the correlation between the estimate f_i and its weight w_i, and $Corr(f_i, w_\cdot)$ be the correlation between estimate f_i and the weight $w_\cdot$, where $w_i = 1/v_i$, $w_\cdot = 1/\bar{v}_{\cdot j}$, and $\rho = Corr(p_i, p_m) < 1$ for all $i,m = 1 \ldots k_j$,. Then, it can be shown that:

(1) $Corr(f_i, w_\cdot) \leq Corr(f_i, w_i)$

(2) $Corr(f_i,w_\cdot) \to 0$ as $k_j \to \infty$ and $\rho \to 0$.

Proof of Theorem 3.2:

First, note when the treatment and control group estimates are independent, the approximate variance of f_i is

$$V_i = V(f(p_{ci},p_{ti})) = V(h(p_{ci})) + V(h(p_{ti})) = G(p_{ci}) + G(p_{ti}).$$

and can be estimated by

$$v_i = v(f(p_{ci},p_{ti})) = v(h(p_{ci})) + v(h(p_{ti})) = g(p_{ci}) + g(p_{ti}).$$

Then the following results apply.

(1) Assume $w_i = 1/v_i$. Then the covariance between the weights and the estimates is

$$Cov(f(p_{ci},p_{ti}),w_i) = Cov(h(p_{ci}),w_i) - Cov(h(p_{ti}),w_i) = \gamma_{ci} - \gamma_{ti}.$$

Recalling that

$$w_i = v_i^{-1} = V_i^{-1} - V_i^{-2}\{(g(p_{ci}) - g(\pi_{ci})) + (g(p_{ti}) - g(\pi_{ti}))\}$$

then for each of the two groups, we have

$$\gamma_i = Cov(h(p_i),w_i) = -V_i^{-2}\, Cov(h(p_i),g(p_i)).$$

(2) Assume $w_i = 1/\bar{v}_j$ instead. Now the covariance can be approximated by

$$\begin{aligned} Cov(f_i,w_\cdot) &= -V_m^{-2}\frac{1}{k_j}\left\{Cov\left(h(p_{ti}),\sum_{m=1}^{k_j} g(p_{tm})\right) - Cov\left(h(p_{ci}),\sum_{m=1}^{k_j} g(p_{cm})\right)\right\} \\ &= \frac{V_i^2}{V_m^2}\frac{1}{k_j}\{\gamma_{ti} - \gamma_{ci}\} - V_m^{-2}\frac{1}{k_j}\sum_{\substack{m=1\\ m\neq i}}^{k_j}\left\{\rho^*_{tim}\sqrt{G(p_{ti})V(g(p_{tm}))} - \rho^*_{cim}\sqrt{G(p_{ci})V(g(p_{cm}))}\right\} \end{aligned}$$

where $\rho^*_{im} = \text{Corr}(h(p_i),g(p_m))$. Note that $\rho^*_{im} \le \rho^*_{ii}$ when $\rho < 1$. Furthermore, in the special case in which all of the outcomes in study j have the same underlying π_{ti} and π_{ci} values and the same ρ^*, this reduces to

$$Cov(f_i, w_\bullet) = \frac{1}{k_j}(\gamma_t - \gamma_c) - \rho^* \frac{k-1}{k} V_m^{-2} \left[\sqrt{G(p_t)V(g(p_t))} - \sqrt{G(p_c)V(g(p_c))} \right].$$

Clearly the results follow.