ON SELECTION BIAS MAGNITUDES

by

Julius Alexander Najab
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Psychology

Committee:

_____ Director

_____

_____

_____ Department Chairperson

_____ Program Director

_____ Dean, College of
Humanities and Social Sciences

Date: _____ Spring Semester 2014
George Mason University
Fairfax, VA

On Selection Bias Magnitudes

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Julius Alexander Najab
Master of Arts
George Mason University, 2007
Bachelor of Arts
University of Arizona, 2002

Director: Patrick E. McKnight, University Professor and Chair
Department of Psychology

Spring Semester 2014
George Mason University
Fairfax, VA

# Acknowledgments

I am who I am and accomplished what I have with the help of many people and I would like to thank just a few of them here. Su-Lin Trepanitis, Kelvin Najab, Steve Ellis, Tu-sing Touchton, and Patrick E. McKnight. Su-Lin, Steve, Kelvin, and Tu-sing prepared and contributed to my development and character in ways that I will never be able to repay or truly express. Any and all of my scientific accomplishments, insights, and skills are due to Patrick E. McKnight and his tireless efforts and encouragement. Each of you contributed much to putting me in this position and I thank you.

# Table of Contents

# List of Tables

# List of Figures

# Abstract

ON SELECTION BIAS MAGNITUDES

Julius Alexander Najab, PhD

George Mason University, 2014

Dissertation Director: Dr. Patrick E. McKnight

Selection bias remains the most prominent threat to validity in social and behavioral sciences. Non-equivalence between groups prior to an intervention reduces our ability to evaluate or infer intervention effects. Some methodologists argue that the effects due to selection bias may be estimated and subtracted from observed effects. If the estimate and subtract method were tenable then social scientists might be able to better understand past, present and future findings by employing this relatively simple procedure. Unfortunately, despite its prominence, selection bias remains largely unknown with respect to its magnitude of effect. The current dissertation aimed to do two things to facilitate the estimate and subtract method. First, I estimated the mean effect for selection bias effects in two different domains. The purpose for the different domains was to ensure that the estimates derived in one domain generalize into at least one other domain. Second, I used a resampling procedure to estimate the distribution of possible effect sizes due to selection bias. The sampling distribution allowed me to estimate the probability of any effect - at least according to the current study and, more importantly, to introduce a method that other researchers may employ in future studies similar to this study. Both aims were met by experimentally manipulating a study to produce selection bias effects. My overall aim was to demonstrate that an experimental procedure to manipulate, estimate,

and model selection bias was both possible and fruitful. Through this demonstration, I encourage other researchers to consider an experimental approach to better understanding threats to validity.

# Chapter 1: Introduction

Selection bias threatens scientific inference because its very presence means that comparisons groups are different for a reason outside of the researcher's manipulation. Since selection bias threatens scientific validity, researchers direct substantial attention, effort and resources to protect against or adjust this bias. Selection bias is theorized to occur and has been empirically observed in studies without random assignment (RA). The observation of selection bias in non-randomized (NR) studies has caused serious concern among social science researchers, with many suggesting that NR studies are unsuitable methodological designs for scientific research (*Procedures and standards handbook (Version 3.0)*, 2013). Research examining the differences in selection bias effects between NR and RA studies has rarely focused on the presence and magnitude of selection bias itself; instead, the research offers an evaluation of the effectiveness of adjustment tools designed to reduce bias. Few studies have directly measured the magnitude of selection bias effects. Of those studies from which a measure of selection bias can be measured or inferred, findings provide mixed messages about the magnitude of these effects - adding further uncertainty about the actual influence of selection bias on scientific results. Still, no previous study explicitly and directly manipulated selection bias in order to evaluate the influence.

Prior work (Donley & Ashcraft, 1992; Shadish, Clark, & Steiner, 2008) manipulated selection bias indirectly via self-selection or other mechanisms but this study is the first to not rely on self-selection to directly manipulate selection bias. Researchers remain deeply concerned about selection bias and with little known about the magnitude of its influence or the probability of its relevance. Without this knowledge, researchers have no empirical basis for their concerns. The current study aims to produce selection bias explicitly and directly to assess the potential severity - as measured by the magnitude and probability of effects.

## 1.1 Selection Bias

Selection bias is defined as any initial difference observed between comparison groups that is relevant to the outcome measure (Campbell, 1957). These initial differences often interact with the treatment and result in a compounded posttest difference from which it is difficult to disentangle the true treatment effect (B. Smith & Sechrest, 1991). The creation of non-equivalent groups is inherent in selection bias due to relevant participant variables that are not statistically controlled for (Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002). For example, consider the evaluation of a new mathematics educational curriculum. If a disproportionate number of participants with high math ability are assigned to the new curriculum relative to the standard curriculum, then the results will be biased towards the adoption of the new mathematics curriculum. The higher scores obtained by the participants in the new curriculum may be attributed to the curriculum being a better program or to the group's personal abilities or the interaction between the two factors. That bias, while present at pretest, becomes relevant at posttest largely due to pretest differences. Although a conceptual definition of selection bias has been firmly established in the literature and throughout social science, the classifications or causes of the bias are less clear and less understood with respect to their varying impacts of the outcome measure.

There are several types of classifications based on how participants are selected or assigned to groups including self, geographic, and administrative (Guo & Fraser, 2010; Shadish et al., 2002). Recent findings suggest that these classifications alone are not beneficial for identifying non-equivalent comparison groups because any of these assignment methods may result in non-equivalent comparison groups. Instead, focusing on assignment variables directly related or proximal to the potential selection bias are better suited for modeling bias in results than assignment variables indirectly related or distal to selection (Steiner, Cook, Shadish, & Clark, 2010). Mounting evidence suggests that typical "off-the-shelf" variables (e.g. demographic variables) are poor predictors of selection bias while variables closer to the selection mechanism (e.g. motivation, ability) perform better in identifying

and reducing causes of selection bias (Cook, Shadish, & Wong, 2008; Glazerman, Levy, & Myers, 2003; Shadish et al., 2008; Steiner et al., 2010). For example, including the influence of math ability on the mathematics treatment mentioned previously would likely produce better estimates of modeling bias than would be yielded from including musical ability. In accord with this perspective, this study focuses on selection variables related to the assignment mechanism (i.e., direct or indirect) rather than the overarching classification types of how (e.g., self, geographic, etc.).

## 1.2 Prior Studies of Selection Bias

Historically, RA studies have been viewed as less susceptible to selection bias relative to NR studies. Still, early predictions and theoretical work stated that NR designs could produce results similar to RA results given a mechanism for creating a suitably matched comparison group (Rubin, 1974). Researchers across the social sciences studied the reliability and validity of these predictions and theories using detailed examinations of field studies, systematic quantitative and qualitative reviews, simulations, and more recently, laboratory experiments (Cook et al., 2008; Dehejia & Wahba, 1999; LaLonde, 1986; Lipsey & Wilson, 1993; Shadish et al., 2008) (see Appendix for greater detail).

Results from an NR study in economics that evaluated a training program's effect on employees' earnings underestimated the program's effectiveness relative to results obtained in a RA study using similar data (Fraker & Maynard, 1987; LaLonde, 1986). Those findings were immediately questioned and tested by other economic researchers due to concerns about the inclusion of poor comparison groups and failure to incorporate modern statistical adjustment tools (e.g. sample selection, propensity score matching) and they found minimal bias in NR study results (Dehejia & Wahba, 1999; Heckman, Hotz, & Dabos, 1987). Still, these attempts to improve causal inference in NR designs by using better comparison groups and modern statistical tools were questioned in a subsequent meta-analysis (Glazerman et al., 2003).

After those initial comparison studies in economics, researchers in other social science fields began to conduct detailed NR and RA design comparisons. As observed within economics, a similar pattern of divergent, inconclusive, and caveated conclusions for the resiliency of NR designs to selection bias effects presented itself in educational (Heinsman & Shadish, 1996), medical (Benson & Hartz, 2000; Ferriter & Huband, 2005) and psychological research (Shadish & Ragsdale, 1996). Comparisons between NR and RA results overwhelmingly showed divergent results presumed to represent selection bias effects (see Appendix A for greater detail).

Given the unreliable effects for selection bias within NR designs and their adjustment tools, no consensus exists regarding the presence and size of these effects. A thorough review of the literature, however, revealed that various researchers across the social sciences agreed upon several aspects of selection bias within NR designs (Cook et al., 2008; Ferriter & Huband, 2005; Glazerman et al., 2003; Shadish, 2010). First, NR studies reliably approximated RA experimental results for both methodological (e.g. comparison groups, Regression Discontinuity) and statistical (e.g. Heckman sample selection, propensity score matching) techniques when the comparison group matched the experimental group and when the statistical assumptions were not violated, respectively. Second, comparisons groups from NR studies worked adequately when the participants in the comparison were matched on variables directly relevant to the outcome measure (Cook et al., 2008; Shadish, 2010). Third, statistical tools to account for selection bias worked better when the selected covariates were not basic demographic variables but instead were indicators of motivation or ability (Dehejia & Wahba, 1999; Heckman et al., 1987; Pohl, Steiner, Eisermann, Soellner, & Cook, 2009; Shadish et al., 2008) (see Appendix A for greater detail). Taken together, NR designs were potentially effective at approximating RA experimental results only when comparison groups were rigorously selected and appropriate covariates were included in the statistical tools - conditions not regularly met by researchers. The failure to meet these conditions resulted in some researchers viewing NR studies as a poor design choice (Glazerman et al., 2003; Shadish, 2010).

Despite the empirical knowledge regarding when NR results approximate RA results, the magnitude of selection bias in NR studies remains unknown. Few studies that present standardized selection bias effects or differences between NR and RA study results provide findings that vary significantly and unpredictably. The estimation of reliable selection bias magnitudes could result in greater confidence in NR study findings, along with decreased reticence to use NR designs.

## 1.3 Prior Estimates of Selection Bias Effects

Shapiro and Shapiro (1983) included assignment mechanism (random or not) as a moderator in their quantitative review of psychotherapies, and yielded a standardized selection bias effect of d = .20. About a decade later, a meta-analysis on psychotherapy outcomes reported a selection bias effect of d = .52 (Shadish & Ragsdale, 1996). Those two effects are the closest to a replication of selection bias for a specific content that exists.

Reviews that incorporated primary studies from a breadth of content areas yielded small to medium selection bias effects. Via a reivew across multiple programs within and outside psychology, Lipsey and Wilson (1993) estimated selection bias to be quite low (d = .06) whereas another review (Heinsman & Shadish, 1996) of selected psychological treatment programs reported a modest but higher effect (d = .25). The higher, .25 effect was not consistent across content areas, with the included outcomes; Scholastic Aptitude Test coaching (d = .01), academic achievement ability groups (d = -.21), juvenile drug use prevention (d = .15), and psychosocial interventions for post surgery outcomes (d =.03) (Heinsman & Shadish, 1996). As would be expected, quantitative reviews that used an average across results are less variable than the differences observed between single studies directly manipulating selection bias.

A well controlled, but methodologically flawed, quasi-experimental study assessed difference due to participant self-selection and produced posttest effect sizes between d = .73 and 1.38 (Donley & Ashcraft, 1992). In contrast, a more methodologically sound design

by Shadish, Clark and Steiner (2008) produced only a medium posttest effect sizes (d = .24). Still further, a direct replication of Shadish et al. (2008) in Germany found only a small effect (d = .06) (Pohl et al., 2009). Across a variety of studies that either comprehensively reviewed the published literature or directly manipulated selection leads to a simple conclusion - there are no stable estimates of selection bias effects other than the basic fact that they may be present and produce small to medium effects. Perhaps prior reviews and manipulation studies were not designed to produce the maximum effect nor were they designed to produce replicable effects. Regardless, the current estimates of selection bias effects appear far smaller than what might warrant the attention.

## 1.4 Why Estimate Selection Bias Effects?

As alluded to previously, estimating selection bias effects enable researchers to better understand the magnitude and probability of the problem. Researchers who fear the unknown may end up expending tremendous (financial, material, personnel, and opportunity) resources over-correcting for a trivial problem and ignore the more pressing concerns. Selection bias - just one of many potential threats to internal validity - garners tremendous attention in the research literature and by researchers themselves. Failure to appreciate the problem adequately leads to our current situation of ritualistic research design use rather than principled and empirically supported best practices.

## 1.5 Using Selection Bias Effects - Estimate & Subtract

Estimating the magnitude and probability of selection bias enables researchers to adhere to a more empirically-driven decision making process. If selection bias distributions and probabilities can be reliably estimated for different assignment related variables, then a statistical adjustment procedure called Estimate & Subtract (E&S; Reichardt & Gollob, 1989) may serve as an alternative to resource intensive tools to reduce the impact of selection bias (e.g. propensity scores, Heckman sample selection, and Instrumental Variable; see

Appendix A for greater detail on the traditional statistical adjustment tools). E&S is a relatively new and untested approach to reducing or eliminating selection bias in NR study results. As a statistical adjustment tool it has received less attention than other adjustment techniques due to the absence of reliable selection bias magnitudes and the absence of any distributions for those magnitudes.

A distinct advantage to the E&S tool over other statistical tools is its simplicity. For example, if selection bias were known to produce a modest effect (e.g, d = 0.3) for a particular population and an NR study published a large effect (e.g., d = 0.9), then the E&S method would estimate the "true" influence of the treatment to be a more moderate effect (i.e., d=0.6 = 0.9 - 0.3). Propensity score matching, Heckman sample selection models, and instrumental variables require researchers to collect additional data during the study itself - a situation that either does not exist or places a burden on archival data analysts that may never be met. Consider these requirements: these statistical adjustment tools require numerous covariates and large sample sizes that make many of these procedures unavailable (Guo & Fraser, 2010; Shadish et al., 2008; Steiner et al., 2010). Even if the requirements of these statistical techniques are met, there are guarantees that useful conditional probabilities or the appropriate selection model will be estimated - both situations that compromise the effectiveness of these statistical tools (Cook et al., 2008; Dehejia & Wahba, 1999; Heckman & Navarro-Lozano, 2004; Steiner et al., 2010).

Estimate & Subtract relies upon the prior estimates of selection bias effects and the probability density functions around those effects. Creating a point estimate and a distribution is beneficial for decision making regarding experimental design selection and empirical evidence evaluation (examples can be read below). Distributions help decision making through using confidence intervals as well as probabilities. A probability density function is useful for calculating the area under the curve and, in so doing the probability for any given value. A selection bias effect size distribution, thus, is valuable for creating a probability density curve and for calculating the area under that curve. Using the area under the curve, an investigator may compute a probability for any given effect size. The reason why a known

distribution is preferred to an unknown or unknowable distribution is simply a matter of efficiency. If the selection bias probability distribution matches a known probability distribution function then the effect size probabilities may be readily looked up in established tables. Tables, however, are no longer essential for estimating areas under curves. There are more advanced mathematical procedures available on standard computers that enable us to estimate the area under a curve. Known distributions are not essential but they are extremely helpful. Using those tables or computer functions, a researcher may estimate the probability of any potential selection bias effect size. The first example is an a priori problem where the researcher must make a forecast of the potential selection bias threat; the second example poses a similar problem but, in this case is a retrospective or post-hoc analysis of a completed study. Both examples illustrate the utility of the estimate and subtract method via the probability density function.

Example 1 (a priori decisions): A researcher considers conducting a study using a NR design. The study assesses an educational skill development programs effect on a specific outcome. Based upon prior research, the researcher expects a treatment effect size of 1.5 (Cohen's d). Now, the researcher asks the question: "what is the probability that my design will produce an effect that may be fully attributable to selection bias?" Given the standard normal distribution we discussed previously, the probability that the selection bias effect would be equal to or greater than .2 is equal to the area under the curve from the mean to infinity. That area, according to the standard function for normal distributions is equal to 0.50. Thus, the researcher has a probability of 0.50 of exceeding the mean selection bias effect. Reichardt and Gollab (1989) proposed estimating the size of validity threats primarily to improve decision making with biased study results. If selection bias effect sizes conform to a normal distribution ($\mu = .20, \sigma = .6$, for this example) then researchers can evaluate the observed study effects in context of the quantified validity threat effects. Utilizing the selection bias distribution a researcher expecting a 1.5 effect size can make an argument that even if the results contain selection bias the difference of the selection effect (an expected

d = .2) does not make a meaningful difference. This information helps researchers make a-priori design decisions. However, if the expected studys result is a smaller effect than that large effect, a selection bias distribution with the previously described properties would assist the researcher in deciding a better avenue for obtaining defensible results will be with the additional resource costs that come with an experimental design or switch to a stronger design like regression discontinuity, or preparing for a biased effect by measuring relevant covariates (like ability) in order to make post hoc statistical adjustments.

Example 2 (post-hoc decisions): A researcher implemented an educational skill development program using a non-randomized design. The researcher found their program's treatment effect to be d = .7. Meta-analyses summarize the average treatment effect size at .5 (Cohen's d). However, much of published primary research utilized experimental designs. Leaving the researcher to question "How much of the observed effect is due to the treatment or to selection bias?" If for this example using estimated selection bias distribution the researcher could broadly answer how much bias is possibly due to the selection threat by creating a confidence interval around the suspected selection bias magnitude and then subtract that empirically derived range from the study's observed effect as intended by Reichardt and Gollab. Returning to the theoretical normal distribution of selection bias effect with a mean of d = .2 and standard deviation of d = .6 the researcher can argue that their observed effect size is larger than previous research possibly due to selection bias and not to the treatment program. Arguing with 95% confidence (obtained by multiplying the selection bias probability distribution's standard deviation by the 1.96 z-score) any selection bias effects range from d = - .98 to d = 1.38. Thus, the observed d = .7 can not only plausibly be described as partly due to selection bias but the entire observed effect is well within a 95% confidence interval for an average selection bias effect. Now the researcher can report on the quantified amount of uncertainty about the effectiveness of their implemented treatment program. The researcher could now report with some confidence and empirical evidence on both the treatment effect and the selection bias effect that make up in part (other potential validity threats notwithstanding) the total observed effect. Estimate &

Subtract was thought up as a method for quantifying validity threat estimates and assist researchers to determine those threat estimates are acceptable using a modicum of empirical justification.

Without these two parameters (i.e., prior estimates and probability density functions of selection bias effects), the E&S technique fails to hold any advantage over the more burdensome and complicated statistical tools. Fortunately, one study does provide easily replicable treatments and measures for estimating selection bias effects across content areas and assignment variables (Shadish et al., 2008). That study formed the basis for the current study; I provide the relevant details of it below.

Shadish et al. (2008) designed and implemented a truly comparable study of NR and RA results from a sample of participants that was not limited to only self-selection and no control over the selection variable like previous quasi-experimental, simulation studies, and meta-analyses. The researchers implemented a doubly randomized preference trial by randomly assigning participants to either a Randomized Control Trial (RCT) or quasi-experimental study. Participants assigned to the RCT were either randomly assigned to a mathematics or vocabulary treatment condition. Treatments and outcome measures for mathematics and vocabulary were selected because of their relevance to many NR studies in education and due to their relatively easy and controllable implementation in a laboratory setting. Participants assigned to the quasi-experimental study chose either a mathematics or vocabulary treatment based on their own preference. Participants received the same measures pre and post treatment.

The objective of the Shadish et al. (2008) study was not to manipulate selection bias or to even measure selection bias effects. Instead, the author's aimed to examine how well various statistical adjustment tools reduced selection bias effects. To do so, they compared the RCT results to adjusted quasi-experiment results, finding that most adjustment procedures (linear regression, propensity score matching, propensity scores as covariates, and various propensity score matching techniques) effectively removed selection bias effects (Pohl et al., 2009; Shadish et al., 2008; Steiner et al., 2010).

The present study does not aim to replicate the methods and objectives of Shadish et al. (2008). However, the treatments and measures from this well controlled and previously replicated experiment provide an important starting point for the establishment of selection bias magnitudes. As such, I have incorporated much of the Shadish et al. (2008) treatments, measures, and procedures in this study.

## 1.6 Purpose of Study

Despite a variety of possible threats to validity, selection bias receives the highest level of attention among psychological researchers. Selection bias arises when groups are non-equivalent by design or by circumstance. Most scientists rely on random assignment (RA) as the gold-standard to counter the potential influence of selection bias. Many researchers presume that without random assignment, selection bias 1) is ubiquitous in non-randomized (NR; observational and quasi-experimental) studies (Shadish et al., 2002), 2) inevitably threatens the study's conclusions, and 3) deserves our attention above all other threats. These presumptions lead to research practices that minimize the use of NR designs and lead to the development of numerous cumbersome NR statistical adjustment procedures. To date, it is unclear how likely (probability) selection bias is to occur or to what extent (magnitude) selection bias influences results. The following study provides the first empirical manipulation and estimation of selection bias' probability and magnitude.

The goal of the present study was two-fold: A) to estimate the magnitude of selection bias effects for direct and indirect selection bias variables, B) resample the estimated magnitudes of selection bias effects to estimate the probability distribution around the point estimates. This work represents an initial step in the estimation of probabilities and magnitudes of the various threats to validity. The purpose of this dissertation is to directly address those goals. In so doing, I expected the following: 1) selection bias can be manipulated and will produce substantial variability in treatment effect sizes, 2) the variability will be moderated by the selection bias mechanism.

# Chapter 2: Methods

The current study consisted of a randomized trial where selection bias was explicitly manipulated. A basic, randomized controlled trial (i.e., experimental design) along with a procedure similar to a prior study by Shadish et al. (2008) enabled me to meet both aforementioned goals. The following chapter provides the details of the design, procedure, measures, preliminary analyses, and computational algorithms. I based my design, measurement model, and analysis on the causal model in Figure 2.1. Each of the model details will be fully explained in the following sections.

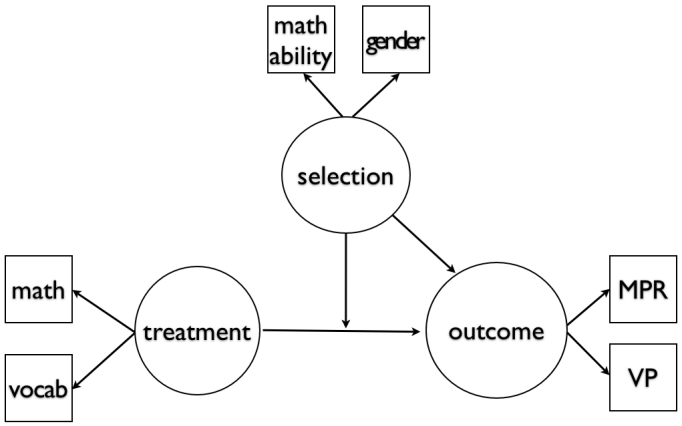



Figure 2.1: Causal model

## 2.1 Participants

One hundred fifteen (N = 115) undergraduate students enrolled in a large, mid-atlantic university volunteered and received course credit for their participation in the study. Students

were informed that they would either receive a vocabulary or mathematics lecture as part of the study. Students met selection criteria for the experiment based on their self-reported math SAT scores, an approximation of math ability. Approximately 95% (N=109) of participants produced their requested verified SAT scores (the other 5% provided unofficial SAT scores but were retained in the data analyses).

## 2.2 Design

The design used to generate selection bias estimates was a two-arm block randomized trial. I chose this design because it allowed me to both experimentally manipulate selection bias as well as measure alternative designs (i.e., quasi-experimental or NR) and their potential selection bias effects. With respect to the experimental manipulation, I used self-reported math ability as a blocking factor. Math ability - even self-reported - served as a variable directly relevant to the treatment. Participants who differ on math ability prior to treatment, by definition, are non-equivalent and their non-equivalence impacts the treatment and its potential outcome. To ensure maximal differences by design, I divided participants into either high or low math ability groups prior to randomization and then randomly assigned to either a mathematics or a vocabulary treatment condition. Thus, I created four groups - Math Ability (2 levels) X Treatment Condition (2 levels) - see Figure 2.2. For example, Participants with high math ability were assigned into either the High-Math (i.e., high math ability participants who received the mathematics treatment) or High-Vocab (i.e., high math ability participants who received the vocabulary treatment) groups. The blocking variable allowed me to manipulate the strength of the selection bias mechanism where that "ability" interacted with the treatment. In short, I created a selection by treatment interaction. As a main effect, the mathematics treatment represented the strong, selection bias-relevant condition while the vocabulary treatment represented the weak, selection bias-irrelevant condition. Participants' math ability (i.e., the blocking variable) and gender (i.e, pseudo selection bias variable that served as my, post-hoc blocking variable for

later analyses) served as the direct and indirect selection variables to produce the maximum and minimum differences between groups (see Figure 2.2). Pretest measurement is at $O_1$ for the High-Math group and the corresponding posttest measurement point is at $O_2$. All pretests have odd numbers (i.e., $O_1$,$O_3$,$O_5$,$O_7$) while all posttests have even numbers (i.e., $O_2$,$O_4$,$O_6$,$O_8$).

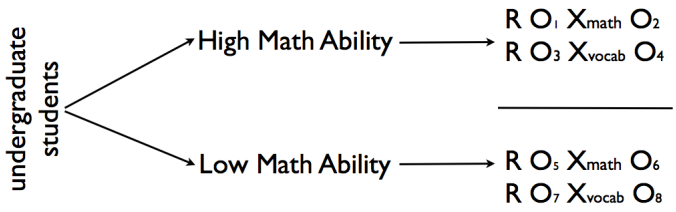



Figure 2.2: Design for ability selection variable

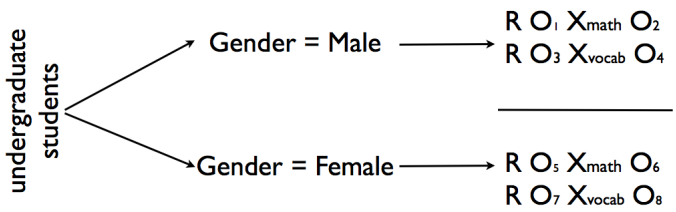



Figure 2.3: Design for gender - the pseudo selection variable

### 2.2.1 Selection Bias Mechanisms

To create both strong and weak selection bias, I used two separate mechanisms based upon the logic provided by Steiner et al. (2010). Steiner et al. (2010) differentiated between direct or indirect relationships for selection variables to treatments and outcomes. In the present study, the selection variable (math ability) related directly to both the treatment and to the outcome, achievement on a mathematics exam. Math ability was assessed using math scores from the Scholastic Aptitude Test (SAT). The indirect mechanism of selection was generated by using gender as a proxy. This study relied upon math SAT scores being a good predictor of math performance (i.e., the MPR; see Figure 2.1). As such, math SAT scores ought to have a greater relationship to pretest MPR scores than to vocabulary SAT, gender, RG-I, and participant's lecture preference. Math SAT had a greater relationship with MPR than gender, RG-I score, and treatment preference (see Table 2.1). I provide further details of these two mechanisms (ability and gender) below.

Table 2.1: Selection and Outcome Variable Correlations

| | Math Performance Revised at Pretest |
|---|---|
| Math SAT | 0.62 |
| Reasoning General-I | 0.46 |
| Vocabulary SAT | 0.41 |
| Gender | 0.11 |
| Treatment Preference | 0.10 |

#### Direct Selection via Math Ability

Participants in the high ability group were required to have a minimum 630 math SAT score while those in the low ability group had a maximum 530 math SAT score. Those values were based upon a rationale for creating maximal differences with the available sample.

An SAT math score of 630 is in the upper quartile (75th percentile) while a score below 530 represents the 50th or lower percentiles. The available sample (see Participants section above) required me to restrict the selection and cut score range because few participants had extremely low scores. Thus, I used the most defensible cut scores to maximize selection bias. The recently reported US national average math SAT score was 516 (SD = 116) for college bound high school students (College Board, 2010). This average varies one or two points from earlier years. Males average 534 (SD = 118) while females average 499 (SD = 112) for the same national sample. The differences between my two math ability groups reflect approximately one standard deviation separation according to those norms.

High ability groups had similar means (682 and 690) and distributions for their math SAT scores across vocabulary and mathematics (see Table 2.2). Low ability groups had similar means (487 and 489) and distributions for their math SAT scores (see Table 2.2). There is nearly a 200 point math SAT score difference between the means of the high and low math ability groups.

Table 2.2: Ability Selection Group Descriptives

| | | Math SAT | | | Vocabulary SAT | | |
|---|---|---|---|---|---|---|---|
| **Ability-Tx** | n | *M* | *SD* | *SE* | *M* | *SD* | *SE* |
| High-Math | 25 | 682 | 39 | 7.89 | 578 | 73 | 14.61 |
| High-Vocab | 22 | 690 | 43 | 9.14 | 623 | 94 | 20.13 |
| Low-Math | 36 | 487 | 39 | 6.55 | 526 | 72 | 12.06 |
| Low-Vocab | 32 | 489 | 36 | 6.45 | 491 | 92 | 16.32 |

Table 2.3: Ability Selection Group Descriptives

| | | Reasoning General-I | | | Vocabulary General-II | | |
|---|---|---|---|---|---|---|---|
| **Ability-Tx** | n | *M* | *SD* | *SE* | *M* | *SD* | *SE* |
| High-Math | 25 | 8.7 | 2.4 | 0.48 | 23.4 | 3.9 | 0.78 |
| High-Vocab | 22 | 8.8 | 2.7 | 0.56 | 24.7 | 3.7 | 0.80 |
| Low-Math | 36 | 5.9 | 1.9 | 0.32 | 21.2 | 3.4 | 0.56 |
| Low-Vocab | 32 | 5.6 | 2.1 | 0.37 | 21.7 | 4.2 | 0.73 |

### Indirect Selection via Gender

Gender served as a pseudo selection variable indirectly related to the treatment and the outcome. Participants were not assigned to treatment conditions according to gender. Previous related studies using gender as a covariate in selection bias reduction procedures found it to be of limited use and weakly related to the outcome variable (Cook, Steiner, & Pohl, 2009; Cook & Steiner, 2010; Steiner et al., 2010). Following data collection, pseudo-comparison groups were created based on participants' gender. The same four-group pre-post design based on participants' math ability groups was simulated using the newly created gender-based pseudo group (See Figure 2.3). Replicating the original design, males were either in the Male-Math (i.e., male participants who received the mathematics treatment; N = 30), or the Male-Vocab (i.e., male participants who received the vocabulary treatment; N = 27), while females were in the Female-Math (i.e., female participants who received the mathematics treatment; N = 31), or the Female-Vocab (i.e., female participants who received the vocabulary treatment; N = 27) groups. The point biserial correlation ($r_{pb}$ = .20) between gender (males = 1; females = 0) indicating that males had higher self-reported SAT scores. Regardless of the direction, the selection mechanism generated by gender ought to be weaker compared to the selection mechanism generated by the more

direct measure of SAT scores.

The pseudo, gender-based groups had sample sizes ranging from 27 to 31. The male groups had similar math SAT scores of 582 and 598 for the math and vocabulary treatments respectively (see Table 2.4. While the female groups also had similar math SAT scores of 552 and 544 for the math and vocabulary treatments respectively. While there is a consistent math SAT score difference between the male and female groups, the difference is well within the standard deviations for either group. Unlike the groups created by math ability having a small difference between the pseudo gender based groups is preferable. Tables 2.2 and 2.4 show clear differences between the math ability based groups and the pseudo gender based groups with respect to their math SAT scores.

Table 2.4: Gender Selection Group SAT Descriptives

| | | Math SAT | | | Vocabulary SAT | | |
|---|---|---|---|---|---|---|---|
| **Gender-Tx** | n | *M* | *SD* | *SE* | *M* | *SD* | *SE* |
| Male-Math | 30 | 582 | 110 | 20.15 | 541 | 63 | 11.58 |
| Male-Vocab | 27 | 598 | 117 | 22.53 | 556 | 99 | 19.09 |
| Female-Math | 31 | 552 | 97 | 17.42 | 554 | 88 | 15.83 |
| Female-Vocab | 27 | 544 | 90 | 17.33 | 533 | 127 | 24.39 |

Table 2.5: Gender Selection Group ETS Descriptives

| | | Reasoning General-I | | | Vocabulary General-II | | |
|---|---|---|---|---|---|---|---|
| **Gender-Tx** | n | *M* | *SD* | *SE* | *M* | *SD* | *SE* |
| Male-Math | 30 | 8.7 | 2.4 | 0.48 | 23.4 | 3.9 | 0.78 |
| Male-Vocab | 27 | 8.8 | 2.7 | 0.56 | 24.7 | 3.7 | 0.80 |
| Female-Math | 31 | 5.9 | 1.9 | 0.32 | 21.2 | 3.4 | 0.56 |
| Female-Vocab | 27 | 5.6 | 2.1 | 0.37 | 21.7 | 4.2 | 0.73 |

### 2.2.2 Treatments

Shadish's (2008) two educational skill area treatments were left unchanged for the present study. The vocabulary lecture included fifty rare words presented via computer on a large projection screen as well as on each participant's computer screen. Each word was presented with its definition, phonetic spelling, and was used in an example sentence. RA's read each word and each example sentence aloud.

The mathematics treatment lecture described 5 rules for solving exponential equations. Each rule was presented with a title, a description of the rule, and an example. RA's read the rule, its description, and the example aloud from a large overhead screen projection while it simultaneously appeared on each participant's personal computer screen. Participants were not permitted to ask questions at anytime during either treatment presentation or during testing.

## 2.3 Procedure

RA's greeted participants upon their arrival and requested participants' official math SAT scores (e.g., a screen print out from the College Board website, or official transcripts). Subsequently, RA's assigned an identification number to each participant, and provided them

with a pen and a blank piece of paper. Identification numbers indicated the participants' treatment group assignment; treatment group assignment was blinded to participants, but not to RA. Participants were randomly assigned into treatment groups, but were matched based on exact math SAT scores to create equal group sizes. For example, if the first participant with a perfect 800 math SAT score was randomly assigned to the math treatment group, then the next participant with a perfect 800 math SAT score would be assigned to the vocabulary treatment group. After being given identification numbers, participants were directed to the proper treatment room for the start of the experiment. RA's in each condition informed the participants that they would be completing a series of questionnaires and tests. All participants received all the questionnaires and tests for each condition (i.e., math and vocabulary tests), regardless of treatment group assignment. Participants were administered half the items from two untimed tests (Vocabulary Performance and Math Performance Revised), followed by a demographics questionnaire, and two timed tests (Vocabulary General-II, Reasoning General-I). Upon completion of this portion of the experiment, RA's gave either a standard vocabulary instructional treatment lecture or a standard mathematics instructional treatment lecture, depending upon treatment group. Participants were thanked for their participation and were fully debriefed. All study procedures were reviewed and approved by the GMU Human Subjects Review Board.

### 2.3.1 Research Assistants

I trained and supervised two female research assistants (RA's) to administer the study protocol. Training and supervision focused on study logistics and session presentation, including overall clarity, pronunciation, vocal pace and volume. Fidelity was not assessed using formal measurements, though presentation style reminders were given periodically in order to limit the amount of test score variance attributable to the RA. Eleven total sessions were conducted, with RA's alternating between the different treatments.

### 2.3.2 Experimental Rooms

The study was carried out in computer laboratory rooms on the university campus. Sessions took place in several rooms between 12 and 4pm over multiple weeks. All computer laboratory rooms contained desks with individual computer stations. Each station had a computer tower, a 15" LCD monitor, a keyboard and mouse. All computers had access to the Internet. All rooms included an overhead projector, a pull down screen, a desk and a computer for the instructor. The RA used the instructor's station to guide participants to the testing website, presented the instructional lecture slides, and observed participant progress throughout each session. Participants' computer use was restricted during the treatment presentation to minimize distraction. All rooms were equivalently lit and were maintained at similar temperatures. No auditory or visual distractions interrupted any of the experimental sessions.

## 2.4 Measures

During pretest (i.e., observation points 1, 3, 5, and 7 as depicted in Figure 2.2) RA's administered demographic, vocabulary (Vocabulary General-II, Vocabulary Performance) and math (Reasoning General-I, Mathematics Performance Revised) related measures. Vocabulary (Vocabulary General-II) and math (Reasoning General-I) measures created by the Educational Testing Services assessed knowledge for their respective knowledge domains and used in this study as an indicator for the assignment differentiation based on math ability. Two measures (Vocabulary Performance and Mathematics Performance Revised) used in this study were created for assessing the treatments and knowledge conveyed within those treatments. As in the Shadish et al. (2008) study, a randomly selected subset of the items from the Vocabulary Performance (15 of 30) and Math Performance Revised (10 of 20) measures were provided during the first administration. All participants received the same items during the first administration. After the treatment lecture (i.e., observation

points 2, 4, 6, and 8 as depicted in Figure 2.2), participants were administered the complete untimed posttests (Vocabulary Performance and Math Performance Revised) with the items being re-ordered.

**Demographics:** I collected standard demographic data on all participants including math and verbal SAT scores, gender, and participant treatment preference. These variables were partly useful in understanding the effects but may be useful in subsequent modeling procedures outside the scope of the current study.

**Vocabulary General-II:** The Educational Testing Services (ETS) created the Vocabulary General-II (VG-II) as a general vocabulary ability measure for students in grades 7-12[1]. Although typically administered via pencil and paper, this measure was adapted for computer administration, and implemented during pretest only as an additional indicator of vocabulary in the present study. The VG-II consists of 30 multiple-choice items, each with 5 response options. Participants were given eight minutes to complete this test. ETS validated their measure on a national sample of 9th grade students and reported a mean score of 15.5 (SD = 5.5) (Ekstrom, French, Harman, & Dermen, 1976).

**Reasoning General-I:** ETS created Reasoning General-I (RG-I) as a mathematics aptitude test measuring arithmetic or very simple algebraic concepts for students in grades 11-16[2]. The measure was administered, via computer, during pretest only as an additional indicator of mathematics knowledge. The RG-I consists of 15 multiple-choice items, each with 5 response options. Participants were given ten minutes to complete this test. ETS published a mean score of 4.6 (SD = 3.6) for those items (Ekstrom et al., 1976).

**Vocabulary Performance:** The Vocabulary Performance (VP) is an untimed 30 item multiple-choice measure (See Appendix B) developed specifically to assess the vocabulary

[1]The VG-II - one part of the Manual for Kit of Factor-Referenced Cognitive Tests - is available to license from ETS Research & Development.

[2]The RG-I - one part of the Manual for Kit of Factor-Referenced Cognitive Tests - is available to license from ETS Research & Development.

treatment in Shadish et al. (2008)'s study. As described above it was administered at pre and posttest as a measure of the vocabulary treatment (see Figure 2.1). Each item required participants to select the definition of a single word from 5 options. Participants were instructed to select the correct definition.

**Mathematics Performance Revised:** The Mathematics Performance Revised (MPR) is an untimed 20 multiple-choice item measure (See Appendix C) developed by Shadish to assess their mathematics treatment. As described above it was administered at pre and posttest as a measure of the mathematics treatment (see Figure 2.1). Each item presents a basic exponential problem (e.g., $(x^a)(x^b) =$) with 5 possible solutions; the participant selected the most correct solution.

The original 20 item Mathematics Performance (MP) test was revised (hence, the acronym MPR) for this study because several of the original items had multiple correct answers or were unrelated to the exponent content. To revise the measure, pilot data was collected at the site of the current study. The measure was piloted with 71 individuals who did not participate in the full study. The pilot study differed from the full study in two ways: 1) participants were included without reference to their math SAT scores, and 2) rather than randomly assigning participants to receive a treatment, sessions were randomly assigned such that on a given day, all participants in the pilot session received the same treatment. Two pilot sessions were conducted. Responsive items were identified using the pre-post change in their Rasch item difficulty. Winsteps version 3.68.2 was used for the item analysis (Linacre, 2011). Six items that decreased in difficulty from pretest to posttest were retained because they showed a treatment effect. All six of the treatment responsive items focused on the treatment content. Uninformative items required outside knowledge that was not included in the treatment (e.g., What is the volume of a cylinder?). Fourteen additional items were written and added to the protocol, creating a total of 20 items. New items matched the retained items by focusing on treatment related information.

### 2.4.1 Measuring Effects

Both aforementioned aims required effect size computations. I began with Cohen's (1992) formula for calculating effects using d. That formula is based upon the z-score where the difference between two means (e.g., treatment and control) gets divided by the pooled standard deviation (see equation 2.1).

$$d = \frac{mean(x) - mean(y)}{\sqrt{\frac{var(x)+var(y)}{2}}} \tag{2.1}$$

Aim 1 focuses on estimating the bias effects for both direct (math) and indirect (gender) selection variables. As a result, it is possible to calculate 2 effect sizes measuring selection bias from different domains. The experimental design (see Figures 2.2 and 2.3) has a clear demarcations for 2 levels (i.e. high and low math, male and female) for each selection variable. Thus, each selection variable can have a maximum effect via participants in the High-Math group compared to the Low-Vocab group and ought to produce the largest selection bias effect for the direct selection variable. In contrast, the minimum effect comes from participants in the Low-Math group compared to the High-Vocab group and that comparison ought to produce the smallest selection bias effect for the direct selection variable.

Each effect size of interest was calculated using Cohen's d using the formula listed above. For example, the maximum selection bias effect for the direct (i.e., Math SAT) selection variable was computed by subtracting the mean MPR scores at posttest ($O_2$) of the participants with the high math ability that received the math treatment from the mean MPR scores at posttest ($O_8$) of the low math ability participants that received the vocabulary treatment. That mean difference then was divided by the pooled average differences of both of those groups. The resulting number is a measured effect for a selection variable.

## 2.5 Analysis

Preliminary analyses checked the manipulation of differences between the direct and indirect selection variable groups' SAT scores and the treatment effect on those groups. Analyses produced 8 selection bias effects for Aim 1 and a resampling procedure on those calculated selection bias effects creating selection bias effects distributions and attempted to predict selection bias magnitudes based upon selection variables for Aim 2. All analyses were conducted using the statistical program R version 2.11.1 (R Development Core Team, 2011). Computations were also completed in R using the R programming language. Specifically, the effect size estimation used a simple function to calculate Cohen's d and the bootstrap (i.e., resampling) estimates used a separate function that called the effect size function. All R code used in the present study appears in the Appendix D.

### 2.5.1 Treatment Effectiveness

The focus of this study is on measuring selection bias. Despite that focus, I chose to include an effect size estimate for the math treatment so that the effects estimated for both aims could be compared to the observed treatment effects manipulated in this study. To estimate these effects, all participants that received the math treatment were combined into a single group and all the participants that received the vocab treatment were combined into a separate single group. The Cohens d was calculated for the math treatment effectiveness measured using the MPR by comparing those that received the math treatment to those that received the vocab treatment.

$$\text{Tx ES} = \text{Math Treatment Groups} - \text{Vocab Treatment Groups} \tag{2.2}$$

### 2.5.2 Aim 1: Estimating Selection Bias Effects

I calculated effect sizes to meet my first aim of estimating the magnitude of selection bias effects for direct and indirect variables. Because there are 2 different observation periods (pretest and posttest) for each experimental group and 2 different selection variables (direct

(i.e. math) and indirect (i.e. gender)) and 2 levels for each selection variable (i.e. high and low math ability, male and female gender) this results in a total of 8 unique effect sizes for the first.

### Pretest and Posttest Effects

A maximum difference for the direct selection variable groups was calculated using the difference of MPR scores between High-Math compared to the Low-Vocab group for both pretest and posttest effects. Comparing the scores from the participants with the high math ability receiving the math treatment to the low math ability participants receiving the vocabulary treatment should produce the greatest or maximum difference scores or the maximum direct selection bias, after the treatment effect's removal (see Equation 2.3). The smallest selection bias effects were calculated to derive from the difference in MPR scores between Low-Math and High-Vocab groups for both pretest and posttest effects. Comparing the scores from the participants with the low math ability receiving the math treatment to the high math ability participants receiving the vocabulary treatment should produce the smallest or minimum difference scores or the minimum direct selection bias, after the treatment effect's removal (see Equation 2.4).

Indirect selection variable based comparison groups used the difference between Male-Math and Female-Vocab groups as the maximum for both pretest and posttest effects with Male-Vocab and Female-Math groups as the minimum for both pretest and posttest effect(see Equations 2.5 and 2.6). These calculations mimicked the respective direct and indirect selection bias equations with gender category male substituting for high ability.

All posttest Effect Sizes (ES's) reflect that the treatment ES was subtracted out. Thus, the posttest effects represent the biased effects rather than a combination of treatment and biased effects. Previous studies used raw difference scores to calculate selection bias point estimates - thus, making it difficult to compare effects across studies. As mentioned previously, all ES presented in the current study used Cohen's d formula (see Equation 2.1)(Cohen, 1988).

$$\text{Maximum Direct Selection Bias} = \text{(High-Math - Low-Vocab) - Tx ES} \quad (2.3)$$

$$\text{Minimum Direct Selection Bias} = \text{(Low-Math - High-Vocab) - Tx ES} \quad (2.4)$$

$$\text{Maximum Indirect Selection Bias} = \text{(Male-Math - Female-Vocab) - Tx ES} \quad (2.5)$$

$$\text{Minimum Indirect Selection Bias} = \text{(Male-Vocab - Female-Math) - Tx ES} \quad (2.6)$$

### 2.5.3 Aim 2: Estimating Distributions of Selection Bias Effects

The second aim focused on producing a distribution around the effects estimated in Aim 1. I provide more details about the precise procedures below.

#### Distributions of Selection Bias Effects

I created pretest and posttest selection variable distributions by resampling the observed selection bias magnitude point estimates. Outcome variable (MPR) scores from the high and low math ability groups represent the extremes from the observed sample. The above selection bias point estimates were re-calculated using the newly created resampled data with the resampling allowing for filling in the sample distribution between the observed extreme scores. The selection bias magnitude distributions were calculated to establish the variability around the observed point estimates. If the resampled distributions fit known distribution properties (e.g., Gaussian, Poisson, Gamma, etc.), then the probability of each selection bias magnitude can be reliably estimated. The minimum and maximum magnitude point estimates were used as starting points for creating a distribution.

A bootstrapping or resampling procedure (random selection with replacement) was used

to fill in the observed point estimates to create the overall distribution. Each of the four selection bias effect sizes (Maximum Direct Selection Bias, Minimum Direct Selection Bias, Maximum Indirect Selection Bias, Minimum Indirect Selection Bias), for both pretest and posttest, received the following resampling procedure. For example, the maximum direct selection bias point estimate was the difference between the High-Math and the Low-Vocab groups, the resampling code started at 10% of the High-Math group scores being replaced with the Low-Math group scores. This process continued increasing replacement by 10% until the final replacement level in which 90% of the High-Math group scores were replaced with Low-Math group scores. The resampling occurred at 10% intervals from 10 - 90% replicated 1,000 times for a total sample of 9,000 resampled effect sizes per selection bias effect size (See Appendix D). The entire replacement process occurred again for the minimum direct selection bias pretest estimate. Following this, both distributions (direct and indirect) were aggregated to create a total of two selection bias distributions, one for pretest and one for posttest.

**Ancillary Analyses: Predicting Effect Sizes via a Regression Model**

An additional bootstrap method was employed to assess the predictive validity of the selection mechanism and the selection bias effect size. This step served as both 1) a check on the hypothesis that direct and indirect selection variables can impact outcome results (see Figure 2.1), and 2) a technique for a practical guide for researchers. The regression model was computed to see if there were a simple calibration between selection mechanism correlation and selection bias. Consider two extreme examples to clarify the use of this method. A selection mechanism that is truly random ($r_{selectionmechanismvariable,pretestscores} = 0$) ought to produce no selection bias (i.e., no differences between groups at pretest). In contrast, a selection mechanism directly and perfectly related to the pretest scores and the treatment ($r_{selectionmechanismvariable,pretestscores} = 1$) ought to produce the maximum effect. We might infer the regression slope by just joining those two extremes with a line and assume no error. Instead, we might fit a line of best fit between those extremes to see if there exists

a calibration curve representative of the predictive validity of selection mechanisms and selection bias. The advantage of this procedure is that it provides a heuristic for future work but it also empirically assesses the relationship between selection mechanism and selection bias.

If there is a distinction between direct and indirect selection variables then the outcome scores should be predictable. The direct selection variable has a .5 larger correlation to the outcome measure than the indirect variable (see Table 2.1) and difference should produce distinct utilities for researchers aiming to predict the amount of potential bias. A bivariate linear regression using the correlation between selection variable (i.e., math ability and gender) and MPR pretest scores predicted selection bias magnitudes.

The data for the regression came from a resampling procedure whereby all high math ability participants were randomly replaced by low math ability participants. In essence, the composition of the sample changed with each bootstrap but the composition changed according to a gradient of correlation ranging from an expected low correlation ($r = 0$, perhaps) to a high correlation ($r \approx 1$). The more high math ability participants included in the sample, the more direct the selection mechanism and, presumably the stronger the selection bias. I used the same resampling procedure for gender to ensure that there was a wide range of correlations between the selection mechanism and the pretest scores. For both resampling procedures, there were 9 levels of selection. Level 1 contained a sample comprised of 10% of the high math ability participants and 90% of the low math ability participants assigned to the math treatment and the opposite composition for the vocabulary treatment. If the selection mechanism favored the vocabulary group with respect to the most direct bias in the study, I expected the effect to be suppressed. In contrast, as the levels increased, I expected the greater shift of high math ability participants toward the math treatment would favorably bias the treatment effect.

# Chapter 3: Results

All preliminary and primary data analyses were from fully observed data from all participants (N = 115). No participants withdrew from the study, and no missing data were observed in the study.

## 3.1 Treatment Effectiveness

### 3.1.1 Ability as a Direct Selection Variable

Math performance was equivalent for the two treatment groups within the different ability categories at pretest. Posttest scores show divergent performances based upon received treatment and ability (see Table 3.1). Figure 3.1 shows a plot of the percentage of correct MPR items by ability comparison group. Regardless of ability group, those that received the math treatment performed better overall at posttest than those that received the vocabulary treatment. Those in the math treatment had greater pre-post increases than vocabulary treatment participants. The treatment effect was d = 1.08 using Equation 2.2.

Table 3.1: Direct Selection Group MPR Descriptives

| | Pretest | | | Postest | | |
|---|---|---|---|---|---|---|
| **Ability-Tx** | *M* | *SD* | *SE* | *M* | *SD* | *SE* |
| High-Math | 7.0 | 1.9 | 0.39 | 16.4 | 2.6 | 0.52 |
| High-Vocab | 6.4 | 2.5 | 0.53 | 12.8 | 5.2 | 1.12 |
| Low-Math | 3.9 | 2.1 | 0.35 | 11.7 | 3.5 | 0.58 |
| Low-Vocab | 3.2 | 1.9 | 0.34 | 6.1 | 2.1 | 0.37 |

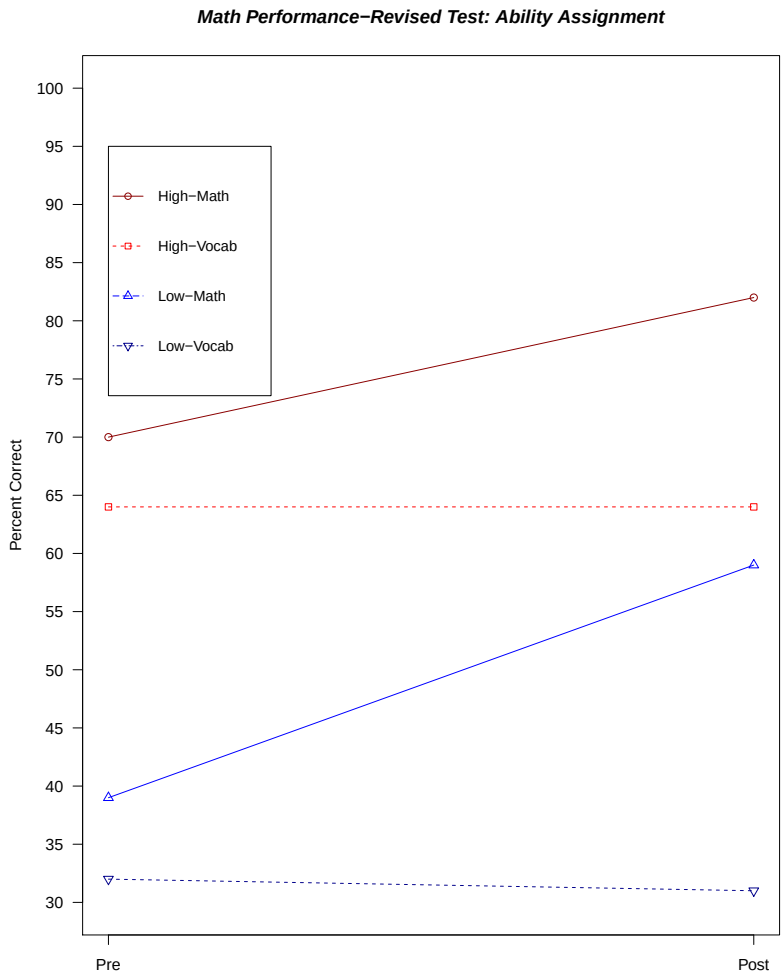



Figure 3.1: MPR for the direct selection variable

### 3.1.2 Gender as an Indirect Selection Variable

Math performance was equivalent for the two treatment groups within the different gender categories at pretest. Posttest scores show divergent performances based upon received treatment and gender (see Table 3.2). Figure 3.2 shows a plot of the percentage of correct MPR items by gender comparison group. Regardless of gender group, participants that received the math treatment performed better overall at posttest than those that received the vocabulary treatment.

Table 3.2: Indirect Selection Group MPR Descriptives

| | Pretest | | | Postest | | |
|---|---|---|---|---|---|---|
| **Gender-Tx** | *M* | *SD* | *SE* | *M* | *SD* | *SE* |
| Male-Math | 5.5 | 2.6 | 0.47 | 14.1 | 3.9 | 0.71 |
| Male-Vocab | 4.7 | 2.9 | 0.57 | 9.9 | 5.6 | 1.08 |
| Female-Math | 4.9 | 2.5 | 0.44 | 13.2 | 3.9 | 0.70 |
| Female-Vocab | 4.3 | 2.4 | 0.46 | 7.8 | 4.1 | 0.78 |

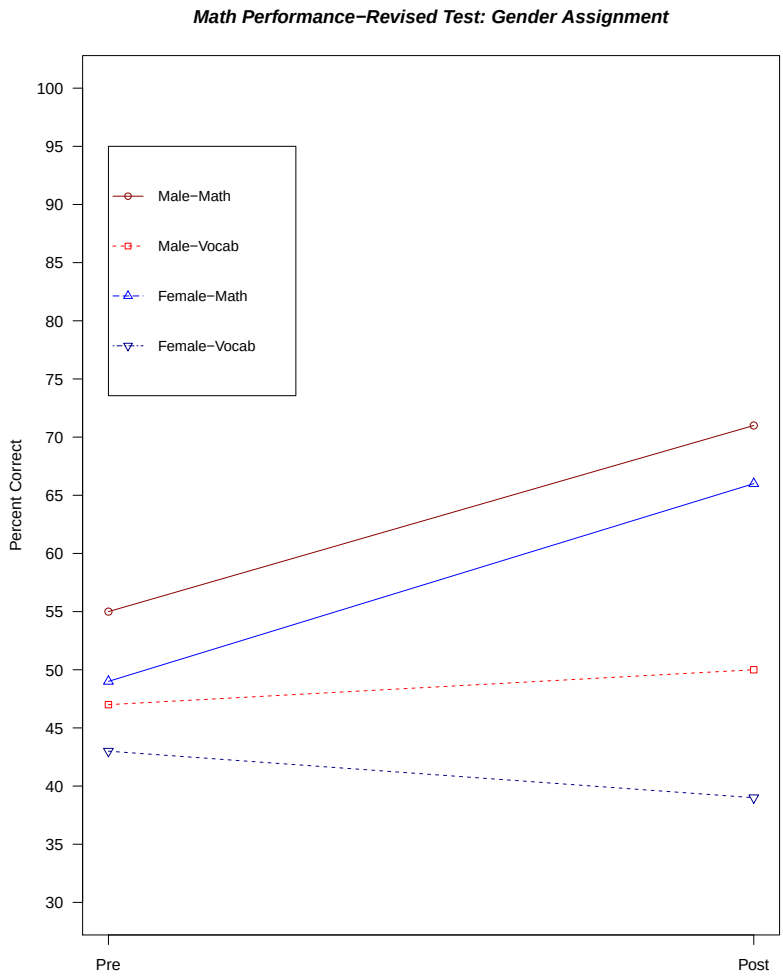



Figure 3.2: MPR for the indirect selection variable

## 3.2 Aim 1: Estimating Selection Bias Effects

### 3.2.1 Pretest Effects

Minimum pretest effects were larger for the direct than the indirect selection variable (see Table 3.3). In fact, the minimum effect for direct selection variable was greater than the guideline for a large ES established by Cohen (1992). The range between the minimum (1.08) and maximum (1.97) selection bias ES's was nearly an entire Cohen's d standardized

unit for the group assigned by their math ability. The indirect selection variable group had pretest ES's ranging from small (-.06) to moderate (.51) by Cohen's standardized effects. The maximize ES from the indirect selection variable group was half the magnitude of the minimum ES from the group assigned by direct selection variable.

Table 3.3: ES for Pretest Effects

| | Minimum | Maximum |
|---|---|---|
| **Direct** | 1.08 | 1.97 |
| **Indirect** | -0.06 | 0.51 |

### 3.2.2 Posttest Effects

All reported posttest effects are after the removal of the treatment effect. The minimum (-.84) posttest effect for the direct selection variable decreased from the pretest effect while the maximum (3.27) posttest effects increased from pretest (see Table 3.4). Direct selection variable groups had mostly larger effect sizes than the indirect selection variable based groups for the posttest effects. The minimum (-.84) posttest effect for direct selection variable was smaller than the minimum (1.08) direct selection variable pretest effect. This smaller effect seems to be related to the treatments effectiveness, since those that received the math treatment improved their MPR scores. While the effect size for the direct selection variable group increased its range due to changes at both the minimum and maximum effects the same is not what happened for the indirect selection variable.

The indirect selection variable posttest effects did not change as dramatically as the direct selection variable posttest effects. In fact the maximum (.51) indirect selection variable effect remained the same as it was at pretest but the minimum decreased from -0.6 to -1.77. Note the negative effects. I originally hypothesized that females would out perform males on the mathematics test but the results suggested exactly the opposite. Since I expected

females to outperform females, I created a binary coding for males = 1 and females = 0 and computed effect sizes assuming females would be higher. Nevertheless, my intent was not to test gender effects but rather to assess absolute magnitudes of selection bias effects. This difference in the indirect selection variable effect was likely due to the effectiveness of the treatment and demonstrates that gender, an indirect variable was not as influential on results as a direct selection bias variable (as well as the data coding - male scores subtracted from female scores). The relatively small maximum indirect selection variable posttest ES was still a medium sized effect by the standards established by Cohen (1992).

Table 3.4: ES for Posttest Effects

| | Minimum | Maximum |
|---|---|---|
| **Direct Selection Variable** | -.84 | 3.27 |
| **Indirect Selection Variable** | -1.77 | .51 |

## 3.3 Aim 2: Estimating Distributions of Selection Bias Effects

### 3.3.1 Distributions of Selection Bias Effects at Pretest

Distributions for the direct and indirect selection variables at pretest are the result of resampling from within the observed effect sizes (see Figures 3.3a and 3.3b). Neither distribution (direct or indirect) conformed to Gaussian or Gamma or Poisson distributions according to Kolmogorov-Smirnov tests. The distributions do show relatively similar means and medians between the direct and indirect selection variable (see Table 3.5). While the distribution averages lean towards the small effect size according to Cohen their standard deviations bring selection bias effects well past the medium classification (Cohen, 1992). However, the extremes and the ranges between the two continue to be quite different (see Figures 3.3a and 3.3b). The direct selection bias mechanism ability has a standard deviation about two

and half times the range of the indirect selection bias mechanism gender. These wide ranges for both selection mechanisms mean the minimum and maximum effects go well beyond the observed point estimates.

Table 3.5: Resampled Pretest Effects Descriptive Statistics

| % Resampled | $ES_M$ | $ES_{Mdn}$ | $ES_{SD}$ | $ES_{SE}$ |
|---|---|---|---|---|
| **Ability** | .30 | .27 | .75 | .01 |
| **Gender** | .29 | 0.28 | 0.30 | .00 |

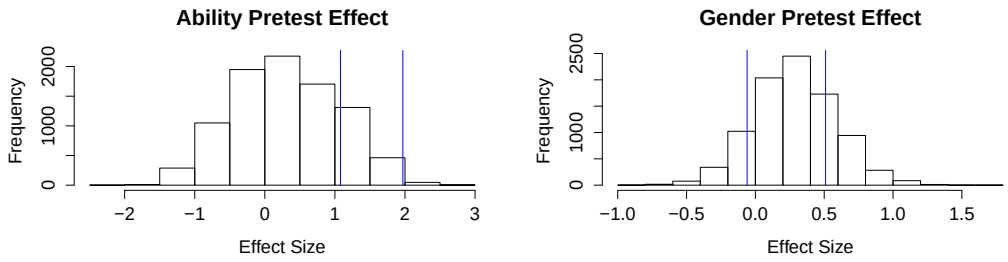


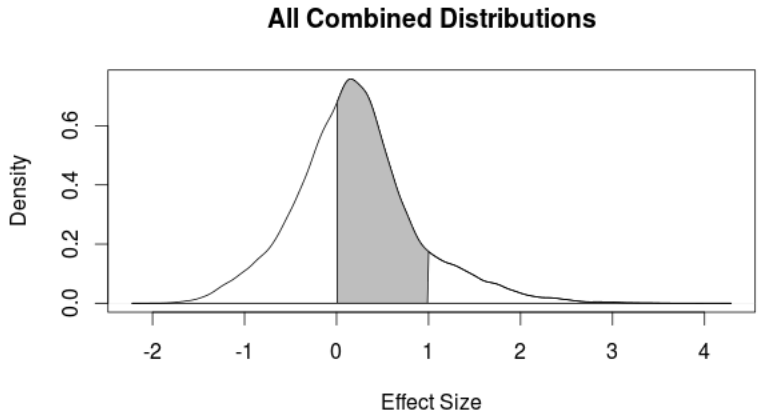

Figure 3.3: Pretest effect size distributions: The two vertical lines are the observed minimum and maximum ES point estimates.

### 3.3.2 Estimating Area Under the Curve (AUC) for probability

As I mentioned in the introduction, estimating the area under the curve (AUC) is an important step to realizing the benefits of the estimate and subtract (or E & S) method. The following figure (Figure 3.4) shows the combined distributions from all resampled statistics. Additionally, the area highlighted under the curve represents the probability of an effect to be greater than 0 - assuming that there is a directional assumption of selection bias - but less that 1. If the resampling method produced a normal distribution then I could easily estimate that area by standard functions built into almost every contemporary statistical

software package. As I noted previously, my results did not produce any known distribution; thus, I need an alternative solution to standard lookup functions. One alternative to these standard functions is a numerical method of approximation - usually by a Reimann sum or Euler method other estimation. Both Reimann sums and the Euler method produce reasonable estimates, however, they fail to produce a generalizable result that I aimed for in this study. I discuss these implications later in the discussion.

Figure 3.4: Combined Distributions AUC




### 3.3.3 Distributions of Selection Bias Effects at Posttest

Distributions for the direct and indirect selection variables at posttest are the result of resampling from within the observed effect sizes (see Figures 3.5a and 3.5b). Neither distribution (direct or indirect) conformed to Gaussian or Gamma or Poisson distributions according to Kolmogorov-Smirnov tests. While there is a greater difference between the two posttest distributions the overall observed selection bias effects are reduced from pretest (see Table 3.6). This is likely due to the effectiveness of the treatment (d = 1.08). And while the averages are negligible the ranges for both the direct (6.01) and indirect (3.17) at posttest are greater than their corresponding direct (4.96) and indirect (2.56) pretest ranges, respectively (see Figures 3.5a and 3.5b).

Table 3.6: Resampled Posttest Effects Descriptive Statistics

| % Resampled | $ES_M$ | $ES_{Mdn}$ | $ES_{SD}$ | $ES_{SE}$ |
|---|---|---|---|---|
| **Ability** | .18 | .01 | .98 | .01 |
| **Gender** | .06 | .02 | .41 | .00 |

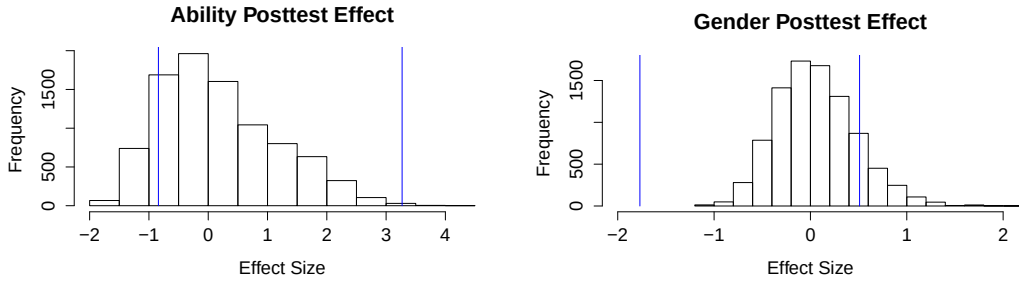



Figure 3.5: Posttest effect distributions: The two vertical lines are the observed minimum and maximum ES point estimates.

### 3.3.4 Predicting Selection Bias via Selection Mechanism

Creating distributions for obtaining selection bias effect size probabilities was in part an attempt at a predicting the magnitudes for the observed effects. Given that none of the resampled distributions fit known distributional properties a secondary method of predicting the selection bias magnitudes was implemented. This prediction method does combine the direct and indirect selection mechanisms assuming that they are not orthogonal classifications but lie along a continuum. Figure 3.6 shows the relationship between the observed correlation for selection (i.e., the correlation between the selection variable and MPR pretest score - a measure of the observable and available selection bias indicator) and the resampled ES distributions. As can be seen in by the difference between the blue cross-hair points and

the empty red points the distributions between the two selection mechanisms had substantial variation in the effect sizes (as represented on the y-axis). Examining the plot and the regression lines it is possible to visualize the impact of the selection mechanism and pretest correlation as impacting the observed selection bias effect. Using a quadratic regression the correlation between selection variable and MPR pretest score significantly predicted resampled ES, $b_1 = -.6, b_2 = 1.6, R^2 = .05, F(2, 17997) = 459, p < .001$.

Assuming the minimum selection bias ought to be 0 (i.e., the regression passes through the origin), the maximum selection bias effect ought to be 1.0 - not quite as high as I estimated in Aim 1 (d=3.27; see above).

$$\text{Posttest Effect Size} = -0.6 + 1.6 - .00 * (\text{selection variable-pretest correlation}) \quad (3.1)$$

This model suggests researchers may use simple correlations to estimate the posttest bias but with somewhat poor precision ($R^2_{adj} = .05$).

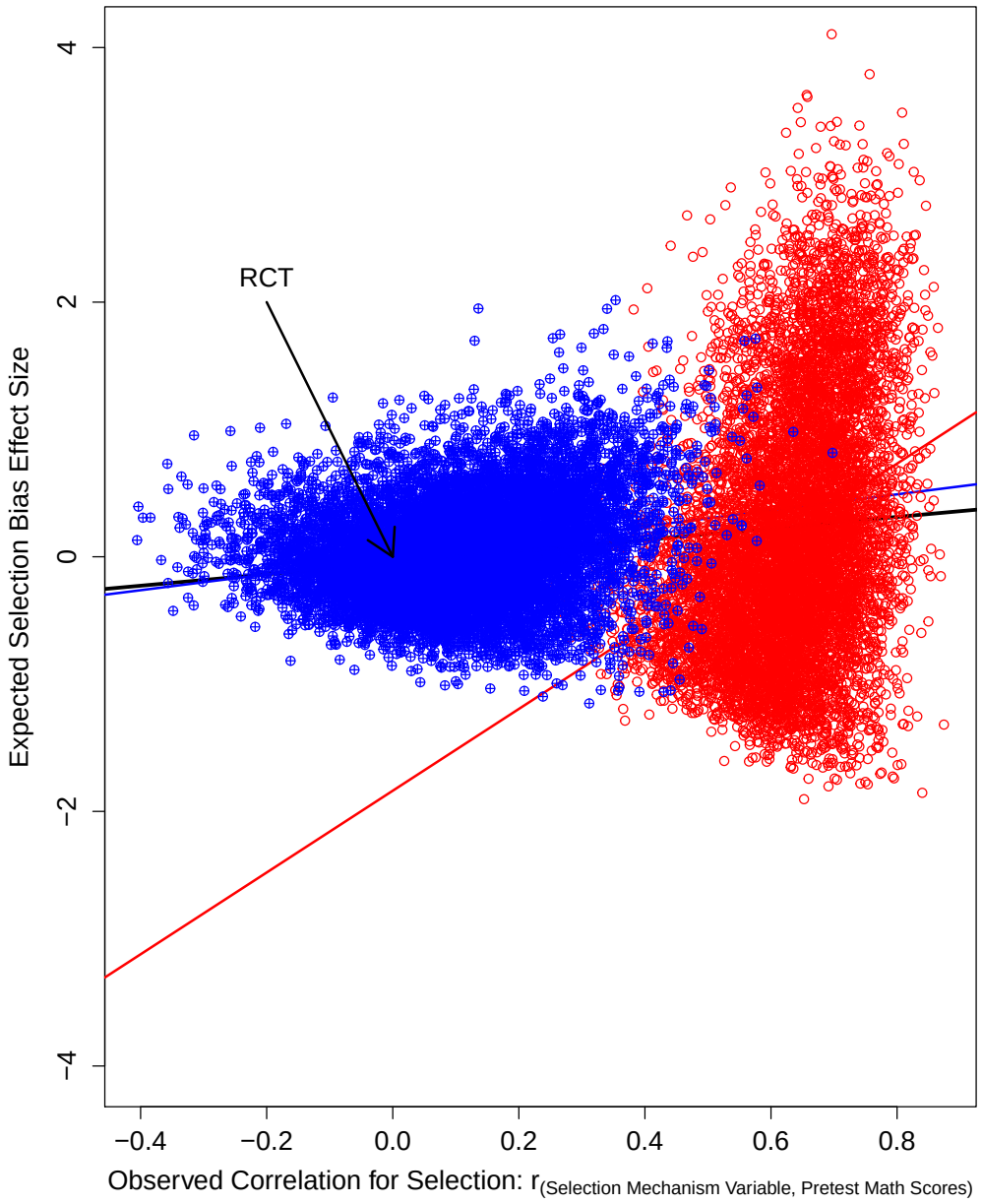



Figure 3.6: The black line is the slope for the combined direct and indirect groups. Direct data points are in red empty points. Indirect data points are in blue cross-hair points. The "x" and "+" points are the posttest point estimates.

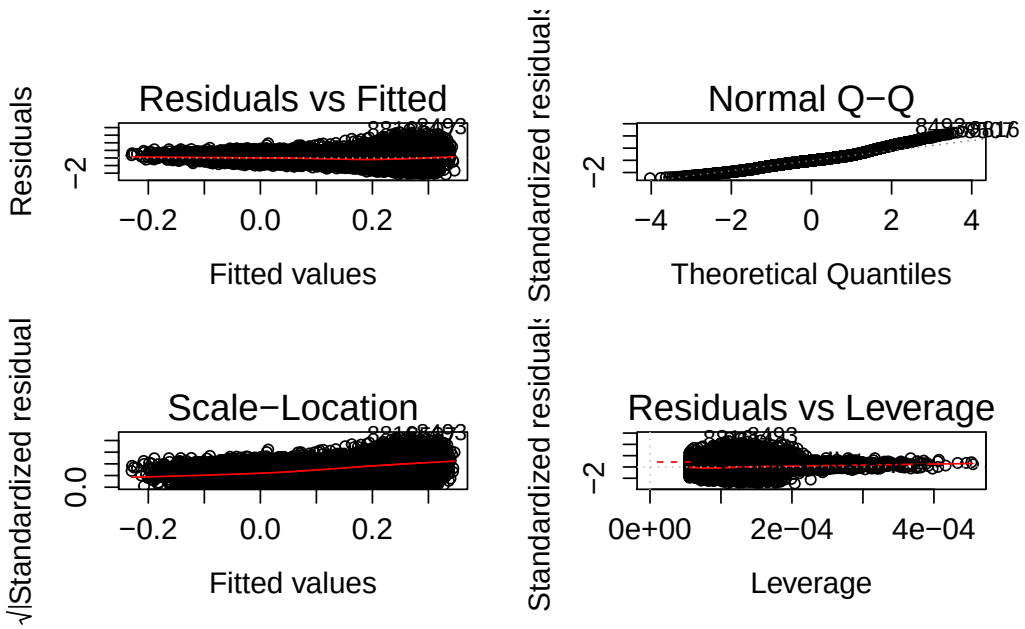



Figure 3.7: Regression diagnostics plot

While this regression may be useful, the regression diagnostics indicate that multiple assumptions were not met. Specifically, the plot and regression diagnostics clearly show heteroskedastic errors along with non-normally distributed errors. These two assumption violations limit the utility of the prediction model; I addressed these limitations later in the discussion.

# Chapter 4: Discussion

The two aims of this study were to 1) estimate selection bias and 2) to estimate the probability distributions around those estimates. Overall, both aims were met. I address each one in turn below along with their implications.

## 4.1 Aim 1: Estimating Selection Bias Effects

Selection variables created bias across conditions at both pretest and posttest. Recall that I created selection bias according to a direct, ability-based selection mechanism (i.e., SAT scores) and an indirect, proxy-based or pseudo selection mechanism (i.e., gender). The purpose of these two selection mechanisms was to deliberately produce greater variability in the selection bias estimates.

### 4.1.1 Direct Selection Mechanism

Selection based on ability resulted in large bias magnitudes at both pretest and posttest - indicating that non-equivalent group designs (i.e., selection bias as a threat to internal validity) produces both differences before treatments as well as afterwards. The pre-treatment effects are precisely what defines selection bias whereas the post-treatment effects are what define the treatment by selection interactions. All calculated bias magnitudes represent large effects (minimum d = .8; maximum d = 3) in the ability-based selection condition. Again, these effects were observed at pretest and at posttest. Although direct selection variable groups produced similar absolute minimum magnitudes at both pretest and posttest, maximum magnitudes increased (from 1.97 to 3.27) from pretest to posttest indicating that the treatment by selection interaction added unique variance to the post-test results. What

can be surmised by these results is that directly relevant differences - that is, differences between groups that are relevant to the treatment and the treatment outcome - may produce an additional standard deviation difference in selection biases and selection by treatment biases.

The effects observed by this direct selection mechanism were large. According to Cohen's rough, qualitative criteria (Cohen, 1992), these effects represent large to extremely large effects. To put these selection bias effects into perspective, the estimated treatment effect was d=1.08. If selection bias - via a direct mechanism - were present then the observed effect at posttest might be as large as d=4.35 provided that these effects were perfectly additive. Thus, the selection bias effects ranged between roughly 78% to 303% of the treatment effects. These percentages were calculated as a function of the selection bias estimate divided by the treatment effect. Taking these estimates into consideration and not generalizing to all treatments, I might consider selection bias to be roughly 1 to 2 times the observed treatment effect. These rough estimates pertain solely to the direct selection mechanism. Now let us consider the indirect effects.

### 4.1.2 Indirect Selection Mechanism

The results for the indirect selection mechanism paralleled the direct mechanism results but the effects were generally smaller. The indirect selection mechanism based on gender resulted in small to medium magnitudes at pretest and small to large magnitudes at posttest. Gender groups produced vastly different minimum magnitudes at pretest (range: d = -.06 to -1.77), while maximum magnitudes decreased remained a medium magnitude at pretest and at posttest. The effects observed for this weaker, indirect selection mechanism resulted in precisely what I did expect - weaker selection bias effects. Females assigned to the mathematics treatment outperformed males assigned to the vocabulary treatment on the MPR measures, which resulted in negative bias magnitudes.

The estimated selection bias effects were roughly half of the effects observed by the direct selection mechanism. Recall that the point-biserial correlation between gender (male = 1,

female = 0) and ability was relatively low ($r_{pb} = .20$); the 50% drop-off in effect size was not entirely expected. I expected the effects would be substantially lower since only about 5% of the variance is shared between the direct and indirect mechanism ($CV = r^2_{pb} = 0.05$). The fact that the effects were only reduced by one-half indicates that selection bias may have complex effects that are not adequately captured in these single-variable mechanisms. As it was shown here and in previous research the indirect selection variables either produce or explain far less selection bias than direct selection variables (Cook & Steiner, 2010). The production of selection bias may relate to how an indirect selection mechanism interacts with other variables affecting selection bias magnitudes. This relationship between the indirect selection variable and other potential variables is still unknown. Ideally the selection mechanism and how it relates to the outcome variable is clear and understood but when the indirect selection variable has no simple casual pathway to the outcome variable its influence is less well understood and less predictable. The comparison I made between selection bias effects and treatment effects may be applied to these indirect effects as well. If I consider only the absolute magnitude of effects instead of the directional effects, I estimate the relative magnitude of selection bias to treatment effects would be roughly 6% to 160%. Note that these ratios are not precisely one-half of the observed direct selection effects but rather a wider range between the minimum and maximum values - a point we shall return to when comparing the distributions.

## 4.2 Aim 2: Distributional Properties of Selection Bias Effects

The second aim of this study focused on estimating the distributions around the effects just mentioned. To refresh the reader, I conducted a resampling procedure via bootstrapping to estimate a sample of selection bias effects from the observed data. Those sampling distributions provided not just the range of effects - as discussed above - but also a reasonable shape around a measure of central tendency (e.g., mean) and the general dispersion about

that single value. The distributions produced by the direct and indirect selection mechanisms in Figures 3.3a, 3.3b, 3.5a, 3.5b appear similar but the x-axis suggests otherwise. Direct selection mechanism samples produced a much wider distribution - compared to the indirect selection mechanism sample distribution - indicating less certainty about any single expected parameter. While the effects were largest for the direct mechanism, those effects are less stable.

An implicit goal of the second aim was to help social scientists make more empirically informed decisions for prioritizing threats to validity. Being unable to mimic known distributions impacts the utility of the classic Expected Value (EV) computation that would assist researchers in prioritizing validity threats (Neumann & Morgenstern, 1944). A researcher, using EV, might consider the resources necessary to treat a threat as well as the impact that the specific threat may have and judge which threat deserves their attention and resources. In order to calculate the EV, the effect and its probability are required (EV = P * ES). After summing the products of the effect distribution the resulting number is an effect size estimate that may be used to compare against other threats. Usually the larger and more positive the EV, the stronger the impetus to act. Comparing different threats is a straightforward approach to weighing decisions and their potential ramifications.

Despite my efforts to characterize the distributional properties of the selection bias mechanism, I was not able to ascertain the mathematical function that explained the sampling distributions. Most bootstrapped estimates conform to a normal distribution (Efron & Tibshirani, 1993) and the distributions produced by the resampling procedure appeared normal but failed to comply with the expected distributional properties tested via standard tools (e.g., Kolmogorov-Smirnov test (Chakravarti & Roy, 1967)). Fitting a specific function to these distributions would be helpful for future work so that the probability of any point estimate may be incorporated into any treatment design. Consider the following scenario as an application of this logic. A researcher wishes to conduct a study where the expected effect of a treatment is large ($d > 1.0$). If selection bias were relevant (i.e., random assignment might not be possible due to logistics) but small due to the indirect nature of

the pre-treatment differences, then the researcher might be able to ascertain the expected value of a small effect using a known distribution. Based upon my results (see Figure 3.3a), a selection bias effect greater than d = .5 or greater only occurred roughly 3000 times out of the 9,000 samples or 1/3rd of the time. My estimate of 3000 comes from adding the heights of the bars that include .5 and above in Figure 3.5a. Thus, the probability of a .5 selection bias effect was .1 and the expected value is .5 × .33 or .17. If I were that researcher, I might focus on other threats than the weak, potential impact of selection bias with an expected effect of .17. The distributions observed in this study allow for rough estimates but if I were able to produce a distribution with certain mathematical properties, then we may apply more robust estimates of area under the distribution curve (i.e., the probability) for more generalizable estimates that can be readily computed for any sample.

My failure to produce a known distribution only weakens the potential impact of the E & S method. I say weakens because there are countless methods for estimating area under the curves (AUC) for known and unknown distributions. For one, numerical integration methods like Reimann sums and the Euler method (among many others) represent a more brute force method for estimating AUC. An alternative is a Bayesian procedure (i.e., MCMC, Metropolis Hastings, Gibbs sampler, etc.) that enables the analyst to estimate these probability density functions without any specific known properties. While the numerical and Bayesian methods are readily available, they both require sufficient mathematical, statistical, and programming knowledge that weaken their appeal to most social scientists. My aim to produce a known distribution was to enable all social scientists to estimate the magnitude and probability of selection bias threats. Failing to produce that known distribution does not eliminate my contribution but it does detract from my aim to simplify a procedure and allow the E&S to gain further ground in social science.

## 4.3 Relating Selection Mechanism and Selection Bias

As previously theorized and empirically supported, different selection variables showed vastly varying selection bias magnitudes. The ability selection variable was directly related

to both the treatment and to the outcome measure, while the gender selection variable was indirectly related to treatment and outcome, as evidenced by the larger correlation observed between math ability and pretest MPR scores relative to the correlations observed between gender and pretest MPR scores (see Table 2.1). Additionally, the ability selection variable (SAT scores) significantly predicted the posttest bias magnitudes. Taken together, this finding lends support to the importance of treatment and outcome related selection variables as contributors to bias in NR study results (Cook et al., 2009; Steiner et al., 2010).

When trying to predict selection bias by the relevance (i.e., correlation) of the selection mechanism, however, I was unable to produce a strong prediction model. The prediction model produced significant results but the overall predictive validity remained low ($R_{adj} = .05$). As indicated by the diagnostic plots, the residuals were heteroskedastic and non-normal. Heteroskedasticity limits the prediction model because the utility of the results depends upon the level of the predictor. As the correlation between the selection variable and the pretest scores increased, the residuals increase - indicating lower predictive validity. We assume that a prediction model works equally for all levels of our predictors but, in the case of predicting selection bias effects, my resampling results indicate the opposite. Furthermore, normally distributed residuals are essential for hypothesis testing. Given the large sample size from the boostrap procedure, neither power nor hypothesis testing were relevant. What remained relevant, however, was the fact that the parameters may be adversely affected by the non-normally distributed observations. The Residual/Leverage plot provides some evidence that the non-normally distributed residuals had little impact (see Figure 3.7). In sum, the predictive validity of these available correlations leads to rather poor prediction of selection bias effects.

## 4.4 Limitations

As mentioned previously, the current study focused on two aims and achieved the first with some success with the second. Despite these successes, there are some limitations worth

noting. The present research represents an improvement over current knowledge regarding selection bias estimates (Aim 1) and their associated probabilities (Aim 2), but the study has some generalizability limitations due, in part, to six potential areas worthy of further consideration and study.

### 4.4.1 Measurement

Selection bias magnitudes and treatment effect sizes calculated in the current study were quite large compared to those typically observed in NR studies. Moreover, the treatment effect observed in the present study was considerably larger than that observed in the Shadish et al. (2008); this finding may have been due to fact that the outcome measure was much better than typical outcome measures. The treatment and the outcome measure were extremely focused. Shadish et al. (2008) allowed for more extraneous error variance in their outcomes and their effects might have been lower due solely to these measurement differences.

### 4.4.2 Treatment Fidelity

Further, the present study had a low participant demand - lasting for fewer than 2 hours. Although typical in laboratory studies, such low participant demand may be less common in field and NR studies limiting any external validity claims. Thus, the selection bias effect estimates may be inflated due to both measurement and treatment fidelity.

### 4.4.3 Treatment Content

The treatment content, math achievement, may limit the generalizability of the point estimates and distributions obtained (Shadish & Ragsdale, 1996). The current study represents only one attempt to estimate selection bias effects, and those effects may be limited to math education or cognitive skills more broadly. Additional information regarding the magnitude of selection bias effects is warranted. Future studies may estimate selection bias for different content areas (e.g. criminal intervention programs), which could provide important

information regarding general or content-specific selection bias variability.

### 4.4.4 Non-Generalizable Posterior Distributions

Establishing probabilities for selection bias magnitudes (Aim 2) was a primary goal of the present study, however, these probability density functions could not be determined because the resampled distribution did not conform to a known distribution. This study's reliance on extreme groups (i.e., high vs low ability) may have created the distorted distribution. While it is possible that selection bias distributions may not fit any known distribution properties, future research should not be dependent solely on resampling procedures but obtain data from an entire range of a selection variables. Furthermore, more sophisticated resampling procedures including Markov Chain Monte Carlo Methods and even simple resampling procedures such as Jackknifing might create more readily understandable distributions.

### 4.4.5 Potential Range Restriction in the Direct Mechanism

The maximum ES point estimate may not truly be the maximum possible given the low ability participants had a restricted lower limit. Participant sampling, therefore, might affect both the maximum and minimum estimates. In the future a sample more representative from the general population might produce both a better, more well-behaved distribution as well as values that generalize to other samples and populations.

Due to the university student sample, the low ability group may not be very low in actual math ability - according to the population. The low ability group was only a standard deviation away from the high ability group with the means of the two groups separated by about two standard deviations. Selection difference should still be and are evident but the range restriction may limit the generalizability to real non-student samples.

#### Design Considerations

The current estimates only pertain to the between-subjects effects and may be entirely different with more sophisticated designs. Sometimes, researchers use split-plot or mixed

designs whereby effects between and within subjects get estimated. Thus, future researchers ought to consider designs with repeated measures, multiple groups, or even combinations of the two that might address the adequacy of these estimates for other designs.

## 4.5 Implications

The present study has three major implications for future research: 1) Selection bias magnitudes may be large enough to account for any observed treatment effect, 2) different selection variables create differing bias magnitudes, and 3) Researchers now have experimentally derived a-priori selection bias magnitudes for the E&S procedure.

First, previous selection bias magnitude findings range from very small (d = .05) to large (d = 1.38) (Donley & Ashcraft, 1992; Lipsey & Wilson, 1993; Shadish et al., 2008; Pohl et al., 2009). Those previous effects fall at the low end of the ability posttest magnitudes or high end of the gender posttest magnitudes results (see Table 3.4). Taken together it is evident that selection bias magnitudes can be quite large, possibly much larger than treatment effects and thus provide additional evidence to the concerns about causal inference in poorly designed and implemented NR studies.

Second, placing the previous findings in context with this study indicates that categorizing selection bias in the traditional terms of self, or administrative, or geographic might not be as useful for estimating or predicting bias magnitudes, but in fact, as Cook et al. (2009) and Steiner et al. (2010) stated, and as I applied, the direct or indirect relationship between the selection variable to the treatment and outcome is a better classification method.

Third, adjusting NR psychology results may be better done by Estimate & Subtract (E&S) than other statistical procedures because of its implementation ease and the trivial burden on researcher resources. Still, a significant limitation to using E&S was the lack of an estimate of selection bias. This study provides a specific selection bias estimate for two types of selection variables, one relevant to the outcome and treatment and one irrelevant, at both pretest and posttest.

Estimating bias from a single experiment along with resampling statistical results offers preliminary findings worth expanding upon. Berelson and Steiner (1964) attempted to produce an inventory of scientific findings that other researchers might find suitable for future research. These findings - they argued - might serve as initial estimates; presumably their utility would be suitable for the purposes of judging current research efforts in light of earlier findings. Unfortunately, their efforts never gained traction with other researchers unlike more general compendiums and inventories including the Guinness Book of World Records (Records, 2013). The present study's findings ought to serve the research community as an initial estimate of selection bias magnitudes - not as a definitive value or static distribution. Further, these results should be considered to be an initial foray into this line of work to be improved upon by further research rather than as a final destination.

Given researchers focus on selection bias as a primary threat to validity, the present study represents an important initial step in providing estimates that can be used to rectify the influence of selection bias. Still, in light of the myriad other threats to validity (i.e. maturation, instrumentality, etc.), future studies may expand the investigation of bias effects for other relevant threats to validity to aid researchers in accounting for conditions that impede the integrity of scientific results.

Researchers and reviewers often assume that selection bias invalidates NR studies. Using these results, researchers may address the impact of selection bias while maintaining NR studies as viable and useful methodological designs. The present findings provide the beginnings for a tool to help address selection bias in a time when empirically-based decisions dominate modern research and policy. We have evidence-based practice (e.g., Cognitive behavioral therapy), empirically supported social policy (e.g., No Child Left Behind), and even scientifically supported social programs (e.g., Race to the Top). Without properly addressing the effects of bias in scientific research, the mechanisms for providing empirical support do not exist to make good decisions. Using estimates obtained in the present study, social scientists may begin to better design and evaluate their and others work, thereby producing more tenable and reliable findings.

# Chapter A: Literature Review

## A.1 Appendix: Literature Review

The following literature review documents relevant studies and summaries pertaining to selection bias. Included in this review are 1) a specific example of selection bias, 2) a clear definition of the term, 3) a justification for my focus on this bias instead of other potential biases, and 4) a review of the methods used to treat selection bias effects. Finally, I end my review with a justification for my dissertation study. I address each of these points below in clearly demarcated sections. The sections follow the aforementioned outline - beginning with the specific example so I could refer to it throughout the review.

## A.2 An example of selection bias

The Minneapolis Spouse Abuse Experiment (Sherman & Berk, 1984) serves as a great example of selection bias. Researchers designed a field experiment to test two competing domestic violence prevention/abatement strategies. They recruited police officers assigned them to implement one of three interventions for domestic violence incidents - either arrest the perpetrator, or mediate between the couple, or temporary separate the couple during the incident call. The researchers determined that the random assignment did not take place in all domestic violence cases. Police officers implemented their preferred intervention strategy, forgot their assigned strategy, or used professional judgement to deviate from the assigned group. Failure to properly administer random assignment left the researchers to resort to post hoc statistical adjustment to their studies to enhance causal inference (Berk, Smyth, & Sherman, 1988; Berk & Sherman, 1988; Sherman & Berk, 1984). Despite their attempts to remediate the problem, no firm conclusions could be drawn from the study -

largely due to selection bias threats.

## A.3 A Clear Definition of Selection Bias

Selection bias stems from initial differences in comparison groups - a situation that comes as a result from either the failure to randomly assign experimental units or the breakdown of the random assignment process. That definition is predominate in psychology it comes from two of its most influential methodologists in Donald Campbell and Julian Stanley. However, definitions for selection bias may differ by researcher and by field (Cronbach, 1982; Rosenbaum, 2002, 2010). Definitional differences include both the active assignment of experimental units and the sampling of experimental units in observational designs - both also differ with respect to the implications selection bias' impact. For the purposes of this review and for the supporting empirical study, I adopt Campbell and Stanley's definition - not for convenience but rather because that is the predominant perspective in areas relevant to my focus.

Campbell (1957, p.5) first described selection bias as "biases resulting in differential selection of respondents for the comparison groups". Their selection bias definition pertained to the main "treatment" effect inherent in non-equivalent groups at pretest and attributable to the systematic difference between groups. Those pretest differences, they argue, affect the posttest scores. For example, if either the responding police officer or reporting person from the Minneapolis Spouse Abuse Experiment choose the treatment, the selection mechanism would be unknown and the individual reasons and motivation for selecting arrest or mediation or separation may uncontrollable effect the results. That bias, while present at pretest, becomes relevant at posttest largely due to pretest differences.

Selection bias may be problematic when researchers rely on a counterfactual for causal inference. Researchers construct comparison groups to establish the counterfactual and failure to establish a reasonable counterfactual limits causal inference. Consider the spouse abuse experiment summarized above. The study aimed to compare three treatments - arrests, mediation, and separation - because the best or even better practical policy action

was unclear. A conceptual view of the counterfactual is "what would have happened" if a treatment had not occurred? More formally the counterfactual is observing experimental unit that went through a treatment as opposed to a similar experimental unit that did not go through a treatment on a dependent variable. Historically, the counterfactual has long been a part of science - even its statistical formulation is a recent addition that does not change the theoretical importance or position in science (Holland, 1986; Rubin, 1974; Winship & Morgan, 1999). Continuing and furthering the Minneapolis Spouse Abuse Experiment example, researchers were interested in the counterfactual where mediation was compared to usual treatment (arresting or separating the couple).

Experimental designs help us infer cause and the counterfactual merely strengthens that causal inference - point I address further below. Strong causal inferences are dependent upon ruling out or eliminating alternative rival hypotheses. Since the counterfactual (aka comparison groups) serves as an alternative to the focal treatment, results from these studies are said to maintain strong causal inference. Using the counterfactual for testing arrest and separation against mediation strengthens the causal inferences if and only if the counterfactual is effectively constructed.

The counterfactual is meant to offer effects comparisons for equivalent groups. That is, the comparison via the counterfactual only holds when the groups assigned to different treatments are equivalent. That group equivalence only holds for between-subject designs but counterfactual comparisons are equally relevant in within-subject designs. The focus of this paper is on the former while the later demonstrates there are multiple methods for creating a counterfactual and maintaining high internal validity. Thus, if no experimental unit can experience two or more treatments at the same time, the counterfactual requires these units to be randomly assigned and equivalent - on all relevant variables - prior to treatment. Initial group equivalence, therefore, is a necessary aspect of all counterfactual comparisons and, in turn, responsible in large part for strong causal inference in experimental designs.

Researchers in social science establish counterfactual comparisons via random assignment. That is, they create equivalent groups by making them probabilistically similar.

Contrast this similarity to a process whereby groups are non-equivalent - the essence of selection bias. Randomly assigning (RA) experimental units to comparison groups removes any pre-treatment differences that might account for the observed effects. Unknown or uncontrolled systematic assignment means units with certain characteristics - relevant to the treatment and/or outcome variable - may bias one group and, in turn, bias the observed effect. Thus, non-random assignment (NR) may adversely affect the initial group equivalence, reduce the effectiveness of the counterfactual, and weaken causal inference. RA can work as a methodological preventive action for not only selection bias but other threats to validity.

### A.3.1 Randomization

Randomization - via random assignment - is essential for preventing selection bias. Randomized controlled designs consist of randomly assigned participants (or experimental units) to two or more groups. These groups often consist of a focal treatment group and either a comparison treatment or no treatment. The basis for comparison remains the counterfactual whereby the investigators wish to infer changes attributable to the focal treatment that would not be observed by any other treatment or no treatment. Again, this description is nothing more than counterfactual reasoning as described above. In the case of the Minneapolis Spouse Abuse Experiment, the investigators compared three equally plausible alternative treatments and the randomization process was implemented to obtain group equivalence and to protect against selection bias. Since chance dictates group assignment, all relevant individual differences that may enhance, inhibit, or interact with the treatment are likely to get balanced between groups (Fisher, 1935). Group equivalence at assignment allows experimenters to make assumptions about the unbiased results. Thus, random assignment produces group equivalence and counters the potential for selection bias. Random assignment, however, is a process so common to social science and yet often misunderstood. Thus, I will explain the rudimentary details of the process as well as the expected outcomes further.

### A.3.2 Randomization - The process

Random assignment is a methodological process. Assignment to comparison groups is done by a random process - uncontrolled by experimenter, participants or context but by chance alone (Fisher, 1935; Shadish et al., 2002). In its most basic application, random assignment can be done by flipping a coin assigning an experimental unit to one of two comparison groups is dependent upon only on which side of the coin lands up. In a simple between groups with two comparison groups, if there were enough experimental units, any unique characteristics of the experimental unit will be approximately equivalent across groups. Differences that do exist are due simply to random error and not a predictable process (Shadish et al., 2002). That random process of assigning participants to one of two groups may be best characterized by the simple equation:

$$P(\text{Assignment} \mid \text{Experimenter, Participant, Context}) = 0.5 \tag{A.1}$$

Any deviation from that probabilistic statement, prior to the study, leads to the inference that either experimenter, participant, or context influenced assignment and may bias the observed effects. Those deviations then lead to an ability to predict assignment, a violation of the randomization concept.

### A.3.3 Randomization - The outcome

The process alone, however, is not what researchers aim to produce. Instead, they aim to produce equivalent groups - the outcome that comes as a result of the randomization process. Unfortunately, as can be seen with the Minneapolis Spouse Abuse Experiment conducting random assignment does not mean it will be effective at establishing equivalence between comparison groups. Failure of random assignment can naturally occur at initial assignment

pre-treatment stage but if the researchers or assistants forget their assignment sheets than the assignment process failed due to real world pragmatic reasons and assignment was done by something other than random (Berk et al., 1988). Failures of random assignment leave alternative systematic reasons for assignment and thus create rival hypotheses for the causal explanations other than the one(s) being explicitly tested and this reduces internal validity and weakens any causal inferences leading to a need to post hoc statistical correction methods to increase internal validity and strengthen any causal inference.

If the process of randomization fails or is never implemented then internal validity and weak causal inferences arise. Thus, randomization equates experimental units by group on all characteristics - regardless of whether they are measured or even relevant to the causal inferences. The process of randomization does not **ensure** the initial group equivalence (i.e., the outcome of the process) however, implementing the process increases the likelihood of equivalence but often depends upon the sampling, sample size, and effect size in the study. I elaborate on each point below.

### A.3.4 Sampling or sample characteristics

Homogeneity of the original sample leads to a greater probability of initial group equivalence. Greater sample diversity means potentially a greater number of relevant variables that need to be balanced between groups. If there are only two experimental units (e.g., people) for two separate experimental conditions and those units or people are replicates of one another then generating equivalent groups will be simple since the process of randomization ensures equivalent groups in that case. That is, if the experimental units are equivalent by nature so it does not matter which experimental unit gets assigned to either condition. However, as the units increase in variability - or differentiate from each other with respect to their inherent characteristics - the probability decreases that randomization as a process will lead to initial group equivalence. The probability for group equivalence with highly heterogeneous groups is dependent upon an appropriately implemented randomization process and sufficient sample size.

### A.3.5 Sample Size

As I mentioned previously, sample size ought to be considered with respect to initial group equivalence. Without sufficient sample size or randomization, non-equivalence or selection bias threats increase. If the Minneapolis Spouse Abuse Experiment only included a total sample size of 9 (i.e., 3 per group), then even proper implementation of a randomization process would lead to likely initial non-equivalence - attributable to the violation of the law of large numbers which according to probability theory requires a large number of observations to obtain a probable and stable value. If however, the total sample size was increased to 90 with 30 per group then the probability of obtaining group equivalence through random assignment increases. The true problem is that researchers do not know how large a sample size must be in order to counter these selection bias threats. Generally, the law of large numbers provides some guidance but how large is large enough remains a long-standing question in social science.

### A.3.6 Effect size

Sample size influence might remain unclear, but the influence of effect size is even less clear. We simply do not know how large of an observed effect will be safe from the threat of selection bias. If the randomization process failed - perhaps due to sampling variability, inadequate sample size, or failure to adequately implement the process - and we are faced with initial non-equivalent groups, the results may not be completely biased. Consider a study that produced an observed effect of 3 standard deviation units (i.e., Cohen's d=3.0) between the treatment and comparison groups. If there were initial group differences, could this large effect size be explained completely by selection bias or a selection by treatment interaction? The answer to that question is simple: we do not know. I suspect that most people might consider that effect so large that selection bias alone might not be sufficient to account for its magnitude. Similarly, small observed effects may be the result of a large true treatment effect offset by a negative selection bias effect. The possibility for these effects to add to or subtract from the true treatment effect exist. Directional effects of these threats

also remains elusive but most researchers can ascertain the direction simply by making some inferences about the differences between the groups and how those differences may directly affect the outcome measures or interact with the treatment. The following equation might help elucidate this point:

$$\mathrm{ES}_O = ES_{TX} + ES_{SB} + ES_{TXxSB} + ES_{OtherBiases} + ES_{error} \quad \text{(A.2)}$$

Equation A.3.6 makes explicit the proposition I introduced above. An observed effect ($ES_O$) comes from a simple additive model where the true treatment effect ($ES_{TX}$) combine with other potential effects - one of which is selection bias ($ES_{SB}$) and the other is error ($ES_{error}$). A third effect comes from the interaction between the treatment and selection bias ($ES_{TXxSB}$). The focus of my point here is to determine whether $ES_O$ can be sufficiently large enough to rule out $ES_{TX}$ and $ES_{TXxSB}$, thereby attributing the entire effect to $ES_{SB}$. As I mentioned previously, there are some instances where selection bias might counteract the treatment effects and produce an underestimate of the observed effect size. I suspect, however, that researchers know enough about their treatment population and treatment characteristics to gauge the direction of these effects. Nevertheless, A.3.6 represents the logic of this point effect sizes remain relevant to the discussion of selection bias and any threat to validity.

Some studies with extremely large effect sizes may overcome initial non-equivalent groups simply by producing an effect that could never be generated by those initial differences. Hence, treatment effects may trump selection bias effects. To what extent can this happen? Researchers (and I) do not know. Selection bias effect sizes remain unclear at best. Meta-analytic studies comparing random assignment to non-random assignment show variable effect size differences between those designs (Lipsey & Wilson, 1993; Shadish

& Ragsdale, 1996; Shapiro & Shapiro, 1983; M. Smith, Glass, & Miller, 1980). Effect sizes for selection bias effects vary widely across social science disciplines (range: d = .05 to d = 1.38) (Donley & Ashcraft, 1992; Lipsey & Wilson, 1993). That range might be a good start, however, there are no studies that explicitly and deliberately produced effects to see how large they may be under any circumstances.

All observed effects include some error ($ES_{error}$). Selection bias effects theoretically exacerbate the inaccuracy of the observed effects reflecting the true treatment effects. Absent knowledge of selection bias effects, researchers are left to infer any deviation from the true treatment effect may be attributable to random measurement error, treatment fidelity problems, or other problems that directly impact $ES_{error}$. Despite the absence of specific direct data about selection bias effect sizes, there are data regarding treatment effects and how treatment effects are impacted through designs that likely have selection bias effects. Effect size estimates for selection bias vary widely across social science disciplines (range: d = .05 to d = 1.38) (Donley & Ashcraft, 1992; Lipsey & Wilson, 1993). That is, these effects that are directly and singularly attributable to selection bias alone remain unclear due to the variability in their estimates. These effects may be large enough to counter selection bias effects but they may be trivial. If we knew the magnitude or range of values for selection bias effects (i.e., $ES_{SB}$), then we might be able to appreciate the limits selection bias influence. Presumably, selection bias may threaten studies with small effects more so than studies with large or extremely large effects. The basis for my empirical study is to estimate these selection bias effects. Before providing details of my study, I justify why I chose selection bias and then review the extant literature to show what we already know about selection bias.

## A.4 Justifying a focus on selection bias

Selection bias is one of many validity threats but tends to get priority over other threats - both in the literature and in practice. More theoretical and empirical attention is paid

to selection bias because of its perceived effects and probable presence in many studies. Selection bias impacts the outcome when initial differences between groups either directly relate to differences in outcome measures or when initial differences interact with treatment effects to produce differences in the outcomes (i.e., a moderated effect). Thus, when selection bias is present, any posttest effect will be subject to not only the influence due to treatment but also the initial differences between the groups. Posttest effects influenced by initial differences are biased effects. We researchers worry that these biased results either 1) underestimate **our** true treatment effects or 2) overestimate **other researchers** true treatment effects ("shifting standards of evidence depending upon whose ox is being gored.") (Meehl, 1973, p.231). In short, selection bias creates posttest results that are a function of the treatment effects, initial differences between groups, and the interaction between the two.

Biased posttest results come from many other factors as well - not just selection bias. Selection bias potentially interacts with numerous other threats ($ES_{OtherBiases}$), further reducing the clarity of the true treatment effects (Campbell & Stanley, 1963; Shadish et al., 2002). The entire foundation of experimental research design focuses on eliminating plausible rival hypotheses and thus, on isolating a specific causal relationship and observing the true treatment effects. When selection bias is present either by limitation in design or failures in the randomization process, the study's effects are biased and any intentions of sound scientific practices are jeopardized.

As documented above, any design or methodological weakness that threatens causal inference is a threat to internal validity. Selection bias, as with any other threat to validity, is a possible alternative explanation for observed effects in a study. Any threat to validity undermines researchers attempt at understanding the true effect. That true effect may be for what constitutes a concept or construct, or the causal relationship between constructs, or even how construct causal relationships vary by settings or persons. There are many threats but I focus on one - selection bias. From this point forward, I will address selection bias as the primary threat to validity and use the term threat and bias interchangeably.

So why selection bias and not another potential threat to validity? Simple. Selection bias is the single threat that dominates all areas of causal inference. As I discussed above, the counterfactual is the central, logical tenet of causal inference. Anything that threatens the counterfactual affects the inferential gains (i.e., utility) of a research project. So the first primary reason for focusing on selection bias is that it relates to a primary and essential ingredient in research - the counterfactual. But that reason alone is not sufficient. Instead, I focused my attention on selection bias because researchers pay more attention to selection bias in the empirical literature than any other bias. Consider the two research programs outlined below and what they aim to contribute to science I elaborate on each for two reasons - to demonstrate a working knowledge and to delineate a rationale for my focus.

## A.5 Methods to Treat Selection Bias

In this section, I document the most widely used statistical and methodological approaches to improve causal inference in non-equivalent group designs. I provided both a historical basis as well as a methodological basis for grouping these approaches. By doing so, I hoped to clearly articulate why my study would contribute to the extant research methodology literature.

Let me begin by introducing the key people in the field - Don Rubin and Don Campbell. Both are (were) luminaries in the field of research methodology and most of their relevant and derivative work get the bulk of attention from the research community (Shadish, 2010). Consider a few metrics just to highlight the attention these researchers garnered over the past few decades. Rubin, for example, has over 15,000 citations to his work on selection bias alone. That work, as documented below, focused on propensity scores, instrumental variables, and causal inference from non- or quasi-experimental studies. Campbell, similarly, has over 20,000 citations to his work on quasi-experimentation, regression discontinuity, and patched-up designs. These works - and their derivations - lead to most, if not all the work I document here and pursue in this study. I begin with Rubin's work - not because it preceded Campbell's contributions but because it provides some clarity to my overall aims.

Following Rubin's contributions, I then detail Campbell's.

### A.5.1 Don Rubin

Don Rubin published a seminal paper on causal inference from non-randomized controlled designs in (1974). That paper lead to a series of papers on what is now referred to as "Rubin's Causal Model" or RCM. The RCM established the first formalized, statistical, and comprehensive approach to establishing causal inferences based upon counterfactual thinking. Other theorists and statisticians had similar ideas (Bacon, 2005; Hume, 1740; Neyman & Pearson, 1933) but none of them offered the same comprehensive approach that Rubin began with in his modeling approach. Rubin boldly held that design trumps analysis for causal inference as reflected in his statement "this [causal inference] enterprise requires careful thought and execution, and not simply running mindless regression programs and looking at coefficients" (2008, p.837). If design trumps analysis and the counterfactual remains central to causal inference then selection bias becomes one, if not **the** most important features of scientific inquiry. Rubin never directly states this position but it becomes apparent from his work on both design and analysis - the counterfactual remains central to all causal inferences. Based upon my prior review and discussion of counterfactual reasoning and the role of selection bias in undermining counterfactual reasoning, I contend that the RCM requires equivalent group designs and minimal selection bias for strong causal inference.

My contention aside, Rubin expended tremendous effort in trying to wring out reasonable inferences from quasi-experimental and non-experimental designs via statistical procedures. He developed a rationale for propensity scores - based upon the logic of instrumental variables (Wright, 1928) - that enabled observational researchers (e.g., economists, psychologists, sociologists, etc.) to draw stronger inferences from weaker non-equivalent group designs. Initially, Rubin (1977) showed that group assignment via a covariate could be modeled sufficiently to account for initial non-equivalence. His logic lead to a simple justification whereby researchers could retrospectively analyze quasi-experimental designs

via an aggregate variable - much like the use of instrumental variables that predated his work. Rosenbaum and Rubin (1983) provided the rationale and empirical justification for group matching based upon a simple logistic regression model they called propensity scores. A propensity score represents the probability that an individual would be assigned to a treatment group. When two individuals - study participants in this case - have an equal probability of being assigned to the treatment but were assigned to different groups, we can consider them to be "matched" and equivalent. Holding that logic across all participants creates a logical framework whereby a data analyst may construct an entire study - retrospectively - as if the individuals were randomly assigned.

Rubin's work on propensity scores lead to a host of research programs focused on the same issue - minimizing selection bias. These efforts were largely empirical tests that showed the importance of comparison group selection in quasi-experimental designs (Campbell, 1969; Rubin, 1974; Heinsman & Shadish, 1996; Steiner et al., 2010). That work showed that covariates can be useful in retrospectively constructing comparison groups because they enable researchers to create group equivalence. Prior to these efforts, there existed only one way to reduce selection bias - randomization. Now, raw significant difference between non-randomized and randomized effect sizes can be reduced by statistical adjustments (covariates in linear regression), careful selection of comparison groups, and eliminating self-selection (Shadish & Ragsdale, 1996).

The first two reduction strategies - statistical adjustment and careful selection - require researchers to design studies to protect against selection bias. Selecting comparison groups requires foresight selecting alternative comparisons at the conclusion of the study does not require foresight but allows other threats to validity to creep in and weaken the causal inference (e.g., maturation, history, etc.). The latter of the three strategies - eliminating self-selection - requires researchers to guard against this confound by carefully monitoring the treatment assignment and treatment fidelity. Monitoring both of these study aspects need to be carried out during the study and, as a result, require researchers to plan ahead. Thus, only one of the three approaches can be carried out retrospectively or in reaction to

the threat of selection bias. I devote the majority of this section from this point forward on retrospectively adjusting effects because these adjustments directly relate to my current study and come from Rubin's work.

Unlike random assignment, equivalence created post-hoc is not hypothetically probable but rather empirically probable based upon the data. In other words, randomization as a process may not deliver the outcome we desire but matching subjects via covariates ensures that between group comparisons maintain reasonable equivalence according to the counterfactual logic I outlined previously (see Figure A.3). Thus, retrospectively constructing comparable groups appears sufficient to deliver unbiased effects. Even approximate matching on covariates leads to acceptable and effective NR designs (Cook et al., 2009; Steiner et al., 2010). The question remains about the approximation of matching required to create a sufficiently strong counterfactual.

$$\begin{aligned} \mathrm{ES}_O &= ES_{TX} + ES_{SB} + ES_{CV} + ES_{error} \\ \mathrm{ES}_{CV} &= ES_{CV1} + ES_{error} \\ \mathrm{ES}_{CV} &= ES_{CV2} + ES_{error} \end{aligned} \tag{A.3}$$

Comparison groups are more effective - as counterfactuals - when individuals in both groups are equivalent across many potentially relevant domains (Cook et al., 2009; Dehejia & Wahba, 1999; Glazerman et al., 2003; Shadish & Ragsdale, 1996; Steiner et al., 2010). As individuals differ across the domains, causal inference suffers. Recent research by Shadish et al. (2008) and subsequent reanalysis by Steiner et al. (2010) suggests an even clearer message the greater the relationship between a covariate (or covariates), the greater the utility in adjusting quasi-experimental effects to reflect reasonable experimental effects (Cook et al., 2009). Re-analysis of meta-analytic data indicated that stable experimental effects may be estimated from quasi- or non-experimental studies via careful selection of

comparison groups (placebo vs. wait-list) thus minimizing pretest differences and avoiding self-selection (Heinsman & Shadish, 1996). These results showed that researchers improved comparison groups through matching attrition levels and sample characteristics. Together, Heinsman and Shadish determined that non-randomized designs with comparisons groups matched for important sample characteristics (i.e. any variable relevant to the treatment or outcome measure) produced reliable and valid effects (Heinsman & Shadish, 1996).

If statistical adjustment can "fix" weak studies, then why don't researchers simply use these methods instead of stringently adhering to randomized designs? Simply put, statistical adjustments require data that are not always available. Specifically, all covariates relevant to assignment must be available and included in the model (Heckman, 1979; Heckman et al., 1987; Rubin, 1974; Steiner et al., 2010). A clear example of these relevant variables comes from Heckman's (1987) re-analysis of Lalonde's (1986) study. Lalonde's contention was that non-randomized designs produce poor causal inferences (i.e., unreliable or invalid effects). Heckman countered that claim by reanalyzing the results and demonstrating that reasonable causal inferences could be gained from non-randomized designs provided relevant variables could be used to form logical matches between treated and untreated participants. Dehejia and Wahba (1999) confirmed Heckman's approach by producing similar effects as Heckman but instead of matching by variable, they matched cases by propensity scores. In both cases, reanalysis of the Lalonde data adjustments approximated the effects observed in randomized designs - albeit at times these estimates were slight underestimates but still reasonable according to the authors. None of these adjustments, however, were possible without the inclusion of relevant covariates. Omission of these variables eliminates the opportunity to adjust and few other options exist after a study has been carried out.

Not only do researchers need access to these relevant variables but also they need to have sufficiently large sample sizes to make reasonable adjustments. Sample size directly affects the reliability of adjustment (e.g., propensity score requirements (Wilde & Hollister, 2007)). Small sample sizes reduce the utility of most adjustment procedures. To expand upon the relevance of both the need for all relevant variables and sample size requirements,

I describe the most popular adjustment procedures - propensity scores, instrumental variables, Heckman selection model, and estimate and subtract - below and then summarize the strengths and weaknesses of each method.

### Propensity Scores

To begin, propensity scores - as Rubin originally envisioned - offer researchers a way to create a counterfactual when the counterfactual does not exist. Specifically, if a study failed to actually randomly assign participants or never implemented random assignment then the use of propensity scores may be a viable statistical adjustment tool to enhance internal validity and draw reasonable valid causal inferences.

Propensity scores are the combination of multiple demographic or otherwise descriptor variables that match or divide observed experimental units into "as-if" counterfactual comparison groups. The creation and application of propensity scores is as follows: 1) select a set of variables relevant to group assignment, 2) use those variables as predictors in (binomial or multinomial if more than two groups) logistic regression to predict group assignment, 3) save probability of assignment - as estimated from logistic regression - as propensity scores (i.e., the propensity to be assigned to a group), 4a) use propensity scores as covariates in statistical analysis or 4b) use propensity scores to create matched pairs from the observed sample and analyze these matched pairs as if they were randomly assigned. These four basic steps serve as the standard procedure for propensity score analysis.

The last two options require a bit more elaboration. Once the researcher computes the propensity scores, she may incorporate those scores directly into a multivariate analysis as covariates (4a above) or use the scores to match participants between groups (4b). As a covariate, the propensity score "adjusts" the partial correlation coefficients by taking into consideration the relevance of assignment. The adjustment does not correct but rather attenuates the potential effects by residualizing out any shared variance between assignment (i.e., selection bias) and the predictor variables. Using propensity scores as covariates improves causal inference, however, an alternative approach tends to be more widely used

- participant matching via propensity score.

As a matching procedure, propensity scores allow the researcher to identify participant pairs to match and analyze as if the pairs were randomized to different groups. The paired participants were indeed assigned or participated in different groups but the matching algorithm simply finds people who might be suitable comparisons by group because they have similar probabilities of assignment. A person assigned to a treatment group, for example, would be matched with a person assigned to the comparison group if both had similar propensities for being assigned to the treatment group. This matching process is similar to what I mentioned earlier about the relevance of constructing useful comparisons. Propensity scores, in short, act as a retrospective method for creating a sound design. After all the study participants are matched, the analyst conducts statistical analyses on these newly formed treatment and control groups "as-if" the participants were randomly assigned to the groups. Propensity score matching has a number of requirements and restrictions including some details I do not intend to address here (e.g., how to create and determine what constitutes a "match" and sample size requirements for complex logistic regression). Despite the limitations of matching, propensity scores used in this manner produce robust and dependable parameter estimates and hypothesis tests - similar to randomized controlled trials (Cook et al., 2008). A few more considerations are necessary to fully appreciate the utility of propensity scores.

First, propensity scores are themselves conditional probabilities P(Assignment | Covariates) = 0.5 for any observed participants' likelihood of being assigned to a treatment or control condition (Rubin, 1974). These conditional probabilities are the result of a logistic regression utilizing numerous observed covariates. In short, propensity scores are conditioned on covariates and ought to be considered estimates of assignment probability.

Second, the quality of those estimates relies heavily on the relationship between the covariates and the assignment mechanism (see subsection under selection bias above). Not all covariates used in these logistic regression models are equal in terms of deriving useful propensity scores. Rubin recommends using only covariates related to the assignment

mechanism and unrelated to the outcome variable. Often, quality propensity score computation requires numerous covariates - partly because the assignment mechanism is unknown. When the assignment mechanism is unknown, researchers might find it difficult to discern the utility of potential covariates to include in the logistic regression. Thus, researchers often resort to including many covariates - perhaps more than necessary - to estimate these propensity scores.

In summary, propensity scores are estimated probabilities from multiple pre-treatment observed covariates (Guo & Fraser, 2010). Those estimated probabilities can then be used as a covariate or as a variable to match participants to form better comparisons. If used as covariates, propensity scores serve as adjustments to the multivariate parameter estimates and, in turn, hypothesis tests. If matching via propensity scores renders sufficient comparisons, then the results ought to come close to the results obtained from randomized trials. The operative word in the previous sentence is "if" because there are limitations including data availability and sample size that affect this conditional statement. Propensity scores require an exhaustive array of relevant variables and large sample sizes to draw strong, valid causal inferences - regardless of the approach. Any deviations from those conditions limit the propensity score approach and fail to "fix" non-randomized designs (Guo & Fraser, 2010; Rubin, 2010).

### Instrumental Variable

An alternative to propensity scores are instrumental variables. As I mentioned previously, propensity scores allow researchers to select a large group of covariates for a logistic regression that selection process does not demand that the researcher know the assignment mechanism. In contrast, instrumental variables are similar to propensity scores - by application - but differ with respect to knowledge of the assignment mechanism. The researcher must have a good rationale for selecting an instrumental variable because that variable must be related to the assignment mechanism. In short, what separates instrumental variables from propensity scores is the knowledge of the assignment mechanism.

Instrumental variables are useful in non-randomized designs, particularly observational designs, because an instrumental variable creates a counterfactual within the observed data by differentiating the statistical model predictor variables - conceptually similar to propensity scores but procedurally different. Instrumental variables are created and applied through the following steps: 1) select a single (instrumental) variable that explains the dependent variable but only as an indirect effect through one or more predictor variables, 2) include the single variable in the statistical model along with the independent variables. A conceptual example is predicting the effect of military service on civilian health and earnings. Using the military draft lottery, as the instrumental variable, and using military service as a predictor of civilian health and earnings. While there is a relationship between being drafted and service there should not necessarily be a connection between a lottery variable and the civilian health outcome variable. Including the instrumental variable in the statistical model is distinct from the propensity score method conceptually as the instrumental variable is theoretically and empirically causing the other independent variables while not directly related to the dependent variable while propensity scores are not conceptually nor empirically causally related to the other predictors in their statistical models. Instrumental variables seem pragmatically less resource intensive than propensity scores but maintain more stringent theoretical and practical restrictions. Those restrictions arise from the how predictor variables are dealt with in a regression model.

An instrumental variable is essentially a predictor that is fully mediated by another predictor. Consider a simple model where the researcher identifies two predictors (A and B) and one outcome (C). If both predictors were related to the outcome then we would not have a suitable instrumental variable among A and B. If, however, A were relevant to C but only as a predictor of B (i.e., A and B are collinear) then A might serve as a good instrumental variable. The utility of instrumental variables is often depicted in the following way:

$$E(C) = E(B \mid A) + E(Error \mid A) \tag{A.4}$$

In standard form, E stands for the expected value of a variable and the pipe ("|") is a conditional statement. Thus, the expected value of C is dependent upon the expected value of B given A along with the expected value of the residual or error given A. The logic of instrumental variables lies in the first term (E(B|A)) where the effect of B on the outcome C is dependent upon A. Some may argue that this effect is no different than a moderation and, in some cases, that makes sense, however, if the effect were a moderation and not a mediation then the residual term would be contingent upon A and, as a result, reduce the utility of our instrumental variable. A moderation effect would look like this

$$E(C) = E(A) + E(B) + E(Error \mid AandB) \tag{A.5}$$

Econometricians view the problem from the equation A.5 I showed previously. The logic seems similar to psychologists statistical definition of moderation but there are important distinctions. First, instrumental variables are not direct effects (i.e., E(A) is zero by that definition). Second, the effect of B (E(B)) can only be reasonably estimated by considering the contingencies of A (E(B|A)). Finally, the error term (E(Error)) must be independent of either A or B - a point I shall return to shortly.

The general use and formulation of instrumental variables tends to be more consistent as a mediation process rather than a moderation process. The causal model below represents the logic of instrumental variables:

$$A \rightarrow B \rightarrow C \tag{A.6}$$

Another important aspect of instrumental variables lies in the second term in the equation A.6 above (E(Error |A)). Ideally, we expect errors to be independent and not contingent upon A - our instrumental variable. When the error term relates to A or any other predictor, we have a problem that econometricians call endogeneity (Heckman, 1979). Specifically, endogeneity results from the correlation between the predictor variable and the error term. Non-randomized designs - especially observational designs - are probabilistically prone to

selection bias that, in turn, probabilistically creates a relationship between predictor variables and the dependent variable's error term. There are other causes of endogeneity (e.g. omission of relevant variables, simultaneity, etc.) but here I focus on the impacts of selection bias.

In order to resolve the endogeneity problem, a data analyst uses a statistical solution whereby an exogenous variable gets included in the regression equation. This exogenous variable must be correlated with the predictor variable (B) and the dependent variable (C) while not being correlated to the error term. The exogenous variable (A), when it satisfies these conditions, serves as an instrumental variable.

Instrumental variables can be used as covariates to "adjust" the parameter estimates of the primary predictor in the same way that any collinear predictor adjusts the parameters of other predictors - via partial regression coefficients. As multicollinearity increases among predictors, partial regression coefficients decrease. Not only do the regression parameters change but so do the standard errors via the variance inflation factor (VIF). Thus, instrumental variables affect not only the parameter estimates for the primary predictors but also the hypothesis tests.

Despite the potential utility of instrumental variables and their relative advantages compared to propensity scores, there are several noteworthy limitations. First, instrumental variables are difficult to find. Identifying them from existing data sources - assuming the instrumental variable process takes place after data collection - can be a challenge. Second, even if a potential instrumental variable were identified, the researcher must be able to reason that the variable is related to the assignment mechanism. Assignment mechanisms are not always readily identifiable let alone knowable either a priori or post hoc. Third, the utility of instrumental variables tends to be related to sample size - the same limitation stated for propensity scores. Small sample sizes produce adjustments via instrumental variables that do not accurately characterize the selection bias effect. Fourth, instrumental variables place a large burden on the variable selection process and, as a result, potentially fail to capture all potentially relevant exogenous variables useful in adjusting the observed

effect. This reliance on a single variable stands as the primary weakness of instrumental variables when compared to propensity scores. Finally, instrumental variables tend to be limited to only situations where the treatment groups are homogeneous (Guo & Fraser, 2010; Winship & Morgan, 1999). When treatment groups have differential fidelity, the adjustment via instrumental variables assumes a fixed effect and may not adequately capture the relative within-group variability. In sum, instrumental variables are statistically and procedurally simpler than propensity scores but their theoretical and empirical requirements often make their application far more difficult.

### Heckman Selection Model

Similar to the two previous methods of addressing selection bias, economist James Heckman devised an alternative statistical adjustment tool that allows researchers to make causal statements using data from non-randomized designs. Heckman's tool is generally referred to as the sample selection model and it can be effective at reducing bias when data is collected through observational and quasi-experimental methods. The sample selection model tends to be a fairly complex, nuanced statistical model that can be summarized simply by comparing the model to propensity scores and instrumental variables. Rather than provide exhaustive details of these complex statistical nuances, I focus most of my attention on the advantages Heckmans model offers in comparison to the other models.

As a simplification, Heckman's sample selection model is a two-step procedure. The first step focuses on modeling the selection mechanism the second step focuses on applying the results from step 1 to a standard prediction model where the parameters are adjusted to accommodate selection (Heckman, 1979). In step 1, the analyst models selection by a least squares equation that results in assignment probability estimates. These probability estimates can be viewed in the same way as propensity scores, however, Heckman does not explicitly state the comparison. The logic between Heckman's method and propensity scores - at least according to my reading and interpretation - appear remarkably similar. Once modeled, the analyst uses that probability of assignment in a second linear model

to predict the outcome variable. Again, this application appears similar to propensity score and instrumental variable applications. Where Heckman's model differs from either propensity scores or instrumental variables is how the selection equation estimates the selection bias. Heckman models the bias as a hazard function or Mills ratio. Use of this hazard function allows for appropriately analyzing a dependent variable that has a truncated normal distribution resulting from the non-randomized design. Including the modelled selection conditional probability produces outcome variables estimates that are reliably comparable to a randomized study design result (Guo & Fraser, 2010).

Heckman's method offers some advantages to the two previously discussed methods in that the method accounts for distributional anomalies that may be inherent in non-randomized designs. The method, however, does have some noteworthy disadvantages. First, Heckman's method requires the inclusion of all possible, relevant covariates in step 1. Failure to include relevant covariates greatly affects bias reduction in Heckman's model (Guo & Fraser, 2010). Second, Heckmans model requires the same large sample sizes as required by propensity score and instrumental variables. The requirements for large sample sizes makes this method less useful for most social scientists - at least those outside the areas where studies rely on large (N > 1,000) sample sizes (Winship & Morgan, 1999). Third, the residuals from both steps must be normally distributed - just as in all ordinary least squares models. Recent derivations of Heckman's model do not solely rely on OLS, however, the original and most widely used application tends to be restricted to OLS procedures. Propensity scores and instrumental variables are not limited to OLS applications. Finally, the residuals from both steps must be independent (hence, residuals as described from the previous limitation and this limitation require *iid* - independent and identically (Gaussian normal) distributions. Again, *iid* is a limitation for most OLS procedures and may be overcome by more modern statistical algorithms.

### A.5.2 Donald T. Campbell

While Rubin's work focused on statistical adjustments for NR and observational designs to approximate RA design effects, Donald T. Campbell worked just as effectively on methodological adaptations to failed RA designs or methodological alternatives to RA designs. Campbell devoted his efforts to either strengthening the counterfactual in NR designs through Regression Discontinuity, numerous comparison group designs, or patch-up designs. Yet, Campbell's most significant contribution may be from categorizing and describing the myriad of threats to validity present for numerous designs (Shadish, 2010).

Campbell, in collaboration with several colleagues (e.g., Julian Stanley, Tom Cook, and Will Shadish), constructed the foundation for social science methodology. Specifically, they acknowledged the limitations of all research designs from RA to NR to observational - leaving none immune to scrutiny and made a concerted effort to pinpoint specific limitations to causal inferences (Campbell & Stanley, 1963). Identifying specific threats to various designs was part of Campbell's approach to explicitly and directly dealing with potentially impaired causal inferences.

The focus on threats to validity soon established selection bias as the primary concern for most social and behavioral scientists. I say "primary" because it attracted the most attention from the long list of validity threats. As support for this contention, I searched the literature for other threats to validity and found few others that gained any attention - not just a modicum but any attention at all. Measurement bias (e.g., reactivity, testing effects, etc.) garnered the second most citations and, ironically, most of those citations came from Don Campbell's direct contribution of unobtrusive measures. Other work that may focus on alternative threats to validity are difficult to find - perhaps because the research is nested within a content domain or because the work is less prevalent than the work focusing on selection bias. In either case, my choice of selection bias stemmed mainly from its prominent attention in the literature. A second reason for focusing on selection bias stems from its potential influence on our inferences from empirical studies. I document that reasoning by the following research program.

Campbell's perspective might be best characterized by the following statement: failure to randomize and prevent selection bias leaves researchers with no alternative than to treat the problem. Even randomization does not ensure initial group equivalence so investigators may need to treat selection bias regardless of the design "...statistical procedures should be used ... [for] any faulty random assignment" (Berk et al., 1988, p.62). Thus, while randomization is necessary for methodological initial group equivalence, it might not always be sufficient. We must treat any situation where selection bias remains probable. Another reason for treatment options stems from the fact that random assignment is not always possible (Shadish et al., 2002). Some studies - non-randomized or NR studies - often begin with non-equivalent groups by design due to feasibility issues. These NR studies are common in many social scientific domains where researchers focus on uncontrollable outcomes (e.g., natural disasters, criminal activity, developmental processes, etc.). The methods I detail below tend to be more consistent with Campbell's notions of causal inference.

Campbell's approach inspired many procedures one in particular was the regression discontinuity design that clearly demonstrated Campbell's contribution to this problem. Another approach more germane to my dissertation was the Estimate & Subtract (Reichardt & Gollob, 1989) technique. I address each of these below.

### Regression Discontinuity

Perhaps the single contribution for estimating causal effects from NR designs consistent with Don Campbell's work comes from the concept of regression discontinuity. This design does not rely on random assignment but maintains high internal validity through a unique assignment mechanism and an *a priori* selected counterfactual. Similar to the above statistical procedures, regression discontinuity only works if the assignment mechanism is known. That mechanism must be taken into consideration by design. Instead of random assignment, the researcher controls group assignment via a pre-specified variable. Basically, regression discontinuity is a pretest-posttest design (see A.5.2) but the comparison or counterfactual is based solely on a cut-score (C) for the assignment method.

C X O
C O

Figure A.1: Regression Discontinuity Design

The cut-score is a bisecting point on a continuous measure; all participants (or experimental units) on one side of the cut-score are assigned to the same experimental group while the remaining participants on the other side of the cut-score are assigned to the other experimental group. For example, if the Minneapolis Spouse Abuse Experiment implemented a regression discontinuity design and a cut-score measure for only imprisonment and separation, then all persons below a threshold on some selected measure would be assigned to imprisonment while all those above the threshold would be in the separation comparison group. The requirements of the cut-score measure are that it be: 1) continuous, and 2) variability on either side of the cut-score (Shadish et al., 2002). Shadish also recommended that the cut-score be the mean from the measure. For causal inferences to be effectively drawn from this design, researchers need to strictly adhere to the assignment based upon the cut-score. After assignment and treatment (X) is complete the participants are measured on the outcome of interest (O).

After participants are assigned to their comparison groups they are observed at posttest (O). There are only two observation points in the basic regression discontinuity design. The first is used to assign participants to a comparison group and the second is to measure the treatment outcome. And upon completion of data collection the analytic process is conducted. As stated through the designs title, determining the effectiveness of the treatment is done using regression. Regression lines are computed for both comparison groups. Any treatment effect is calculated from the difference between regression lines. When plotted, with the assignment measure (C) on a horizontal axis and the outcome variable (O) on a vertical axis, the treatment effect is visually confirmed by the split between regression lines, this split is the discontinuity referenced in the methods title.

Regression discontinuity is a relatively simple design but difficult to implement because it requires assigning participants on a predetermined basis that is often not feasible for

the same reasons that random assignment may not be possible or feasible. However, the advantage of regression discontinuity is that treatments can be targeted to those in need while still maintaining high causal inferences and avoiding or reducing most threats to validity (Campbell & Stanley, 1963; Shadish et al., 2002).

**Estimate & Subtract**

Estimate and Subtract (E&S; Reichardt & Gollob, 1989) is one alternative to those aforementioned complicated and resource intensive statistical matching procedures. E&S aims to improve causal inferences from NR studies by subtracting the estimated bias from the observed effect. For example, if selection bias were known to produce a modest effect (e.g, d=0.3) for a particular population and an NR study published a large effect (e.g., d=0.9), then the E&S method would estimate the "true" effect of the treatment to be a more moderate effect (i.e., d=0.6 = 0.9 - 0.3). Effect size bias estimates may come from any study, rarely, however, do we have them readily available. E&S seems simple and useful to counter selection bias. Unfortunately, this procedure has not enjoyed the overall popularity as the more complicated statistical matching procedures mentioned above. One of the potential reasons for E&S's lack of popularity might stem from lack of readily available bias estimates. Furthermore, all published E&S reports use a specialized methodological design (e.g., interrupted time series) (Reichardt & Gollob, 1989; Reichardt, 2000, 2006, 2011). Current E&S implementations limit researchers to a few types of designs - many are not applicable to a typical two-group comparison study. Readily available bias estimates, however, may increase the viability of the E&S procedure to psychological research. In fact, unlike the statistical matching procedures, E&S implementations with using robust bias estimates for self-selection would be well suited for a majority of psychological NR studies due to its lack of resource demands and computational ease. E&S using pre-existing information treats biased results by subtracting a biased estimate from the observed results. If, for example a NR study had an observed Cohen's d of 1.5, the analyst would simply subtract the expected selection bias from available sources (e.g., .24 a value based on Shadish's (2008) study) from

the observed 1.5 to obtain the adjusted or unbiased 1.26 Cohen's d. As potentially useful as E&S may be for adjusting psychology NR study results its use is restricted by the absence of selection bias estimates.

All NR studies are generally considered flawed when compared to RA studies. This consideration has been observed in various studies comparing results from NR and RA designs. Because NR studies do not replicate exact directional and magnitudes of RA study results several adjustment tools have been developed. These adjustment tools usually adjust for selection bias. Their effectiveness is dependent on numerous factors but can be effective given the right circumstances and data. Despite theoretical and empirical knowledge on causality, threats to validity, and adjustment tools theres is no solid understanding or any empirical evidence for selection bias magnitudes themselves. Theoretical selection bias differences are modeled via simulation and comparison group differences were observed in separate instances but no independent estimates of selection bias magnitudes currently exist. Knowing what the bias magnitude is likely to be can be quite useful to researchers when deciding if it is necessary to conduct an RA study or fatally flawed to implement a NR study without an adjustment tool or a real requirement for an adjustment to that NR study result.

### A.5.3 A brief history of estimating selection bias

A reasonable justification for the current study may come from its historical roots - one that I provide below. Selection bias, as mentioned previously, became a relevant concern in the early to mid 1960's with the publication of Donald T. Campbell and Julian Stanleys monograph on Experimental and Quasi-Experimental Designs for Research (1963). There were other events that preceded that publication - ones that I document in Table A.2 - but contemporary social scientists largely attribute selection bias to Campbell and Stanley. Since their publication, social scientists slowly progressed toward estimating selection bias effects some efforts were subtle while others were explicit. I present these events in tabular form to allow the reader to appreciate the brief but productive history of selection bias. The

Table A.1: Table of selection bias treatment procedures

| Procedure Name | When Implemented | Strengths | Limitations |
|---|---|---|---|
| Propensity Scores | Post-hoc | 1. Effective treatment in bias reduction<br>2. Ease in implementation<br>3. Flexibly applied as either as a covariate or via a matching procedure | 1. Large sample size requirements<br>2. Large array of covariates required to create strong propensity score bias adjustment |
| Instrumental Variables | Post-hoc | 1. Solution to endogeneity problem | 1. Large sample size requirements<br>2. Difficult to implement |
| Heckman | Post-hoc | 1. Effective within constraints - large sample size, assumption of normality, and appropriate variables selected. | 1. Large sample size requirements<br>2. Susceptible to non-normality |
| Regression Discontinuity | a priori | 1. Strong internal validity - consistent with RCT<br>2. Ideal for targeting treatment to those in need | 1. Difficult to implement<br>2. Similar restrictions as RCT |
| Estimate and Subtract | Post-hoc | 1. Ease of implementation<br>2. No sample size requirements<br>3. Directly interpretable by and applicable for non-statistically oriented research consumers | 1. No a priori estimates currently available<br>2. unknown effectiveness |

events portrayed in the following table do not represent a census of all events but rather a comprehensive list that I use to establish the research programs associated with my research questions.

Table A.2: Table of keystone selection bias related contributions

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 1940 | Pratt and Rhine published a book containing the first aggregation of independent study results. | First publication of similar studies - conceptually identically but methodologically distinct. | Psychologys first compilation and review of similar constructs broke ground for future psychological meta-analyses. |
| 1963 | Campbell and Stanley published their Monograph. | Organized threats to validity for all social and behavioral scientists. | Starting point for most modern reviews and the historical starting point where the threat label for non-equivalent group designs gets assigned "selection bias." |
| 1960's | Social programs became more scrutinized by US Congress. | Congress' demands made quasi-experimental designs more popular and than experimental designs. | Created demand for research and, in particular for NR designs due to convenience. These demands also provided the opportunity to question the validity of inferences from these weaker designs. |

Table A.2 – *continued from previous page*

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 1974 | Rubin published his causal model. | Established a statistical framework for examining cause. | Merged multiple fields and encouraged causal inference from both design and statistical perspectives. |
| 1977 | Smith and Glass published the first applied meta-analysis in an area hotly debated by multiple stakeholder groups. | First application of meta-analysis where social scientists and the public cared about the outcome. | Inspired and motivated researchers within clinical psychology and beyond to examine meta-analytic procedures in greater detail. |
| 1981 | Glass, McGraw, and Smith publisher their meta-analysis book. | This is the first instructive and complete text on meta-analysis | |
| 1982 | Hunter and Schmidt published their book on meta-analysis. | The publication expanded meta-analysis and made it available for a broader audience. | Fully articulated meta-analysis for social scientists - particularly in industrial and organizational psychology. |

Table A.2 – *continued from previous page*

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 1983 | Shapiro and Shapiro re-examined Smith and Glass' meta-analysis and included more studies. | Incorporated study design into their meta-analysis and found that it had minimal impact. | Drew the attention of methodologists and other stakeholders who remained steadfast in their conviction that research design affected the quality and magnitude of the research findings. |
| 1986 | Major evaluation compared work programs (LaLonde, 1986). | Single study that provided the first direct comparison of RA and NR designs - they concluded that NR designs produced unreliable and potentially invalid causal inferences. | Made social scientists wary of non-randomized designs. |

Table A.2 – *continued from previous page*

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 1987 | Fraker and Maynard re-examined Lalonde's (1986) work programs comparison. | Replicated Lalonde's work but used a different sampling procedure and comparison group selection but found similar results as Lalonde - non-randomized designs produced inferior causal inferences. | Showed that even different analytic approaches produced similar results when comparing RA and NR designs. |
| 1987 | Heckman et al. re-examined Lalondes (1986) data. | Re-examined the data with more sophisticated statistical tools and arrived at different conclusions than Lalonde. | With two studies demonstrating the inferiority of non-randomized designs, Heckman showed that a statistical procedure could correct the weaknesses of poor designs and gave hope to social scientists that these designs might lead to reasonable inferences. |

Table A.2 – *continued from previous page*

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 1992 | Donley and Ashcraft published a study on the impact of self-selection on treatment outcomes. | Demonstrated in a solomon-four group design that self-selection had varying and, in some cases non-significant impact on treatment outcomes. | Almost no impact (cited 11 times at this point) but relevant to the current study. They demonstrated in a classroom setting that self-selection had minimal impact on effect sizes. Also, their results were scattered and unfocused but served as a good example of the selection bias estimation literature. |
| 1993 | Lipsey and Wilson published meta-meta-analysis. | First synthesis of meta-analytic results from multiple fields. Allowed them to estimate the effect of design characteristics (randomized vs. non-randomized studies) across multiple disciplines. Design had minimal impact on the mean estimate but the variance of effect sizes were influenced by design. | Provided greater justification for some researchers to use non-randomized designs. Despite their authors' intent, many researchers use this study as a justification for using non-randomized designs - suggesting that the effects probably do not differ in the long run. |

Table A.2 – *continued from previous page*

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 1996 | Heinsman and Shadish published a meta-analysis that focused on assessing whether non-randomized studies provided approximately the same effects as randomized studies. | Results indicated that a well-conducted NR design approximated an RA design for effect size estimates and causal inferences. | Focused methodologists' and social scientists' attention on comparison group selection. |
| 1996 | Shadish and Ragsdale published a meta-analysis on psychotherapy effectiveness. | Focused on assignment method unlike previous meta-analyses and results showed NR to be reliable if lower estimates of RA results, due to greater inter primary study result variance. | Revealed that differences between NR and RA results to be minimized with appropriate methodological or statistical tools. |

Table A.2 – *continued from previous page*

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 2000 | Benson and Hartz published a meta-analysis of medical studies. | Focused on a the perceived concern from the field that observational designs are inadequate are incomparable to RA designs. | Even in the medical literature a perception that observational designs are invalid was tested and proven with caveats to be an inaccurate bias against NR designs. |
| 2003 | Glazerman, Levy, and Myers published a review of the many examinations of the work program studies that brought to light concerns about NR designs. | Finding mixed results from NR studies approximating RA studies in the economic literature. | Years of economic research has no clear and definitive answer about NR result reliability and validity. |
| 2005 | Ferriter and Huband published a systematic review of medical studies included and excluded from the Cochrane Collaboration database. | Their review concluded that NR designs are not de facto inferior studies in comparison to RA designs. | Evidence from this systematic review further lends credence to the quality of the study being as vital to reliable and valid results as the assignment mechanism. |

Table A.2 – *continued from previous page*

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 2008 | Shadish, Clark, and Steiner published an experiment assessing the effectiveness of various statistical adjustment tools on NR designs. | Results indicated that the various statistical tools were adequate for adjusting NR results in order to replicate RA findings. | This experiment provided a template for treatment and measures used in my investigation of of selection bias effect sizes. |
| 2008 | Cook, Shadish, and Wong published a review summarizing current knowledge about NR designs results updating Glazerman et al. (2003). | Concluding NR results can replicate RA results with statistical adjustments or methodological alternatives. They Advocate for careful comparison group creation and specific covariate selection in statistical tools. | Particularly important is the authors recommendation to avoid "off-the-shelf" covariates (e.g. demographics) for either creating comparison groups or statistical covariates. This helped inform my studys use of assignment variables. |

Table A.2 – *continued from previous page*

| Year | Contribution | Relevance | Impact |
|---|---|---|---|
| 2009 | Pohl, Steiner, Eisermann, Soellner, and Cook published a replication of Shadish et al. (2008). | An independent replication of that came to similar conclusions as the original study but noted their differences were probably due to covariate selection. | Their work not only provided useful estimates of selection bias effects but also demonstrated how "off-the-shelf" covariates are not effective for all samples. |

Table A.2 provided me with several key markers to move forward with my research. Social programs produced through the Great Society initiative created the demand for evaluating hypotheses in field settings (Shadish & Cook, 1991). That demand facilitated the increased attention and consideration for threats to validity (Shadish et al., 2002). Specifically, multiple stakeholders - those for and against certain programs - started to pay attention to the "evidence" and carefully monitored what might affect the evidentiary basis of any program - particularly programs that were expensive or less aligned with their individual preferences. Social programs shifted research design from laboratory based experimental procedures (i.e., randomized designs) to field, observational, or quasi-experimental designs. These designs demanded methods to strengthen causal inference - something lacking from "weak" designs (Fraker & Maynard, 1987; Glazerman et al., 2003; LaLonde, 1986) - and the eventually lead the first popular statistical adjustment tools (Heckman et al., 1987). Then, meta-analysis became more prominent and almost mandated that researchers, program evaluators, and other stakeholders examine the relative comparability of research designs. Comparing these designs or even combining them to form the basis for policy became an essential step in almost all areas of research (e.g., medicine, psychology, economics, policy analysis, political science, and even physics). These comparisons lead researchers to focus on either solving or estimating the magnitude of the comparability problem. The aforementioned research programs focused on the solving the problem while my research program focuses on the latter - estimating the magnitude. Starting with Lipsey and Wilson (1993), researchers began to offer reasonable estimates of selection bias - just one marker of the comparability problem I refer to here. Despite these early contributions, I remained convinced that more refined estimates needed to be published to ensure adequate coverage of selection bias. Consider the following historical basis for my convictions.

### Quantitative, Meta-Analytic Reviews

Many reviews exist across a variety of the social sciences that provide a complicated overview of our understanding for NR effectiveness and selection bias effect sizes. Before enumerating

the many influential primary studies that brought about the current understanding of NR and RA differences and similarities along with specific experimental work, I first detail several quantitative systematic reviews.

A quantitative review or meta-analysis works as a method for aggregating effects across different studies. Smith and Glass (1977) began with a controversial meta-analysis that was followed up by no fewer than 10 re-analyses of the same data. Later, in (1982) Hunter and Schmidt and (1985) Hedges and Olkin brought meta-analysis to the forefront of multiple social science disciplines by publishing seminal books on meta-analysis. These books influenced the field greatly - often time polarizing sub-disciplines by placing values on research quality and scientific contribution.

Prior to meta-analysis, research summaries relied heavily on authors' discretion and implicit weighting schemes. Meta-analysis, however, required authors to explicitly state their selection criteria but, more importantly, allowed researchers to estimate the influence of design quality on observed experimental effects. Lipsey and Wilson (1993) provided a meta-meta-analysis that broke down research design effect sizes from meta-analyses of psychological treatment effectiveness outcomes. They concluded that "mean effect sizes for studies rated high and low for methodological quality found little difference" and, as a result, pushed all social scientists to question the relevance of research design to individual or aggregated effects (Lipsey & Wilson, 1993, p.1193).

The use of aggregated results to assess methodological quality or assignment mechanism is not restricted to economics or psychology. In medical research, NR studies and specifically observational studies, were considered inferior to RA studies and that observational methods inflated the observed effects. Benson and Hartz (2000) conducted a systematic review to assess those commonly held beliefs. To differentiate this review from other medical reviews, Benson and Hartz included studies from a variety of topics and included better sample aggregation through modern quantitative techniques (i.e., meta-analysis). They reported that observational studies did not produce effects dramatically higher than RA studies and that observational designs often meet the standards of classical experimental designs. Those

conclusions were similar to conclusions drawn from other systematic reviews in other fields including the economics.

Glazerman, Levy, and Myers (2003) provided a systematic review of several published economic studies. Those economic studies focused on whether NR studies can replicate RA study results. Unlike other reviews of these NR versus RA comparisons, Glazerman et. al.'s review included only primary studies with randomly assigned groups and some form of created non-randomly assigned comparison group. These primary studies included the Lalonde (1986) and Fraker and Maynard (1987) studies along with several subsequent studies that focused on job training programs. Their summary provided mixed results. NR studies under- and over-estimated the effects on individual earnings after people participated in an employment training program. Since Glazerman et al. (2003) relied upon actual dollar earnings instead of a metricless measure (e.g., probability of employment), the parameter ranges from the NR studies were difficult to estimate beyond rough and large yearly earnings. They were able to draw some tentative conclusions that NR studies do not accurately replicate RA studies. Furthermore, they concluded that a majority of the NR studies underestimated the RA results. They also offered some recommendations about how to create an NR study that closely replicated RA results. Those recommendations included larger sample sizes, statistical adjustment tools, and careful selection of comparison groups. Interestingly - simply due to the fact that Glazerman et. al. (2003) were economists - the statistical tools they recommended did not include the traditional economic tools; they found those traditional tools to be the least effective at reducing selection bias. According to their report, propensity scores and even basic regression with covariates did an adequate job of reducing selection bias.

Updating the work of Glazerman et al. (2003) and asking more focused questions Cook, Shadish, and Wong (2008) conducted a systematic review on a dozen published studies examining the differences between NR and RA results. Specific attention was paid to methodological or statistical tools that could be used to replicate results from RA studies and that comparison groups characteristics were beneficial to NR designs. They found that

Regression Discontinuity designs approximated RA design results while statistical adjustments (including propensity scores, heckman sample modeling, etc) were only effective if the selection process is known and accounted for in the data. Regarding NR design comparison group selection, Cook et al. (2008) echoed what many researchers previously stated: the comparison group ought to match the treatment for greater internal validity. Specifically, the authors argued that "off-the-shelf" variables for comparison group selection were neither beneficial nor effective in reducing selection bias. Comparison groups, according to Cook et al. (2008), need to created on variables relevant to the selection assignment (e.g. motivation, outcome pretest scores).

More relevant to psychology, Ferriter and Huband (2005) conducted a systematic review as a pilot study for examining the results of NR and RA Schizophrenia treatment outcome studies. In their review, they used the Cochrane Collaboration Database for compiling the primary treatment studies. Ferriter and Huband distinguished their review by categorizing the quality of the primary studies beyond the assignment mechanism. They reported that NR approximated the RA results. Additionally, they argued that the study quality - regardless of assignment - may influence the treatment effect magnitude. Specifically, higher quality studies resulted in lower treatment effects. Ferriter and Huband qualified their findings by acknowledging that since the Cochrane Database excludes most NR studies, their findings might not generalize to other NR studies (Ferriter & Huband, 2005).

The aforementioned studies were published after 2000 but other evidence existed prior to that time. Lipsey and Wilson (1993) gathered treatment outcome studies across behavioral, educational, and psychological fields. Instead of the traditional approach to meta-analysis where analysts combine primary empirical studies, Lipsey and Wilson conducted a meta-meta-analysis whereby they analyzed previously published meta-analyses. That is, prior published meta-analyses were the primary sources. Lipsey and Wilson (1993) analyzed the direction and magnitude of treatment effects along with the methodologically relevant predictors of those parameters. Their results showed the mean NR study design slightly underestimated the RA results - a conclusion that stood across treatment types and research

fields. Moreover, NR designs produced more variability in effect size estimates compared to RA designs.

With the intent to significantly increase the number of primary studies included in a meta-analysis from previous NR and RA comparison meta-analyses, Heinsman and Shadish (1996) focused on four distinct content areas (Scholastic Aptitude Test coaching, juvenile drug use prevention, psychosocial interventions for post surgery outcomes, intellectual ability groups). Across these content areas in this meta-analysis, the NR design mean effect size (d = .03) was significantly different from the RA mean effect size (d = .28). This difference, however, did not generalize across content areas. The juvenile drug use prevention and intellectual ability groups had substantially different mean results - based on assignment mechanism - but the Scholastic Aptitude Test coaching and psychosocial intervention research areas produced NR and RA mean results that were near replicates of one another. There were some qualifications to this finding. Regardless of the content area, if attrition levels or control group type were controlled for then NR results were increasingly similar to RA results. Using various breakdowns of the meta-analysis data, the authors concluded that NR effect sizes approximated RA effect sizes with careful selection of comparison groups (placebo vs. wait-list), with minimizing pretest differences and with avoiding self-selection.

With the same intent, Shadish and Ragsdale (1996) collected 100 primary studies that focused on psychotherapy outcomes. The meta-analysis was conducted in parallel with Heinsman and Shadish (1996) but contained a different content area and a more specific goal of reducing error by incorporating methodological quality indices. They found similar differences between NR and RA effect sizes as Heinsman and Shadish (1996). The absolute effect sizes differed slightly from the previous study but the conclusion remained the same; NR designs produced a significantly lower mean effect size (d = .08) compared to the meant effects produced by RA designs (d = .60). Again, they concluded that the raw significant difference between NR and RA effect sizes could be reduced by statistical adjustments (covariates in linear regression), careful selection of comparison groups, and when possible the elimination of self-selection according to the authors.

**Direct Comparisons between NR and RA Study Results**

What lead to those quantitative reviews began with researchers asking two questions: 1) Do NR studies approximate RA studies and if not 2) How much difference is there between NR and RA results. Those questions were first addressed by economic researchers almost three decades ago using federal work program samples.

Lalonde's (1986) study started in program evaluation with an expected finding that NR and RA studies did not produce equivalent effects nor were they equivalent with respect to causal inference. They examined separate employment and job training programs, National Supported Work Program (NSWD - an RA design), Manpower Development and Training Act, Comprehensive Employment and Training Act, and Job Training Partnership Act. One program used RA (NSWD) while the others did not. Their comparison of the four programs showed that each program produced somewhat different financial earnings outcomes - differences that may be potentially attributable to the assignment of individuals to their respective treatment programs. Individuals were either placed in a program based on pre-existing qualifications or they self-selected a treatment. The outcomes were significantly affected by selection method that Lalonde concluded that program evaluations could not rely upon NR studies to produce reliable and valid outcomes.

Similarly, Fraker and Maynard (1987) empirically examined the adequacy of comparison group designs for program evaluations. They utilized published field study data on employment and training programs - the same programs used by Lalonde. Fraker and Maynard (1987) created another comparison group from the separate but similar dataset CPS. The CPS comparison group were not randomly assigned participants and as such serve as the NR comparison group. The employment salaries from the NSWD data vary significantly from the salaries of the CPS participants in the comparisons. Similar to Lalonde, Fraker and Maynard (1987) concluded that NR designs cannot be relied on to estimate program effectiveness and any field studies using NR comparison group designs should be done so with great caution if at all. In short, the evidence was starting to accumulate in favor of RA designs and undermine the utility of NR designs.

Heckman, Hotz, and Dabos (1987) took issue with Lalonde (1986) and Fraker and Maynards (1987) conclusions. Heckman et al. (1987) pointed out the sample in the National Supported Work Demonstration (NSWD), Current Population Survey (CPS) and other government sponsored work programs identified as appropriate for field studies have selection bias concerns. While the absence of RA lead to the idea that selection bias was inherent in the comparison groups used in the NSWD and CPS samples, Heckman et al. (1987) pointed out that selection bias was not considered in the estimate of the outcome variable or participant earnings post training program. Heckman et al. contended that these relevant variables were excluded and the NR results were inappropriately described as "unreliable." With that concern, Heckman and his colleagues re-analyzed the samples and groups used by Lalonde (1986) and Fraker and Maynard (1987) but incorporated selection bias into the statistical models as well as examining the reliability of earnings estimates (i.e., the outcome of interest). Heckman concluded that the unreliability was largely due to the selection bias concerns in subsets of the comparison groups and that when selection bias was modelled the NR studies, the results approximated the RA study results. The authors recommended that those prior and their current analyses be used as examples of the important role selection bias plays in NR study results. Furthermore, they concluded that NR studies could be used to approximate classical experimental results in field studies.

While those economic field studies focused on comparing readily available samples that included self-selection groups and randomly assigned groups an experiment by psychological researchers sought better control of the assignment mechanism to obtain to understand the differences between NR and RA results. Shadish et al. (2008) implemented a strategy they termed as "Doubly Randomized Preference Trial" or DRPT. Researchers randomly assigned participants to either a Randomized Control Trial (RCT) or quasi-experimental study. Participants assigned to the RCT were either randomly assigned to a math or vocabulary treatment condition. Participants assigned to the quasi-experimental study choose either a math or vocabulary treatment based on their own preference. Participants received the same measures pre and post their assigned treatment. Quasi-experimental

results were adjusted by either basic linear regression, and various propensity score methods. These adjusted results were compared to the RCT determining which adjustment methods best approximated RCT results. Their intention was to determine if statistical tools can adjust NR study results to approximate experimental results not estimating selection bias. They found differences between NR and RA results but any of the statistical adjustment method could and did remove the selection bias effects from NR designs.

The majority of these studies focused on determining if there was a difference between NR and RA studies. While they concluded that there were selection bias differences between the two design types, they also concluded that adjustment techniques did an acceptable job of removing the biases and that the selection bias magnitudes themselves may not be as great as previously thought. For more detail on how much of a difference there is between NR and RA designs I will now turn my focus from these reviews and experimental conclusions above to the specific effect sizes reported in those studies.

### Selection Bias Effects

Estimating selection bias effects can be quite tricky. Early attempts that resulted in these estimates were not directed solely at this goal. Instead, researchers compared NR designs to RA designs almost as an after-thought rather than as a planned comparison via meta-analysis. Specifically, the meta-analysis by Shapiro and Shapiro (1983) provided a perfect springboard to this comparison because they included the assignment design as a moderator in their analysis. Their results indicated that NR studies produced lower effect sizes (d=.76) but within the range of RA effect sizes (d=.96). While these effect sizes differences may be indicative of an overall effect for study quality, Lipsey and Wilsons (1993) larger meta-meta-analysis suggested less striking results they stated that "the mean effect size for non-randomized control or comparison group designs [d=.41] is actually slightly smaller than that for randomized designs [d=.46]" (p.1193). Other researchers found mean differences by design that were much more striking. The .20 difference, for example, observed by Shapiro and Shapiro (1983) was similar to the aggregate .25 difference found

by Heinsman and Shadish (1996). A point worth noting is that Heinsman and Shadishs (1996) study was similar to Lipsey and Wilson (1993) with respect to content - a point I shall return to shortly. Even more striking, Shadish and Ragsdale (1996) found a dramatic difference (d=.52) between NR (d=.08) and RA (d=.60) designs their study also relied on content similar to Shapiro and Shapiro (1983). Thus, several researchers estimated selection bias effects by way of comparing effects via meta-analysis using different scientific content (psychotherapy, educational, medicine and other social science outcomes).

Focusing more on education with specifically designed studies, Donley and Ashcraft (1992) estimated selection bias effects ranging between d =.91, and d=.53. Their NR study examined these effects on a series of university physics test items. Participants self-selected into one of several comparison groups. They estimated selection bias effects item by item and reported that many items had no significant group difference the items that were different, however, produced medium to large effects (d > .90). Unfortunately, Donley and Ashcraft's (1992) study did not contain true RA results so the comparison might not be directly relevant to my interests but at least these results suggest a broad range of selection bias effects. Shadish, Clark and Steiner (2008), however, directly compared RA and NR designs using a similar educational outcomes in math and reading. Participants were randomly assigned to either a RA study or a NR study with both receiving the same procedures, experimental manipulation, and measures. The only difference between the groups was the assignment mechanism. Shadish et al., 2008 estimated a modest selection bias effect size (d = .24) that was similar to medium effect sizes estimated in prior meta-analyses (Shadish et al., 2008; Steiner et al., 2010). A subsequent replication (Pohl et al., 2009) of Shadish et. al's study produced a much smaller effect (d = .06). Taken together, these findings suggest that educational outcomes produce a wide array of selection bias effects that may be difficult to generalize across content domains.

The same conclusion drawn from the educationally-focused studies can be extended to the broader domain of estimating selection bias effects. There is no clear pattern or magnitude apparent in the corpus of studies focused on estimating selection bias effects (see

Table A.5.3). Selection bias effects appear to be quite variable between content domains and even within content domains this general conclusion may suggest that estimating selection bias effects may be so content-specific and perhaps even study-specific that an effort to estimate a single point-prediction or even a distribution of potential effects might appear to be a fools errand. Nevertheless, point estimates are relevant to all disciplines and must begin even when contingencies or conditional statements such as these are known a priori. In fact, the lack of reliable selection bias estimates is the basis for my research focus. Estimating selection bias effects across disciplines in carefully controlled trials will ensure that a broad array of researchers may be able to form expectations about selection bias and plan their studies accordingly.

Table A.3: Table of selection bias effect sizes

| Selection bias Effect Size (Cohen's d) | Study Type | Author & Year |
|---|---|---|
| 1.38 | self-selection experiment | Donley 1992 |
| .73 | self-selection experiment | Donley 1992 |
| .52 | meta-analysis | Shadish 1996 |
| .25 | meta-analysis | Heinsman 1996 |
| .24 | Doubly Randomized Preference Trial (DRPT) | Shadish 2008 |
| .20 | meta-analysis | Shapiro 1983 |
| .06 | DRPT replication | Pohl 2009 |
| .05 | meta-analysis | Lipsey 1993 |

## A.6 Conclusion

The literature review above provides a clear rationale for the importance of establishing a reasonable counterfactual for causal inference. The counterfactual comes from randomly assigning individuals to contrasting groups failure to do so leads to potential biases - referred to as selection biases - that are not clearly estimated in the literature. Many different approaches exist for treating selection bias but most require large sample sizes not readily available in social and behavioral sciences. Thus, I propose to follow a much simpler route - estimate the effect and subtract it from the observed effect. What remains unknown is a reasonable selection bias effect and I endeavor to estimate these effects. A single study will not produce the definitive effect but rather a series of studies - each improving and expanding upon the previous - will enable us to gain better insights into this effect. Just as researchers studying a specific content area progress in their research, I intend to follow suit. My first attempt in this research program is to estimate the effect - a replication of prior work with some refinements. Later, I hope to expand these efforts to estimate the stability across different disciplines. The end goal is to create a series of selection bias estimates that allow researchers across a wide array of content domains to apply a simple method of subtraction whereby they simply subtract out any potential bias from an observed effect. My first study in this research program begins with the documented study in this dissertation.

## Chapter B: Vocabulary Performance

1. UNDULATORY

1 - secretive

2 - motion characterized by successive rise and fall, like waves

3 - an underground, government base

4 - a process in which a political official is forcibly removed from office

5 - vulgar or offensive gestures

2. SNOFF

1 - powdery substance snorted through the nose

2 - an unruly child

3 - to ignore

4 - a long wicked candle used to light dynamite fuses

5 - to put oneself above others

3. GLEET

1 - a large woody hedge

2 - a small sheep

3 - a malfunction in a computers hard drive

4 - to impersonate another person or create a second identity

5 - a microscopic organism

4. GHERKIN

1 - a mans formal vest

2 - a literary term for shifting from one scene to another throughout a story

3 - a breed of horse trained specifically for racing

4 - a cucumber

5 - to speak in manner with the intention of confusing the listener

5. HARRIDAN

1 - a prostitute

2 - to run frantically

3 - a well made piece of furniture

4 - an antique

5 - a haggard, old woman

6. ONEIROCRITIC

1 - a medical procedure in which the liver is extracted

2 - someone who interprets dreams

3 - someone who suffers from hallucinations

4 - having great bearing upon a situation

5 - repeating the same series of behavior over and over

7. EPHEMERIS

1 - upper most layer of skin

2 - a compilation of household socioeconomic statuses taken from the U. S. census

3 - the atmospheric layer closest to Earth

4 - a table showing the predicted positions of heavenly bodies

5 - using a poetic way to express oneself

8. DENTILS

1 - small, tasteless beans

2 - tools used by dentists

3 - highway dividers

4 - political campaign volunteers

5 - an architectural style resembling teeth

9. FILIBEG

1 - a musical instrument developed in Iceland

2 - a large suitcase

3 - a tartan kilt

4 - to speak in a manner with the intention of confusing the listener

5 - a homeless person who begs for money

10. MOPOKE

1 - a treacherous ski slope

2 - a small cap worn in religious ceremonies

3 - a long, walking stick

4 - a thick blanket made out of woven animal hair

5 - an Australian bird whose call sounds like “mopoke, mopoke”

11. BESMIRCH

1 - to soil with smoke, soot or mud

2 - to be smug or sarcastic

3 - a tropical fish most often found off shores in the Eastern Caribbean

4 - the crank that was used to start the early Model T Ford

5 - the highest point of a mountain peak

12. THOB

1 - to hit forcefully with a blunt object

2 - to explain your beliefs and opinions

3 - a large portion of gelatinous material

4 - a clothing accessory worn about the neck

5 - to walk sluggishly

13. UNDERWRITER

1 - an employer who does not maintain a stable staff, but pays workers on a daily basis

2 - one accused of tax fraud

3 - someone who writes using a false name or as “Anonymous”

4 - one who does editing for advertising executives

5 - a subscriber or shareholder in a mercantile venture

14. FOOLSCAP

1 - a cap flamboyantly decorated

2 - the tip of a probing instrument

3 - the limit that a person can drink, determined by a bartender

4 - a group of unruly teenagers

5 - a medieval dance

15. NARGILE

1 - a reptile, usually three feet long, that is found in South American jungles

2 - silt found a the bottom of a river

3 - an elaborate hookah pipe

4 - a long fur piece worn around the shoulders

material that is made to resemble leather

16. RONDEAU

1 - a French prostitute

2 - procedures for vacating a rental home

3 - a small boat

4 - a short poem that uses refrains

5 - legal precedence protecting persons from slander

17. VERSIFY

1 - to write poetry

2 - to sing without musical accompaniment

3 - to talk continuously without making sense

4 - to express individuality

- to harm with intent

18. JEJUNE

1 - a small brown insect

2 - an automobile design to travel across rugged terrain

3 - to hit with great force

4 - a slang expression for being beautiful

5 - to go without food

19. CHITON

1 - the leader of a tribe

2 - an ancient Greek tunic

3 - a British toilet

4 - a short sword with a wide, curved blade

5 - marital rituals of Eastern India

20. WASSAILER

1 - a plant eating mammal that inhabits the southern Atlantic

2 - a narrow sailboat

3 - one who invests in stocks and bonds

4 - a violent sea storm

5 - one who takes part in riotous festivities

21. LOVELORN

1 - saddened or distressed by love

2 - a small bird

3 - one who is engaged to be married

4 - a gardening tool similar to a shovel

5 - lawn decorations

22. QUINSY

1 - to run very fast

2 - feeling of dizziness and nausea

3 - a place for sick people to be held until they recover

4 - inflammation of the throat

5 - an unsuccessful television show

23. IDEOLOGY

1 - the study of religions

2 - the study of infectious diseases

3 - the study of ideas

4 - the study of aesthetic placement

5 - the study of micro organisms

24. MUZHIK

1 a stringed musical instrument

2 a Russian peasant

3 a variation of popular tunes that is often played in elevators

4 a Middle Eastern rice dish

5 a traditional dance performed at pagan wedding ceremonies

25. BYSSUS

1 - an Egyptian monument

2 - reckless or wildly

3 - a financial transaction made between foreign countries

4 - a fine linen fabric

5 - an expensive French wine

26. SOBRIQUET

1 - a heavy wine sauce

2 - a small couch or settee

3 - a type of jewelry setting

4 - a ballet movement

5 - a nickname

27. ZITHER

1 - a stringed musical instrument

2 - a brothel

3 - ladies undergarment worn in the 18th century

4 - to exaggerate

5 - having no substance

28. ORYX

1 - a semi-precious gem

2 - an insect that is commonly found in Africa

3 - a chemical element found in iron that is often used as a black dye

4 - a type of antelope or gazelle

5 - a physical change that heavenly bodies undergo

29. VITUPERATION

1 a dangerous frenzy overtaking male elephants and camels

2 to be sarcastic

3 - using violent or abusive language

4 to vomit

5 a medical process that elements waste from the body

30. CHINQUAPIN

1 a vaccine used to treat blood disorders

2 - a nut tree found in Virginia and North Carolina

3 emphasizing unimportant characteristics

4 grossly distorted

5 a small island in the Eastern Pacific

## Chapter C: Mathematics Performance Revised

1. $(\mathrm{x}^a)(x^b) =$

1.)$xab$

2.)$2x^{ab}$

3.)$x^{a+b}$

4.)$2x^{a+b}$

5.)$x^{2ab}$

2. $(\mathrm{x}^a)^{ab} =$

1)$(x)^{ba2}$

2)$x^{a+b}$

3)$x^{ab}$

4)$2abx$

5)$x^{2a+b}$

3. $(\mathrm{x}^a)^{(a+b)} =$

1)$x^{(2a+b)}$

2)$x^{a*(a+b)}$

3)$x^{3a+b}$

4)$x^{a+ab}$

5)$(x2a)^b$

4. $(\mathrm{x}^a)(y^a) =$

1)$(xy)^{2a}$

2)$(xy)^a$

3)$xy^2$

4)$xy/a$

5)$xy^a$

5. $(\mathrm{x}^{2c})(y^{2c}) =$

1)$(xy)^{4c}$

2)$x^{4c} + y^{4c}$

3)$(xy)^{2c}$

4)$xy2c$

5)$(2xy)^c$

6. $(\mathrm{x}^{-c})(y^{-c}) =$

1)$xy^{2c}$

2)$1/x^c y^c$

3)$xy^c$

4)$1/(xy)^c$

5)$-xy^c$

7. $(\mathrm{x}^a)(x^b)(x^c)(x^d) =$

1)$x^{abcd}$

2)$x^{(ab+cd)}$

3)$x^{ab} + x^{cd}$

4)$(x^{ab})(x^{cd})$

5)$x^{a+b+c+d}$

8. $\mathrm{x}^a/x^b =$

1)$x^{ab}$

2)$x^{(a-b)}$

3)$x^{a/b}$

4)$2x^{(a+b)}$

5)$x^{(a+b)}$

9. $(\mathrm{x}^a)(x^b)(x^c)(x^a) =$

1)$x^{2a+b+c}$

2)$x^{a+b+c}$

3)$x^{2a+bc}$

4)$x^{abc}$

5)$x^{2a(bc)}$

10. $(\mathrm{x})^a(y)^a(z)^a =$

1)$x^a y^a z^a$

2)$(xyz)^a$

3)$xyz^{3a}$

4)$x^a + y^a + z^a$

5)$(xyz)^{a+a+a}$

11. $(2\mathrm{x}^a)(y^a)(3x^a) =$

1)$5x^{aya}$

2)$6xy^a$

3)$6x^{2a}y^a$

4)$5xy^a$

5) $-x^{2aya}$

12. $(\mathrm{x}^a)^{bc} =$

1)$x^{a+bc}$

2)$x^{a+b+c}$

3)$1/x^{abc}$

4)$x^{abc}$

5)$1/x^{a+b+c}$

13. $(\mathrm{x}^a)^a =$

1)$x^a$

2)$1/x^a$

3)$1/x^{a+a}$

4)$x/a$

5)$x^{a^2}$

14. $(\mathrm{x}^a)/(x^a) =$

1)$x^0$

2)$2ax$

3)$1/x^a$

4)$x^1$

5)$1/x^a$

15. $(\mathrm{x}^{-a})(x^{-a}) =$

1)$x^{-a}$

2)$1/x^{2a}$

3)$2x^a$

4) $-2x^a$

5)$1/x$

16. $(\mathrm{x}^a)^{-b} =$

1)$x^{ab}$

2)$xa-b$

3)$ax^b$

4)$1/x^{ab}$

5) $-x^{a+b}$

17. $1/(\mathrm{x}^a) =$

1)$x^{2a}$

2) $-x$

3)$(x^{2a)}(x^{-3a})$

4)$x-(x^a)$

5)$(1/x)(1/x)$

18. $(4^5)^{-3} =$

1) $-4^8$

2)$4^2$

3)$1/4^2$

4)$(4)(15)$

5)$1/4^{15}$

19. $\mathrm{x}^a/x^a =$

1)$x^{2a}$

2)$x^{(a+a)}$

3) $-1$

4) $x^{(a-a)}$

5) $1/x^{2a}$

20. $(3^3)(3^5)/(3^7) =$

1) $1/9^{15}$

2) 3

3) $9^8/3^7$

4) $3^{15} - 3^7$

5) $3^{8/7}$

# Chapter D: R Code

```
## Validity Study 1: selection bias pilot data from 12-3-10 and 12-10-10


## NOTES:  subjects with ID's between 1000-1800 are from my November pilot data collection
##         subjects with ID's between 1900-1999 are from my training session for the ura's in March;
##          all received math tx
##         subjects with ID's between 2000-2999 are from ura's in March


#######random assignment
?randomize()
id <- 2000:3000
scores <- c(350:500,650:800,by=10)
adf


ra_grps <- randomize(,group= c("MATH","VOCAB"), match=scores, complete=TRUE)


ra <- sample(0:1, 40, TRUE)
 ## [1] 1 1 0 0 1 0 0 0 0 1 0 1 0 0 1 0 1 1 1 1 1 1 0 1 0 1 1 0 1 0 1 1 0 1 1 0 1 1
## [39] 1 1


library(foreign)
library(plotrix)
library(psych)
library(psychometric)
library(experiment)
lirabrary(car)
```

```
library(boot)

library(MASS)

library(xtable)




## Dissertation Data

dissd <- read.csv("~/vs_4.29.csv",T)

head(dissd)

str(dissd)

names(dissd)

## DATA CLEANING

dissd$id_1



## remove my test run of cases: ids in 1900s and id 0

dissd <- dissd[-c(1:5,15),]

## remove study procedural variables: stop_n, id(limesurvey''s), start_n,

dissd <- dissd[,-c(1,2,3,5,6,55,92,108)]

## recode case for id_1

dissd[12,1] <- 2507 ## 2705 to 2507

dissd[17,1] <- 2512 ## 102512 to 2512

dissd[18,1] <- 2511 ##  102511 to 2511

dissd[19,1] <- 2510 ## 102510 to 2510

dissd[27,1] <- 2011 ##  10 to 2011

dissd[29,1] <- 2016 ## 102016 to 2016

dissd[38,1] <- 2521 ##  102521 to 2521

dissd[43,1] <- 2019 ## 102019 to 2019

dissd[85,1] <- 2043 ## 20403 to 2043



## recode sat score

dissd[87,35] <- 520 ## 526 to 520

dissd[97,35] <- 330 ## 303 to 330
```

```
dissd[53,36] <- 550 ## 5503 to 550


## remove cases with SAT's btwn 530-630
dissd <- dissd[-c(11,15,16,105,117),]




####math intervention cases (1:38) #cleaned rows
####vocab intervention cases (39:70) #cleaned rows
####cases 71 #after cleaning is me for my testing the scoring keys


########export cleaned file
## write.csv(pilot, "/Users/itwasi/coursework/dissertation/data/pilot_clean.csv", sep=",")


## Scoring the test




## Vocabulary Pre-test: V1_1-15
v1.key <- c(3,5,5,1,3,5,3,2,5,4,4,1,4,4,1)
## V2 Part 1 and 2 Scoring key ##in ets order
v2.key <- c(4,5,2,4,3,2,3,3,1,3,3,2,4,3,3,4,1,1,2,5,5,5,4,1,3,4,2,3,5,5,2,2,5,2,4,1)
## Vocabulary Post-test: V1_1.1-30 is the original Shadish order
## "V1_7.1" "V1_8.1"   "V1_14.1"  "V1_16"    "V1_11.1"  "V1_3.1"   "V1_4.1"
## "V1_17"    "V1_18"    "V1_19"    "V1_5.1"   "V1_2.1"   "V1_20"
## "V1_9.1"   "V1_15.1"  "V1_21"    "V1_22"    "V1_23"    "V1_24"
## "V1_6.1"   "V1_12.1"  "V1_13.1"  "V1_25"    "V1_26"    "V1_27"
## "V1_28"    "V1_29"    "V1_10.1"  "V1_1.1"   "V1_30"
v3.key <-c(3,2,4,1,4,5,1,2,4,5,3,5,2,5,1,4,1,5,2,5,1,4,3,2,4,5,1,4,3,2)


## precentage correct
```

```
# score.multiple.choice(v1.key, dissd[,c(2:16)], score=TRUE, totals=FALSE, missing=FALSE, short=FALSE)
# score.multiple.choice(v2.key, dissd[,c(50:85)], score=TRUE, totals=FALSE, missing=FALSE, short=FALSE)
# score.multiple.choice(v3.key, dissd[,c(102:131)], score=TRUE, totals=FALSE, missing=FALSE, short=FALSE)


v1 <- score.multiple.choice(v1.key, dissd[,c(2:16)], score=TRUE, totals=TRUE, missing=FALSE, short=FALSE)
v2 <- score.multiple.choice(v2.key, dissd[,c(50:85)], score=TRUE, totals=TRUE, missing=FALSE, short=FALSE)
v3 <- score.multiple.choice(v3.key, dissd[,c(102:131)], score=TRUE, totals=TRUE, missing=FALSE, short=FALSE)
##math intervention
# mean(v3$scores[1:38,])
# mean(v1$scores[1:38,])
# str(dissd$id_1)
# ##vocab inervention
# mean(v3$scores[39:70,])
# mean(v1$scores[39:70,])


## Math Pre-test: A1_1-10
# a1.key <- c(3,2,1,2,3,3,3,1,4,4)
a1r.key <- c(3,2,4,1,1,5,1,2,4,3)
## RG1 Part 1 Scoring key ##in ets order
rg1.key <- c(2,5,1,4,3,2,3,4,4,1,2,3,5,2,1)
## Math Post-test: A1_1.1-20 is the orginal Shadish order
## "A1_1.1"  "A1_11"    "A1_12"    "A1_2.1"   "A1_13"    "A1_3.1"   "A1_14"
## "A1_4.1"   "A1_5.1"   "A1_15"    "A1_16"    "A1_17"    "A1_6.1"
## "A1_7.1"   "A1_8.1"   "A1_9.1"   "A1_10.1"  "A1_18"    "A1_19"   "A1_20"
# a3.key <- c(3,4,3,2,2,1,2,2,3,3,5,1,3,3,1,4,4,5,3,2)
a3r.key <- c(3,1,2,2,3,4,5,1,1,2,3,4,5,1,2,4,3,5,4,2)


## percentage correct
# score.multiple.choice(a1r.key, dissd[,c(17:26)], score=TRUE, totals=FALSE, missing=FALSE, short=FALSE)
# score.multiple.choice(rg1.key, dissd[,c(86:100)], score=TRUE, totals=FALSE, missing=FALSE, short=FALSE)
# score.multiple.choice(a3r.key, dissd[,c(132:151)], score=TRUE, totals=FALSE, missing=FALSE, short=FALSE)
```

```
m1 <- score.multiple.choice(a1r.key, dissd[,c(17:26)], score=TRUE, totals=TRUE, missing=FALSE, short=FALSE)

m2 <- score.multiple.choice(rg1.key, dissd[,c(86:100)], score=TRUE, totals=TRUE, missing=FALSE, short=FALSE)

m3 <- score.multiple.choice(a3r.key, dissd[,c(132:151)], score=TRUE, totals=TRUE, missing=FALSE, short=FALSE)


##math intervention

# mean(m3$scores[1:38,])

# mean(m1$scores[1:38,])

#

# ##vocab inervention

# mean(m3$scores[39:70,])

# mean(m1$scores[39:70,])


#######combine scored results to data

dissd$v1sc <- v1$scores

dissd$v2sc <- v2$scores

dissd$v3sc <- v3$scores


dissd$m1sc <- m1$scores

dissd$m2sc <- m2$scores

dissd$m3sc <- m3$scores


############################### CORRELATIONS

## math SAT & math pretest


## math SAT & math pretest

# cor(dissd$D_9[ which(dissd$D_9 >= '610' | dissd$D_9 <= '530')],

# dissd$m1sc[ which(dissd$D_9 >= '610' | dissd$D_9 <= '530')])


#### ets & math pretest
```

```
m2.m1.r <- cor(dissd$m2sc, dissd$m1sc)


#### math SAT & math pretest

d9.m1.r <- cor(dissd$D_9, dissd$m1sc)


#### math SAT & math posttest

d9.m3.r <- cor(dissd$D_9, dissd$m3sc)


#### vocab SAT & math pretest

d10.m1.r <- cor(dissd$D_10, dissd$m1sc)


#### gender & math pretest

d2.m1.r <- cor(as.numeric(dissd$D_2), dissd$m1sc)


#### ets & math SAT

d9.m2.r <- cor(dissd$m2sc, dissd$D_9)


#### perference & pretest (math == 1, vocab == 2)

d14.m1.r <- cor(as.numeric(dissd$D_14), dissd$m1sc, use="complete.obs")


#####gender & math SAT

d2.d9.r <- cor(as.numeric(dissd$D_2), dissd$D_9,)


#####gender & preference

cor(as.numeric(dissd$D_2), as.numeric(dissd$D_14), use="complete.obs")


corr <- data.frame(d9.m1.r, m2.m1.r, d10.m1.r, d2.m1.r, d14.m1.r)

str(corr)
```

```
xtable(corr, caption="Pretest and Math Performance Correlations", label="corr")


##################Descriptives
## Looking at differences pre and post math scores
str(m1)
mean(m3$score)
## summary(dissd[1:38,35])
## summary(dissd[39:70,35])


## Math Performance Scores
# one sd is 100pt difference, then 2 sd-480 & 680
# demo graphics of gmu students


str(o1m)
o1m <- dissd$m1sc[ which(dissd$D_9 >= "610" & dissd$id_1 < "2100") ]
o2m <- dissd$m3sc[ which(dissd$D_9 >= '610' & dissd$id_1 < 2100) ]
o3m <- dissd$m1sc[ which(dissd$D_9 >= '610' & dissd$id_1 > 2100) ]
o4m <- dissd$m3sc[ which(dissd$D_9 >= '610' & dissd$id_1 > 2100) ]
o5m <- dissd$m1sc[ which(dissd$D_9 <= '530' & dissd$id_1 < 2100) ]
o6m <- dissd$m3sc[ which(dissd$D_9 <= '530' & dissd$id_1 < 2100) ]
o7m <- dissd$m1sc[ which(dissd$D_9 <= '530' & dissd$id_1 > 2100) ]
o8m <- dissd$m3sc[ which(dissd$D_9 <= '530' & dissd$id_1 > 2100) ]


## Vocabulary Performance Scores
o1v <- dissd$v1sc[ which(dissd$D_9 >= '610' & dissd$id_1 < 2100) ]
o2v <- dissd$v3sc[ which(dissd$D_9 >= '610' & dissd$id_1 < 2100) ]
o3v <- dissd$v1sc[ which(dissd$D_9 >= '610' & dissd$id_1 > 2100) ]
o4v <- dissd$v3sc[ which(dissd$D_9 >= '610' & dissd$id_1 > 2100) ]
o5v <- dissd$v1sc[ which(dissd$D_9 <= '530' & dissd$id_1 < 2100) ]
o6v <- dissd$v3sc[ which(dissd$D_9 <= '530' & dissd$id_1 < 2100) ]
o7v <- dissd$v1sc[ which(dissd$D_9 <= '530' & dissd$id_1 > 2100) ]
```

```
o8v <- dissd$v3sc[ which(dissd$D_9 <= '530' & dissd$id_1 > 2100) ]


###### EXPERIMENTAL GROUPINGS
dissd$grp1[dissd$D_9 >= '610' & dissd$id_1 < '2100'] <- "hi.m"
dissd$grp1[dissd$D_9 >= '610' & dissd$id_1 > '2100'] <- "hi.v"
dissd$grp1[dissd$D_9 <= '530' & dissd$id_1 < '2100'] <- "lo.m"
dissd$grp1[dissd$D_9 <= '530' & dissd$id_1 > '2100'] <- "lo.v"


dissd$grp2[dissd$D_2 == 'M' & dissd$id_1 < '2100'] <- "male.m"
dissd$grp2[dissd$D_2 == 'M' & dissd$id_1 > '2100'] <- "male.v"
dissd$grp2[dissd$D_2 == 'F' & dissd$id_1 < '2100'] <- "female.m"
dissd$grp2[dissd$D_2 == 'F' & dissd$id_1 > '2100'] <- "female.v"




table(dissd$grp2)
# om <- cbind(o1m, o2m, o3m, o4m, o5m, o6m, o7m, o8m)
########################## ABILITY




#group 1
# summary(dissd$D_9[ which(dissd$D_9 >= "610" & dissd$id_1 < "2100") ])
# sd(dissd$D_9[ which(dissd$grp1 == 'hi.m') ])  ###this code works
# sd(dissd$D_10[ which(dissd$D_9 >= "610" & dissd$id_1 < "2100") ])
# ## ETS MATH
# mean(dissd$m2sc[ which(dissd$D_9 >= "610" & dissd$id_1 < "2100") ])
# ## ETS VOCAB
# mean(dissd$v2sc[ which(dissd$D_9 >= "610" & dissd$id_1 < "2100") ])


# #group 2
```

```
# sd(dissd$D_9[ which(dissd$D_9 >= '610' & dissd$id_1 > 2100) ])
# summary(dissd$D_10[ which(dissd$D_9 >= '610' & dissd$id_1 > 2100) ])
# ## ETS MATH
# mean(dissd$m2sc[ which(dissd$D_9 >= '610' & dissd$id_1 > 2100) ])
# ## ETS VOCAB
# mean(dissd$v2sc[ which(dissd$D_9 >= '610' & dissd$id_1 > 2100) ])
#
# # group 3
# summary(dissd$D_9[ which(dissd$D_9 <= '530' & dissd$id_1 < 2100) ])
# > summary(dissd$D_10[ which(dissd$D_9 <= '530' & dissd$id_1 < 2100) ])
# ## ETS MATH
# mean(dissd$m2sc[ which(dissd$D_9 <= '530' & dissd$id_1 < 2100) ])
# ## ETS VOCAB
# mean(dissd$v2sc[ which(dissd$D_9 <= '530' & dissd$id_1 < 2100) ])
#
# # group 4
# summary(dissd$D_9[ which(dissd$D_9 <= '530' & dissd$id_1 > 2100) ])
# > summary(dissd$D_10[ which(dissd$D_9 <= '530' & dissd$id_1 > 2100) ])
# ## ETS MATH
# mean(dissd$m2sc[ which(dissd$D_9 <= '530' & dissd$id_1 > 2100) ])
# ## ETS VOCAB
# mean(dissd$v2sc[ which(dissd$D_9 <= '530' & dissd$id_1 > 2100) ])


hi.m.desc.d9 <- describe(dissd$D_9[ which(dissd$grp1 == 'hi.m') ], skew=FALSE, ranges=FALSE)
hi.v.desc.d9 <- describe(dissd$D_9[ which(dissd$grp1 == 'hi.v') ], skew=FALSE, ranges=FALSE)
lo.m.desc.d9 <- describe(dissd$D_9[ which(dissd$grp1 == 'lo.m') ], skew=FALSE, ranges=FALSE)
lo.v.desc.d9 <- describe(dissd$D_9[ which(dissd$grp1 == 'lo.v') ], skew=FALSE, ranges=FALSE)
desc.d9 <- rbind(hi.m.desc.d9,hi.v.desc.d9,lo.m.desc.d9,lo.v.desc.d9)


hi.m.desc.d10 <- describe(dissd$D_10[ which(dissd$grp1 == 'hi.m') ], skew=FALSE, ranges=FALSE)
hi.v.desc.d10 <- describe(dissd$D_10[ which(dissd$grp1 == 'hi.v') ], skew=FALSE, ranges=FALSE)
lo.m.desc.d10 <- describe(dissd$D_10[ which(dissd$grp1 == 'lo.m') ], skew=FALSE, ranges=FALSE)
```

```
lo.v.desc.d10 <- describe(dissd$D_10[ which(dissd$grp1 == 'lo.v') ], skew=FALSE, ranges=FALSE)
desc.d10 <- rbind(hi.m.desc.d10,hi.v.desc.d10,lo.m.desc.d10,lo.v.desc.d10)


hi.m.desc.m2 <- describe(dissd$m2sc[ which(dissd$grp1 == 'hi.m') ], skew=FALSE, ranges=FALSE)
hi.v.desc.m2 <- describe(dissd$m2sc[ which(dissd$grp1 == 'hi.v') ], skew=FALSE, ranges=FALSE)
lo.m.desc.m2 <- describe(dissd$m2sc[ which(dissd$grp1 == 'lo.m') ], skew=FALSE, ranges=FALSE)
lo.v.desc.m2 <- describe(dissd$m2sc[ which(dissd$grp1 == 'lo.v') ], skew=FALSE, ranges=FALSE)
desc.m2 <- rbind(hi.m.desc.m2,hi.v.desc.m2,lo.m.desc.m2,lo.v.desc.m2)


hi.m.desc.v2 <- describe(dissd$v2sc[ which(dissd$grp1 == 'hi.m') ], skew=FALSE, ranges=FALSE)
hi.v.desc.v2 <- describe(dissd$v2sc[ which(dissd$grp1 == 'hi.v') ], skew=FALSE, ranges=FALSE)
lo.m.desc.v2 <- describe(dissd$v2sc[ which(dissd$grp1 == 'lo.m') ], skew=FALSE, ranges=FALSE)
lo.v.desc.v2 <- describe(dissd$v2sc[ which(dissd$grp1 == 'lo.v') ], skew=FALSE, ranges=FALSE)
desc.v2 <- rbind(hi.m.desc.v2,hi.v.desc.v2,lo.m.desc.v2,lo.v.desc.v2)


gp1.desc <- cbind(desc.d9[,-c(1)], desc.d10[,-c(1,2)])
xtable(gp1.desc, caption="Ability Selection Group Descriptives", label="satability")
str(gp1.desc)
gp1.desc.ets <- cbind(desc.m2[,-c(1)], desc.v2[,-c(1,2)])
xtable(gp1.desc.ets, caption="Ability Selection Group Descriptives", label="etsability")




#####MPR pre and post test descriptives for ABILITY
hi.m.desc.m1 <- describe(dissd$m1sc[ which(dissd$grp1 == 'hi.m') ], skew=FALSE, ranges=FALSE)
hi.v.desc.m1 <- describe(dissd$m1sc[ which(dissd$grp1 == 'hi.v') ], skew=FALSE, ranges=FALSE)
lo.m.desc.m1 <- describe(dissd$m1sc[ which(dissd$grp1 == 'lo.m') ], skew=FALSE, ranges=FALSE)
lo.v.desc.m1 <- describe(dissd$m1sc[ which(dissd$grp1 == 'lo.v') ], skew=FALSE, ranges=FALSE)
desc.m1 <- rbind(hi.m.desc.m1,hi.v.desc.m1,lo.m.desc.m1,lo.v.desc.m1)
```

```
hi.m.desc.m3 <- describe(dissd$m3sc[ which(dissd$grp1 == 'hi.m') ], skew=FALSE, ranges=FALSE)
hi.v.desc.m3 <- describe(dissd$m3sc[ which(dissd$grp1 == 'hi.v') ], skew=FALSE, ranges=FALSE)
lo.m.desc.m3 <- describe(dissd$m3sc[ which(dissd$grp1 == 'lo.m') ], skew=FALSE, ranges=FALSE)
lo.v.desc.m3 <- describe(dissd$m3sc[ which(dissd$grp1 == 'lo.v') ], skew=FALSE, ranges=FALSE)
desc.m3 <- rbind(hi.m.desc.m3, hi.v.desc.m3, lo.m.desc.m3, lo.v.desc.m3)


m1m3.desc <- cbind(desc.m1[,-c(1,2)], desc.m3[,-c(1,2)])
xtable(m1m3.desc, caption="Ability Selection Group MPR Descriptives", label="mprability")




#################### GENDER
# #group 1
# summary(dissd$D_9[ which(dissd$D_2 == 'M' & dissd$id_1 < "2100") ])
# > summary(dissd$D_10[ which(dissd$D_2 == 'M' & dissd$id_1 < "2100") ])
# ## math pretest
# describe(o1mg)
# ## math posttest
# describe(o2mg)
# ## ETS MATH
# mean(dissd$m2sc[ which(dissd$D_2 == 'M' & dissd$id_1 < "2100") ])
# ## ETS VOCAB
# mean(dissd$v2sc[ which(dissd$D_2 == 'M' & dissd$id_1 < "2100") ])
#
# # group 2
#  summary(dissd$D_9[ which(dissd$D_2 == 'M' & dissd$id_1 > 2100) ])
#  summary(dissd$D_10[ which(dissd$D_2 == 'M' & dissd$id_1 > 2100) ])
# ## math pretest
# describe(o3mg)
# ## math posttest
# describe(o4mg)
```

```
# ## ETS MATH
# mean(dissd$m2sc[ which(dissd$D_2 == 'M' & dissd$id_1 > 2100) ])
# ## ETS VOCAB
# mean(dissd$v2sc[ which(dissd$D_2 == 'M' & dissd$id_1 > 2100) ])
#
# # group 3
#  summary(dissd$D_9[ which(dissd$D_2 == 'F' & dissd$id_1 < 2100) ])
#  summary(dissd$D_10[ which(dissd$D_2 == 'F' & dissd$id_1 < 2100) ])
# ## math pretest
# describe(o5mg)
# ## math posttest
# describe(o6mg)
# ## ETS MATH
# mean(dissd$m2sc[ which(dissd$D_2 == 'F' & dissd$id_1 < 2100) ])
# ## ETS VOCAB
# mean(dissd$v2sc[ which(dissd$D_2 == 'F' & dissd$id_1 < 2100) ])
#
# #group 4
# summary(dissd$D_9[ which(dissd$D_2 == 'F' & dissd$id_1 > 2100) ])
# summary(dissd$D_10[ which(dissd$D_2 == 'F' & dissd$id_1 > 2100) ])
# ## math pretest
# describe(o7mg)
# ## math posttest
# describe(o8mg)
# ## ETS MATH
# mean(dissd$m2sc[ which(dissd$D_2 == 'F' & dissd$id_1 > 2100) ])
# ## ETS VOCAB
# mean(dissd$v2sc[ which(dissd$D_2 == 'F' & dissd$id_1 > 2100) ])


male.m.desc.d9 <- describe(dissd$D_9[ which(dissd$grp2 == 'male.m') ], skew=FALSE, ranges=FALSE)
male.v.desc.d9 <- describe(dissd$D_9[ which(dissd$grp2 == 'male.v') ], skew=FALSE, ranges=FALSE)
female.m.desc.d9 <- describe(dissd$D_9[ which(dissd$grp2 == 'female.m') ], skew=FALSE, ranges=FALSE)
female.v.desc.d9 <- describe(dissd$D_9[ which(dissd$grp2 == 'female.v') ], skew=FALSE, ranges=FALSE)
```

```
desc.d9.gen <- rbind(male.m.desc.d9, male.v.desc.d9, female.m.desc.d9, female.v.desc.d9)


male.m.desc.d10 <- describe(dissd$D_10[ which(dissd$grp2 == 'male.m') ], skew=FALSE, ranges=FALSE)
male.v.desc.d10 <- describe(dissd$D_10[ which(dissd$grp2 == 'male.v') ], skew=FALSE, ranges=FALSE)
female.m.desc.d10 <- describe(dissd$D_10[ which(dissd$grp2 == 'female.m') ], skew=FALSE, ranges=FALSE)
female.v.desc.d10 <- describe(dissd$D_10[ which(dissd$grp2 == 'female.v') ], skew=FALSE, ranges=FALSE)
desc.d10.gen <- rbind(male.m.desc.d10, male.v.desc.d10, female.m.desc.d10, female.v.desc.d10)


male.m.desc.m2 <- describe(dissd$m2sc[ which(dissd$grp2 == 'male.m') ], skew=FALSE, ranges=FALSE)
male.v.desc.m2 <- describe(dissd$m2sc[ which(dissd$grp2 == 'male.v') ], skew=FALSE, ranges=FALSE)
female.m.desc.m2 <- describe(dissd$m2sc[ which(dissd$grp2 == 'female.m') ], skew=FALSE, ranges=FALSE)
female.v.desc.m2 <- describe(dissd$m2sc[ which(dissd$grp2 == 'female.v') ], skew=FALSE, ranges=FALSE)
desc.m2.gen <- rbind(male.m.desc.m2, male.v.desc.m2, female.m.desc.m2, female.v.desc.m2)


male.m.desc.v2 <- describe(dissd$v2sc[ which(dissd$grp2 == 'male.m') ], skew=FALSE, ranges=FALSE)
male.v.desc.v2 <- describe(dissd$v2sc[ which(dissd$grp2 == 'male.v') ], skew=FALSE, ranges=FALSE)
female.m.desc.v2 <- describe(dissd$v2sc[ which(dissd$grp2 == 'female.m') ], skew=FALSE, ranges=FALSE)
female.v.desc.v2 <- describe(dissd$v2sc[ which(dissd$grp2 == 'female.v') ], skew=FALSE, ranges=FALSE)
desc.v2.gen <- rbind(male.m.desc.v2, male.v.desc.v2, female.m.desc.v2, female.v.desc.v2)


gp2.desc <- cbind(desc.d9.gen[,-c(1)], desc.d10.gen[,-c(1,2)])
xtable(gp2.desc, caption="Gender Selection Group SAT Descriptives", label="satgender")


gp2.desc.ets <- cbind(desc.m2[,-c(1)], desc.v2[,-c(1,2)])
xtable(gp2.desc.ets, caption="Gender Selection Group ETS Descriptives", label="etsgender")


#####MPR pre and post test descriptives for GENDER
male.m.desc.m1 <- describe(dissd$m1sc[ which(dissd$grp2 == 'male.m') ], skew=FALSE, ranges=FALSE)
male.v.desc.m1 <- describe(dissd$m1sc[ which(dissd$grp2 == 'male.v') ], skew=FALSE, ranges=FALSE)
```

```
female.m.desc.m1 <- describe(dissd$m1sc[ which(dissd$grp2 == 'female.m') ], skew=FALSE, ranges=FALSE)
female.v.desc.m1 <- describe(dissd$m1sc[ which(dissd$grp2 == 'female.v') ], skew=FALSE, ranges=FALSE)
desc.m1.gen <- rbind(male.m.desc.m1, male.v.desc.m1, female.m.desc.m1, female.v.desc.m1)

male.m.desc.m3 <- describe(dissd$m3sc[ which(dissd$grp2 == 'male.m') ], skew=FALSE, ranges=FALSE)
male.v.desc.m3 <- describe(dissd$m3sc[ which(dissd$grp2 == 'male.v') ], skew=FALSE, ranges=FALSE)
female.m.desc.m3 <- describe(dissd$m3sc[ which(dissd$grp2 == 'female.m') ], skew=FALSE, ranges=FALSE)
female.v.desc.m3 <- describe(dissd$m3sc[ which(dissd$grp2 == 'female.v') ], skew=FALSE, ranges=FALSE)
desc.m3.gen <- rbind(male.m.desc.m3, male.v.desc.m3, female.m.desc.m3, female.v.desc.m3)

m1m3.desc.gen <- cbind(desc.m1.gen[,-c(1,2)], desc.m3.gen[,-c(1,2)])
xtable(m1m3.desc.gen, caption="Gender Selection Group MPR Descriptives", label="mprgender")




############## set up the percentage plots


grp1.ab <- c(70,82)
grp2.ab <- c(64,64)
grp3.ab <- c(39,59)
grp4.ab <- c(32,31)


plot(grp1.ab, type="o", col="dark red", ylim=c(30,100), axes=FALSE, ann=FALSE)
axis(1, at=1:2, lab=c("Pre","Post"))
axis(2, las=1, at=c(30,35,40,45,50,55,60,65,70,75,80,85,90,95,100))
box()
lines(grp2.ab, type="o", pch=22, lty=2, col="red")
lines(grp3.ab, type="o", pch=24, lty=1, col="blue")
lines(grp4.ab, type="o", pch=25, lty=2, col="dark blue")
```

```
title(main="Math Performance-Revised Test: Ability Assignment", font.main=4)
title(ylab="Percent Correct")
legend(1, 95, c("High-Math","High-Vocab", "Low-Math", "Low-Vocab"), cex=0.9,
 col=c("dark red","red", "blue", "dark blue"), pch=c(21,22,24,25), lty=c(1,2,5,6))


grp1.gen <- c(55,71)
grp2.gen <- c(47,50)
grp3.gen <- c(49,66)
grp4.gen <- c(43,39)


plot(grp1.gen, type="o", col="dark red", ylim=c(30,100), axes=FALSE, ann=FALSE)
axis(1, at=1:2, lab=c("Pre","Post"))
axis(2, las=1, at=c(30,35,40,45,50,55,60,65,70,75,80,85,90,95,100))
box()
lines(grp2.gen, type="o", pch=22, lty=2, col="red")
lines(grp3.gen, type="o", pch=24, lty=1, col="blue")
lines(grp4.gen, type="o", pch=25, lty=2, col="dark blue")
title(main="Math Performance-Revised Test: Gender Assignment", font.main=4)
title(ylab="Percent Correct")
legend(1, 95, c("Male-Math","Male-Vocab", "Female-Math", "Female-Vocab"), cex=0.9,
  col=c("dark red","red", "blue", "dark blue"), pch=c(21,22,24,25), lty=c(1,2,5,6))



### Test RE-test for Math Performance test
cor(o1m,o2m)
cor(o3m,o4m)
cor(o5m,o6m)
cor(o7m,o8m)


########################   EFFECT SIZES FOR ABILITY
```

```
##from r help listserv
cohens.d <- function (x, y) {(mean(x)-mean(y))/sqrt((var(x)+var(y))/2) }


##########  EFFECT SIZES
############## INITIAL DIFFERENCES
###math tx
cohens.d(o1m, o5m)


###MIN es
ab.min.mn <-cohens.d(o3m, o5m)
###MAX es
ab.max.mn <-cohens.d(o1m, o7m)


############## SELECTION BIAS INTERACTION
###math tx


##MIN es
ab.min.int <-cohens.d(o4m, o6m)
###MAX es
ab.max.int <- cohens.d(dissd$m3sc[ which(dissd$grp1 == 'hi.m') ],
   dissd$m3sc[ which(dissd$grp1 == 'lo.v') ]) ##this works
cohens.d(o2m, o8m)


### Expected Treatment Effect Size
ExpESTx <- (cohens.d(o2m,o4m) + cohens.d(o6m,o8m))/2
# cohens.d(o2m,o4m)
# cohens.d(o6m,o8m)


#####MIN. bias ES posttest
min.ES.post <- ab.min.int - ExpESTx
```

```
#####MAX. bias ES posttest
max.ES.post <- ab.max.int - ExpESTx


######################### Simple Bar Plot
counts <- table(mtcars$gear)
barplot(counts, main="SBI ES Distribution",
   xlab="Resampling Percentage", ylab="Effect Size")







#################   RESAMPLING AND EFFECT SIZE DISTRIUBTIONS


##resample 10% low in hi group, 20%lo in hi group to 80% etc...


###creating groups for resampling
o1m.d <- as.data.frame(o1m)
o2m.d <- as.data.frame(o2m)
o3m.d <- as.data.frame(o3m)
o4m.d <- as.data.frame(o4m)
o5m.d <- as.data.frame(o5m)
o6m.d <- as.data.frame(o6m)
o7m.d <- as.data.frame(o7m)
o8m.d <- as.data.frame(o8m)


# #### MINimum effect size groups
# o3o5m <- cbind(o3m.d, o5m.d)
# o4o6m <- cbind(o4m.d, o6m.d)
# #### MAXimum effect size group
# o1o7m <- cbind(o1m.d, o7m.d)
# o2o8m <- cbind(o2m.d, o8m.d)
```

```
######resampling functions and loop


###resample function
my.resample <- function (x, y, percent) {c(x[sample(1:nrow(x),
    nrow(x)*(1 - percent), replace=TRUE),], y[sample(1:nrow(y), nrow(y)*percent, replace=TRUE),])}


########## RESAMPLING


dissd$female <- recode(dissd$D_2,"'F'=0;'M'=1",F)


iter <- 1
i <- 1
j <- 2
x <- dissd




jn.resample <- function(x,iter=10,TxES=1.4){ ##see ExpESTx above - needed to take out Tx ES
  out <- rep(NA,6)
  for (j in 1:9){
    for (i in 1:iter){
      HImath <- x[x$grp1=='hi.m' | x$grp1=='hi.v',]
      LOmath <- x[x$grp1=='lo.m' | x$grp1=='lo.v',]
      male <- x[dissd$D_2=="M",]
      female <- x[dissd$D_2=="F",]
      datM <- rbind(HImath[sample(1:nrow(HImath),nrow(HImath)*(j/10)),],
                LOmath[sample(1:nrow(LOmath),nrow(LOmath)*(1-(j/10))),])
      datS <- rbind(male[sample(1:nrow(male),nrow(male)*(j/10)),],
                female[sample(1:nrow(female),nrow(female)*(1-(j/10))),])
```

```
      out <- rbind(out,c(i,j/10,cor(datM$D_9,datM$m1sc,use="complete.obs"),
                   cor(datS$female,datS$m1sc,use="complete.obs"),
                        cohens.d(datM$m3sc[datM$grp1=='hi.m' | datM$grp1=='hi.v'],
                        datM$m3sc[datM$grp1=='lo.m' | datM$grp1=='lo.v'])-TxES,
                        cohens.d(datS$m3sc[datS$D_2=="M"],datS$m3sc[datS$D_2=="F"])-TxES))
      ## the output dataset should have the following parts:
      # i, and j, along with rMath, rSex, ESmath, ESsex


    }
  }
  out <- as.data.frame(out)[-1,]
  names(out) <- c("iter","prop","rMath","rSex","ESmath","ESsex")
  return(out)
}


testing <- jn.resample(dissd,1000)
str(testing)
summary(testing)


plot(testing$ESmath~testing$rMath,)
abline(lm(testing$ESmath~I(testing$rMath^2),data=testing), col = "red")
points(.11, 1.59,pch=21, bg=20)
points(.62, 4.35,pch=21, bg=20)
summary(lm(abs(testing$ESmath)~abs(testing$rMath), data=testing))


plot(testing$ESsex~testing$rSex,) # ylim=c(0,4.5), xlim=c(0,.65)
abline(lm(abs(testing$ESsex)~abs(testing$rSex), data=testing), col = "red")
points(.11, 1.59, pch=21,bg=20)
points(.62, 4.35,pch=21, bg=20)
summary(lm(testing$ESsex~testing$rSex, data=testing))
```

```
####COMBINE so a single plot is available

es.df <- c(testing$ESmath, testing$ESsex)

r.df <- c(testing$rMath, testing$rSex)

head(es.df)

str(r.df)



plot(es.df~r.df,  ylim=c(-3,3), ylab="Effect Size",

       xlab= expression(paste("r" ["selection variable, pretest"])))

# abline(lm(es.df~ I(r.df^2)), col="red")

abline(lm(es.df~ abs(r.df)), lwd =3)

abline(lm(testing$ESmath~testing$rMath, data=testing), lwd=2, col="red" )

abline(lm(testing$ESsex~testing$rSex, data=testing), lwd=2, col="blue" )

points(r.df[c(1:9000)], es.df[c(1:9000)],  col="red") ##ability red

points(r.df[c(9001:18000)], es.df[c(9001:18000)], col="blue") ##gender blue

points(.11, min.ES.post.mg, pch=4,col="blue")

points(.11, max.ES.post.mg, pch=4,col="blue")

points(.62, max.ES.post,pch=4, col="red")

points(.62, min.ES.post,pch=4, col="red")




summary(lm(es.df~ abs(r.df)))

summary(lm(es.df~ r.df))

summary(lm(es.df~ I(r.df^2)), col="red")

summary(lm(abs(es.df)~ abs(r.df)))

summary(lm(es.df~ abs(r.df)))
```

```
#################### MINIMUM E.S. ABILITY INITIAL DIFFERENCES
o3o5.m.0 <- cohens.d(my.resample(o3m.d, o5m.d,0),o5m.d)
for (i in 1:999){
   o3o5.m.0 <- c(o3o5.m.0,cohens.d(my.resample(o3m.d, o5m.d,0),o5m.d))
 }


o3o5.m.10 <- cohens.d(my.resample(o3m.d, o5m.d, .10),o5m.d)
for (i in 1:999){
   o3o5.m.10 <- c(o3o5.m.10,cohens.d(my.resample(o3m.d, o5m.d, .10),o5m.d))
 }


o3o5.m.20 <- cohens.d(my.resample(o3m.d, o5m.d, .20),o5m.d)
for (i in 1:999){
    o3o5.m.20 <- c(o3o5.m.20,cohens.d(my.resample(o3m.d, o5m.d, .20),o5m.d))
  }

o3o5.m.30 <- cohens.d(my.resample(o3m.d, o5m.d, .30),o5m.d)
  for (i in 1:999){
     o3o5.m.30 <- c(o3o5.m.30,cohens.d(my.resample(o3m.d, o5m.d, .30),o5m.d))
   }

o3o5.m.40 <- cohens.d(my.resample(o3m.d, o5m.d, .40),o5m.d)
for (i in 1:999){
   o3o5.m.40 <- c(o3o5.m.40,cohens.d(my.resample(o3m.d, o5m.d, .40),o5m.d))
 }

o3o5.m.50 <- cohens.d(my.resample(o3m.d, o5m.d, .50),o5m.d)
for (i in 1:999){
   o3o5.m.50 <- c(o3o5.m.50,cohens.d(my.resample(o3m.d, o5m.d, .50),o5m.d))
 }


o3o5.m.60 <- cohens.d(my.resample(o3m.d, o5m.d, .60),o5m.d)
```

```
for (i in 1:999){
   o3o5.m.60 <- c(o3o5.m.60,cohens.d(my.resample(o3m.d, o5m.d, .60),o5m.d))
 }

o3o5.m.70 <- cohens.d(my.resample(o3m.d, o5m.d, .70),o5m.d)
for (i in 1:999){
   o3o5.m.70 <- c(o3o5.m.70,cohens.d(my.resample(o3m.d, o5m.d, .70),o5m.d))
 }

o3o5.m.80 <- cohens.d(my.resample(o3m.d, o5m.d, .80),o5m.d)
for (i in 1:999){
   o3o5.m.80 <- c(o3o5.m.80,cohens.d(my.resample(o3m.d, o5m.d, .80),o5m.d))
 }

o3o5.m.90 <- cohens.d(my.resample(o3m.d, o5m.d, .90),o5m.d)
for (i in 1:999){
   o3o5.m.90 <- c(o3o5.m.90,cohens.d(my.resample(o3m.d, o5m.d, .90),o5m.d))
 }

o3o5.m.100 <- cohens.d(my.resample(o3m.d, o5m.d, 1.0),o5m.d)
for (i in 1:999){
    o3o5.m.100 <- c(o3o5.m.100,cohens.d(my.resample(o3m.d, o5m.d, 1.0),o5m.d))
  }

o3o5.m <- c(o3o5.m.0, o3o5.m.10, o3o5.m.20, o3o5.m.30, o3o5.m.40, o3o5.m.50,
             o3o5.m.60, o3o5.m.70, o3o5.m.80, o3o5.m.90, o3o5.m.100)
hist(o3o5.m, main = "Minimum Ability Initial Difference", xlab = "Effect Size")
summary(o3o5.m)
describe(o3o5.m)

o3o5.df <- data.frame(o3o5.m.0, o3o5.m.10, o3o5.m.20, o3o5.m.30, o3o5.m.40, o3o5.m.50,
                      o3o5.m.60, o3o5.m.70, o3o5.m.80, o3o5.m.90, o3o5.m.100)

 str(o3o5.df)
o3o5.desc <- describe(o3o5.df, skew=FALSE, ranges=FALSE)
```

```
xtable(o3o5.desc,
      caption="Resampled Minimum Main Effects Descriptive Statistics",
      label="o3o5.m")


#################### MAXIMUM E.S. ABILITY INITIAL DIFFERENCES
o1o7.m.0 <- cohens.d(my.resample(o1m.d, o7m.d,0),o7m.d)
for (i in 1:999){
   o1o7.m.0 <- c(o1o7.m.0,cohens.d(my.resample(o1m.d, o7m.d,0),o7m.d))
 }


o1o7.m.10 <- cohens.d(my.resample(o1m.d, o7m.d, .10),o7m.d)
for (i in 1:999){
   o1o7.m.10 <- c(o1o7.m.10,cohens.d(my.resample(o1m.d, o7m.d, .10),o7m.d))
 }


o1o7.m.20 <- cohens.d(my.resample(o1m.d, o7m.d, .20),o7m.d)
for (i in 1:999){
    o1o7.m.20 <- c(o1o7.m.20,cohens.d(my.resample(o1m.d, o7m.d, .20),o7m.d))
  }

o1o7.m.30 <- cohens.d(my.resample(o1m.d, o7m.d, .30),o7m.d)
  for (i in 1:999){
     o1o7.m.30 <- c(o1o7.m.30,cohens.d(my.resample(o1m.d, o7m.d, .30),o7m.d))
   }

o1o7.m.40 <- cohens.d(my.resample(o1m.d, o7m.d, .40),o7m.d)
for (i in 1:999){
   o1o7.m.40 <- c(o1o7.m.40,cohens.d(my.resample(o1m.d, o7m.d, .40),o7m.d))
 }

o1o7.m.50 <- cohens.d(my.resample(o1m.d, o7m.d, .50),o1m.d)
for (i in 1:999){
```

```
   o1o7.m.50 <- c(o1o7.m.50,cohens.d(my.resample(o1m.d, o7m.d, .50),o7m.d))
 }


o1o7.m.60 <- cohens.d(my.resample(o1m.d, o7m.d, .60),o7m.d)
for (i in 1:999){
   o1o7.m.60 <- c(o1o7.m.60,cohens.d(my.resample(o1m.d, o7m.d, .60),o7m.d))
 }

o1o7.m.70 <- cohens.d(my.resample(o1m.d, o7m.d, .70),o7m.d)
for (i in 1:999){
   o1o7.m.70 <- c(o1o7.m.70,cohens.d(my.resample(o1m.d, o7m.d, .70),o7m.d))
 }

o1o7.m.80 <- cohens.d(my.resample(o1m.d, o7m.d, .80),o7m.d)
for (i in 1:999){
   o1o7.m.80 <- c(o1o7.m.80,cohens.d(my.resample(o1m.d, o7m.d, .80),o7m.d))
 }

o1o7.m.90 <- cohens.d(my.resample(o1m.d, o7m.d, .90),o7m.d)
for (i in 1:999){
   o1o7.m.90 <- c(o1o7.m.90,cohens.d(my.resample(o1m.d, o7m.d, .90),o7m.d))
 }

o1o7.m.100 <- cohens.d(my.resample(o1m.d, o7m.d, 1.0),o7m.d)
for (i in 1:999){
  o1o7.m.100 <- c(o1o7.m.100,cohens.d(my.resample(o1m.d, o7m.d, 1.0),o7m.d))
  }

o1o7.m <- c(o1o7.m.0, o1o7.m.10, o1o7.m.20, o1o7.m.30, o1o7.m.40,
            o1o7.m.50, o1o7.m.60, o1o7.m.70, o1o7.m.80, o1o7.m.90, o1o7.m.100)
hist(o1o7.m, main = "Maximum Ability Initial Difference", xlab = "Effect Size")
summary(o1o7.m)
describe(o1o7.m)
```

```
o1o7.df <- data.frame(o1o7.m.0, o1o7.m.10, o1o7.m.20, o1o7.m.30,
             o1o7.m.40, o1o7.m.50, o1o7.m.60, o1o7.m.70, o1o7.m.80, o1o7.m.90, o1o7.m.100)

str(o1o7.df)
o1o7.desc <- describe(o1o7.df, skew=FALSE, ranges=FALSE)
xtable(o1o7.desc,
     caption="Resampled Maximum Main Effects Descriptive Statistics",
       label="o1o7.m")




###Ability Initial Difference COMBINED
ability.id <- c(o3o5.m, o1o7.m)
id.df <- cbind(o3o5.desc, o1o7.desc, o3o5.mg.desc, o1o7.mg.desc)
xtable(id.df[,c(3,4,8,9,13,14,18,19)],
    caption="Resampled Main Effects Descriptive Statistics",label="ability.id")
names(id.df)


hist(ability.id, main = "Ability Pretest Effect", xlab = "Effect Size")
lines(1.05, 3000, type="h", col="blue")
lines(1.97, 3000, type="h", col="blue")


describe(ability.id)
ks.test(ability.id, "pnorm", mean=mean(ability.id), sd=sd(ability.id))
ks.test(ability.id, "pgamma", 1000, 5)
fitdistr(ability.id, "logistic")


ab.id <- xtable(describe(as.data.frame(ability.id)))
# print(ab.id)
```

```
#################### MINimum E.S. ABILITY SELECTION BIAS INTERACTIONS
o4o6.m.0 <- cohens.d(my.resample(o4m.d, o6m.d,0),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.0 <- c(o4o6.m.0,cohens.d(my.resample(o4m.d, o6m.d,0),o6m.d)-ExpESTx)
 }


o4o6.m.10 <- cohens.d(my.resample(o4m.d, o6m.d, .10),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.10 <- c(o4o6.m.10,cohens.d(my.resample(o4m.d, o6m.d, .10),o6m.d)-ExpESTx)
 }


o4o6.m.20 <- cohens.d(my.resample(o4m.d, o6m.d, .20),o6m.d)-ExpESTx
for (i in 1:999){
    o4o6.m.20 <- c(o4o6.m.20,cohens.d(my.resample(o4m.d, o6m.d, .20),o6m.d)-ExpESTx)
  }

o4o6.m.30 <- cohens.d(my.resample(o4m.d, o6m.d, .30),o6m.d)-ExpESTx
  for (i in 1:999){
     o4o6.m.30 <- c(o4o6.m.30,cohens.d(my.resample(o4m.d, o6m.d, .30),o6m.d)-ExpESTx)
   }

o4o6.m.40 <- cohens.d(my.resample(o4m.d, o6m.d, .40),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.40 <- c(o4o6.m.40,cohens.d(my.resample(o4m.d, o6m.d, .40),o6m.d)-ExpESTx)
 }

o4o6.m.50 <- cohens.d(my.resample(o4m.d, o6m.d, .50),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.50 <- c(o4o6.m.50,cohens.d(my.resample(o4m.d, o6m.d, .50),o6m.d)-ExpESTx)
 }
```

```
o4o6.m.60 <- cohens.d(my.resample(o4m.d, o6m.d, .60),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.60 <- c(o4o6.m.60,cohens.d(my.resample(o4m.d, o6m.d, .60),o6m.d)-ExpESTx)
 }

o4o6.m.70 <- cohens.d(my.resample(o4m.d, o6m.d, .70),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.70 <- c(o4o6.m.70,cohens.d(my.resample(o4m.d, o6m.d, .70),o6m.d)-ExpESTx)
 }

o4o6.m.80 <- cohens.d(my.resample(o4m.d, o6m.d, .80),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.80 <- c(o4o6.m.80,cohens.d(my.resample(o4m.d, o6m.d, .80),o6m.d)-ExpESTx)
 }

o4o6.m.90 <- cohens.d(my.resample(o4m.d, o6m.d, .90),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.90 <- c(o4o6.m.90,cohens.d(my.resample(o4m.d, o6m.d, .90),o6m.d)-ExpESTx)
 }


o4o6.m.100 <- cohens.d(my.resample(o4m.d, o6m.d, 1.0),o6m.d)-ExpESTx
for (i in 1:999){
   o4o6.m.100 <- c(o4o6.m.100,cohens.d(my.resample(o4m.d, o6m.d, 1.0),o6m.d)-ExpESTx)
 }

o4o6.m <- c(o4o6.m.0, o4o6.m.10, o4o6.m.20, o4o6.m.30, o4o6.m.40,
               o4o6.m.50, o4o6.m.60, o4o6.m.70, o4o6.m.80, o4o6.m.90, o4o6.m.100)
hist(o4o6.m, main = "Minimum Ability Selection Bias Interaction", xlab = "Effect Size")
summary(o4o6.m)
describe(o4o6.m)


o4o6.df <- data.frame(o4o6.m.0, o4o6.m.10, o4o6.m.20, o4o6.m.30,
```

```
            o4o6.m.40, o4o6.m.50, o4o6.m.60, o4o6.m.70, o4o6.m.80, o4o6.m.90, o4o6.m.100)

 str(o4o6.df)
o4o6.desc <- describe(o4o6.df, skew=FALSE, ranges=FALSE)
xtable(o4o6.desc,
     caption="Resampled Minimum Posttest Effects Descriptive Statistics",
      label="o4o6.m")




#################### MAXimum E.S. ABILITY SELECTION BIAS INTERACTIONS
o2o8.m.0 <- cohens.d(my.resample(o2m.d, o8m.d,0),o8m.d)-ExpESTx
for (i in 1:999){
   o2o8.m.0 <- c(o2o8.m.0,cohens.d(my.resample(o2m.d, o8m.d,0),o8m.d)-ExpESTx)
 }


o2o8.m.10 <- cohens.d(my.resample(o2m.d, o8m.d, .10),o8m.d)-ExpESTx
for (i in 1:999){
   o2o8.m.10 <- c(o2o8.m.10,cohens.d(my.resample(o2m.d, o8m.d, .10),o8m.d)-ExpESTx)
 }


o2o8.m.20 <- cohens.d(my.resample(o2m.d, o8m.d, .20),o8m.d)-ExpESTx
for (i in 1:999){
    o2o8.m.20 <- c(o2o8.m.20,cohens.d(my.resample(o2m.d, o8m.d, .20),o8m.d)-ExpESTx)
  }

o2o8.m.30 <- cohens.d(my.resample(o2m.d, o8m.d, .30),o8m.d)-ExpESTx
  for (i in 1:999){
     o2o8.m.30 <- c(o2o8.m.30,cohens.d(my.resample(o2m.d, o8m.d, .30),o8m.d)-ExpESTx)
   }

o2o8.m.40 <- cohens.d(my.resample(o2m.d, o8m.d, .40),o8m.d)-ExpESTx
```

```
for (i in 1:999){
   o2o8.m.40 <- c(o2o8.m.40,cohens.d(my.resample(o2m.d, o8m.d, .40),o8m.d)-ExpESTx)
 }

o2o8.m.50 <- cohens.d(my.resample(o2m.d, o8m.d, .50),o2m.d)-ExpESTx
for (i in 1:999){
   o2o8.m.50 <- c(o2o8.m.50,cohens.d(my.resample(o2m.d, o8m.d, .50),o8m.d)-ExpESTx)
 }


o2o8.m.60 <- cohens.d(my.resample(o2m.d, o8m.d, .60),o8m.d)-ExpESTx
for (i in 1:999){
   o2o8.m.60 <- c(o2o8.m.60,cohens.d(my.resample(o2m.d, o8m.d, .60),o8m.d)-ExpESTx)
 }

o2o8.m.70 <- cohens.d(my.resample(o2m.d, o8m.d, .70),o8m.d)-ExpESTx
for (i in 1:999){
   o2o8.m.70 <- c(o2o8.m.70,cohens.d(my.resample(o2m.d, o8m.d, .70),o8m.d)-ExpESTx)
 }

o2o8.m.80 <- cohens.d(my.resample(o2m.d, o8m.d, .80),o8m.d)-ExpESTx
for (i in 1:999){
   o2o8.m.80 <- c(o2o8.m.80,cohens.d(my.resample(o2m.d, o8m.d, .80),o8m.d)-ExpESTx)
 }

o2o8.m.90 <- cohens.d(my.resample(o2m.d, o8m.d, .90),o8m.d)-ExpESTx
for (i in 1:999){
   o2o8.m.90 <- c(o2o8.m.90,cohens.d(my.resample(o2m.d, o8m.d, .90),o8m.d)-ExpESTx)
 }


o2o8.m.100 <- cohens.d(my.resample(o2m.d, o8m.d, 1.0),o8m.d)-ExpESTx
for (i in 1:999){
   o2o8.m.100 <- c(o2o8.m.100,cohens.d(my.resample(o2m.d, o8m.d, 1.0),o8m.d)-ExpESTx)
 }
```

```
o2o8.m <- c(o2o8.m.0, o2o8.m.10, o2o8.m.20, o2o8.m.30, o2o8.m.40,
            o2o8.m.50, o2o8.m.60, o2o8.m.70, o2o8.m.80, o2o8.m.90, o2o8.m.100)
hist(o2o8.m, main = "Maximum Ability Selection Bias Interaction", xlab = "Effect Size")
describe(o2o8.m)

o2o8.df <- data.frame(o2o8.m.0, o2o8.m.10, o2o8.m.20, o2o8.m.30, o2o8.m.40,
                      o2o8.m.50, o2o8.m.60, o2o8.m.70, o2o8.m.80, o2o8.m.90, o2o8.m.100)

 str(o2o8.df)
o2o8.desc <- describe(o2o8.df, skew=FALSE, ranges=FALSE)
xtable(o2o8.desc,
       caption="Resampled Maximum Selection by Treatment Effects Descriptive Statistics",
        label="o2o8.m")




###Ability SBI COMBINED
ability.sbi <- c(o4o6.m, o2o8.m)
hist(ability.sbi, main = "Ability Posttest Effect", xlab = "Effect Size")
lines(min.ES.post, 6000, type="h", col="blue")
lines(max.ES.post, 6000, type="h", col="blue")


describe(ability.sbi)

########Descriptives for ability and gender posttest ES resamples
sbi.df <- cbind(o4o6.desc, o2o8.desc, o4o6.mg.desc, o2o8.mg.desc)
xtable(sbi.df[,c(3,4,8,9,13,14,18,19)],
       caption="Resampled Posttest Effects Descriptive Statistics",
         label="sbti.df")
```

```
ks.test(ability.sbi, "pnorm", mean=mean(ability.sbi), sd=sd(ability.sbi))




######################### Gender differences

## Math Performance Scores

# o1m <- dissd$m1sc[which(dissd$D_9 >= 610 & dissd$id_1 < 2100). ]

str(o7mg)

o1mg <- dissd$m1sc[ which(dissd$D_2 == 'M' & dissd$id_1 < 2100) ]

o2mg <- dissd$m3sc[ which(dissd$D_2 == 'M' & dissd$id_1 < 2100) ]

o3mg <- dissd$m1sc[ which(dissd$D_2 == 'M' & dissd$id_1 > 2100) ]

o4mg <- dissd$m3sc[ which(dissd$D_2 == 'M' & dissd$id_1 > 2100) ]

o5mg <- dissd$m1sc[ which(dissd$D_2 == 'F' & dissd$id_1 < 2100) ]

o6mg <- dissd$m3sc[ which(dissd$D_2 == 'F' & dissd$id_1 < 2100) ]

o7mg <- dissd$m1sc[ which(dissd$D_2 == 'F' & dissd$id_1 > 2100) ]

o8mg <- dissd$m3sc[ which(dissd$D_2 == 'F' & dissd$id_1 > 2100) ]

## Vocabulary Performance Scores

o1vg <- dissd$v1sc[ which(dissd$D_2 == 'M' & dissd$id_1 < 2100) ]

o2vg <- dissd$v3sc[ which(dissd$D_2 == 'M' & dissd$id_1 < 2100) ]

o3vg <- dissd$v1sc[ which(dissd$D_2 == 'M' & dissd$id_1 > 2100) ]

o4vg <- dissd$v3sc[ which(dissd$D_2 == 'M' & dissd$id_1 > 2100) ]

o5vg <- dissd$v1sc[ which(dissd$D_2 == 'F' & dissd$id_1 < 2100) ]

o6vg <- dissd$v3sc[ which(dissd$D_2 == 'F' & dissd$id_1 < 2100) ]

o7vg <- dissd$v1sc[ which(dissd$D_2 == 'F' & dissd$id_1 > 2100) ]

o8vg <- dissd$v3sc[ which(dissd$D_2 == 'F' & dissd$id_1 > 2100) ]



##########  EFFECT SIZES

############### INITIAL DIFFERENCES

###math tx

cohens.d(o1mg, o5mg)

###MIN es

gen.min.mn <-cohens.d(o3mg, o5mg)
```

```
###MAX es
gen.max.mn <-cohens.d(o1mg, o7mg)
## within male differences
cohens.d(o1mg, o3mg)
##within female differences
cohens.d(o5mg, o7mg)


############### SELECTION BIAS INTERACTION
###math tx


##MIN es
gen.min.int <- cohens.d(o4mg, o6mg)
###MAX es
gen.max.int <- cohens.d(dissd$m3sc[ which(dissd$grp2 == 'male.m') ],
               dissd$m3sc[ which(dissd$grp2 == 'female.v') ]) ##this works
gen.max.int <- cohens.d(o2mg, o8mg)
cohens.d(o2mg, o8mg)
#####Biased ES posttest gender
min.ES.post.mg <- gen.min.int - ExpESTx
max.ES.post.mg <- gen.max.int - ExpESTx

## within male differences
cohens.d(o2mg, o4mg)
##within female differences
cohens.d(o6mg, o8mg)


######ES MAINS
ab.es.mn <- cbind(ab.min.mn, ab.max.mn)
gen.es.mn <- cbind(gen.min.mn, gen.max.mn)
mn.es <- rbind(ab.es.mn, gen.es.mn)
xtable(mn.es, caption="ES for Main Effects", label="mpres1")
```

```
######ES INTERACTIONS
ab.es.int <- cbind(ab.min.int, ab.max.int)
gen.es.int <- cbind(gen.min.int, gen.max.int)
int.es <- rbind(ab.es.int, gen.es.int)
xtable(int.es, caption="ES for Interaction Effects", label="mpres2")


######################### Simple Bar Plot
counts <- table(mtcars$gear)
barplot(counts, main="SBI ES Distribution",
   xlab="Resampling Percentage", ylab="Effect Size")



###vocab tx
## biased treatment effect
## unbiased treatment effect




#################   RESAMPLING AND EFFECT SIZE DISTRIUBTIONS


##resample 10% low in hi group, 20%lo in hi group to 80% etc...


###creating groups for resampling
o1mg.d <- as.data.frame(o1mg)
o2mg.d <- as.data.frame(o2mg)
o3mg.d <- as.data.frame(o3mg)
o4mg.d <- as.data.frame(o4mg)
o5mg.d <- as.data.frame(o5mg)
o6mg.d <- as.data.frame(o6mg)
o7mg.d <- as.data.frame(o7mg)
o8mg.d <- as.data.frame(o8mg)
```

```
# #### MINimum effect size groups
# o3o5mg <- cbind(o3mg.d, o5mg.d)
# o4o6mg <- cbind(o4mg.d, o6mg.d)
# #### MAXimum effect size group
# o1o7mg <- cbind(o1mg.d, o7mg.d)


########## RESAMPLING
## MINimum E.S. GENDER INITIAL DIFFERENCES
o3o5.mg.0 <- cohens.d(my.resample(o3mg.d, o5mg.d,0),o5mg.d)
for (i in 1:999){
   o3o5.mg.0 <- c(o3o5.mg.0,cohens.d(my.resample(o3mg.d, o5mg.d,0),o5mg.d))
 }


o3o5.mg.10 <- cohens.d(my.resample(o3mg.d, o5mg.d, .10),o5mg.d)
for (i in 1:999){
   o3o5.mg.10 <- c(o3o5.mg.10,cohens.d(my.resample(o3mg.d, o5mg.d, .10),o5mg.d))
 }


o3o5.mg.20 <- cohens.d(my.resample(o3mg.d, o5mg.d, .20),o5mg.d)
for (i in 1:999){
    o3o5.mg.20 <- c(o3o5.mg.20,cohens.d(my.resample(o3mg.d, o5mg.d, .20),o5mg.d))
  }

o3o5.mg.30 <- cohens.d(my.resample(o3mg.d, o5mg.d, .30),o5mg.d)
  for (i in 1:999){
     o3o5.mg.30 <- c(o3o5.mg.30,cohens.d(my.resample(o3mg.d, o5mg.d, .30),o5mg.d))
   }

o3o5.mg.40 <- cohens.d(my.resample(o3mg.d, o5mg.d, .40),o5mg.d)
for (i in 1:999){
```

```
   o3o5.mg.40 <- c(o3o5.mg.40,cohens.d(my.resample(o3mg.d, o5mg.d, .40),o5mg.d))
 }

o3o5.mg.50 <- cohens.d(my.resample(o3mg.d, o5mg.d, .50),o5mg.d)
for (i in 1:999){
   o3o5.mg.50 <- c(o3o5.mg.50,cohens.d(my.resample(o3mg.d, o5mg.d, .50),o5mg.d))
 }


o3o5.mg.60 <- cohens.d(my.resample(o3mg.d, o5mg.d, .60),o5mg.d)
for (i in 1:999){
   o3o5.mg.60 <- c(o3o5.mg.60,cohens.d(my.resample(o3mg.d, o5mg.d, .60),o5mg.d))
 }

o3o5.mg.70 <- cohens.d(my.resample(o3mg.d, o5mg.d, .70),o5mg.d)
for (i in 1:999){
   o3o5.mg.70 <- c(o3o5.mg.70,cohens.d(my.resample(o3mg.d, o5mg.d, .70),o5mg.d))
 }

o3o5.mg.80 <- cohens.d(my.resample(o3mg.d, o5mg.d, .80),o5mg.d)
for (i in 1:999){
   o3o5.mg.80 <- c(o3o5.mg.80,cohens.d(my.resample(o3mg.d, o5mg.d, .80),o5mg.d))
 }

o3o5.mg.90 <- cohens.d(my.resample(o3mg.d, o5mg.d, .90),o5mg.d)
for (i in 1:999){
   o3o5.mg.90 <- c(o3o5.mg.90,cohens.d(my.resample(o3mg.d, o5mg.d, .90),o5mg.d))
 }

o3o5.mg.100 <- cohens.d(my.resample(o3mg.d, o5mg.d, 1.0),o5mg.d)
for (i in 1:999){
    o3o5.mg.100 <- c(o3o5.mg.100,cohens.d(my.resample(o3mg.d, o5mg.d, 1.0),o5mg.d))
  }

o3o5.mg <- c(o3o5.mg.0, o3o5.mg.10, o3o5.mg.20, o3o5.mg.30, o3o5.mg.40,
```

```
            o3o5.mg.50, o3o5.mg.60, o3o5.mg.70, o3o5.mg.80, o3o5.mg.90, o3o5.mg.100)
hist(o3o5.mg, main = "Minimum Gender Initial Difference", xlab = "Effect Size")
summary(o3o5.mg)
describe(o3o5.mg)


o3o5.mg.df <- data.frame(o3o5.mg.0, o3o5.mg.10, o3o5.mg.20, o3o5.mg.30, o3o5.mg.40,
            o3o5.mg.50, o3o5.mg.60, o3o5.mg.70, o3o5.mg.80, o3o5.mg.90, o3o5.mg.100)

 str(o3o5.mg.df)
o3o5.mg.desc <- describe(o3o5.mg.df, skew=FALSE, ranges=FALSE)
xtable(o3o5.mg.desc,
      caption="Resampled Minimum Main Effects Descriptive Statistics",
       label="o3o5.mg")


## MAXimum E.S. GENDER INITIAL DIFFERENCES
o1o7.mg.0 <- cohens.d(my.resample(o1mg.d, o7mg.d,0),o7mg.d)
for (i in 1:999){
   o1o7.mg.0 <- c(o1o7.mg.0,cohens.d(my.resample(o1mg.d, o7mg.d,0),o7mg.d))
 }


o1o7.mg.10 <- cohens.d(my.resample(o1mg.d, o7mg.d, .10),o7mg.d)
for (i in 1:999){
   o1o7.mg.10 <- c(o1o7.mg.10,cohens.d(my.resample(o1mg.d, o7mg.d, .10),o7mg.d))
 }


o1o7.mg.20 <- cohens.d(my.resample(o1mg.d, o7mg.d, .20),o7mg.d)
for (i in 1:999){
    o1o7.mg.20 <- c(o1o7.mg.20,cohens.d(my.resample(o1mg.d, o7mg.d, .20),o7mg.d))
  }

o1o7.mg.30 <- cohens.d(my.resample(o1mg.d, o7mg.d, .30),o7mg.d)
```

```
  for (i in 1:999){
     o1o7.mg.30 <- c(o1o7.mg.30,cohens.d(my.resample(o1mg.d, o7mg.d, .30),o7mg.d))
   }

o1o7.mg.40 <- cohens.d(my.resample(o1mg.d, o7mg.d, .40),o7mg.d)
for (i in 1:999){
   o1o7.mg.40 <- c(o1o7.mg.40,cohens.d(my.resample(o1mg.d, o7mg.d, .40),o7mg.d))
 }

o1o7.mg.50 <- cohens.d(my.resample(o1mg.d, o7mg.d, .50),o1mg.d)
for (i in 1:999){
   o1o7.mg.50 <- c(o1o7.mg.50,cohens.d(my.resample(o1mg.d, o7mg.d, .50),o7mg.d))
 }


o1o7.mg.60 <- cohens.d(my.resample(o1mg.d, o7mg.d, .60),o7mg.d)
for (i in 1:999){
   o1o7.mg.60 <- c(o1o7.mg.60,cohens.d(my.resample(o1mg.d, o7mg.d, .60),o7mg.d))
 }

o1o7.mg.70 <- cohens.d(my.resample(o1mg.d, o7mg.d, .70),o7mg.d)
for (i in 1:999){
   o1o7.mg.70 <- c(o1o7.mg.70,cohens.d(my.resample(o1mg.d, o7mg.d, .70),o7mg.d))
 }

o1o7.mg.80 <- cohens.d(my.resample(o1mg.d, o7mg.d, .80),o7mg.d)
for (i in 1:999){
   o1o7.mg.80 <- c(o1o7.mg.80,cohens.d(my.resample(o1mg.d, o7mg.d, .80),o7mg.d))
 }

o1o7.mg.90 <- cohens.d(my.resample(o1mg.d, o7mg.d, .90),o7mg.d)
for (i in 1:999){
   o1o7.mg.90 <- c(o1o7.mg.90,cohens.d(my.resample(o1mg.d, o7mg.d, .90),o7mg.d))
 }
```

```
o1o7.mg.100 <- cohens.d(my.resample(o1mg.d, o7mg.d, 1.0),o7mg.d)
for (i in 1:999){
  o1o7.mg.100 <- c(o1o7.mg.100,cohens.d(my.resample(o1mg.d, o7mg.d, 1.0),o7mg.d))
  }

o1o7.mg <- c(o1o7.mg.0, o1o7.mg.10, o1o7.mg.20, o1o7.mg.30, o1o7.mg.40,
             o1o7.mg.50, o1o7.mg.60, o1o7.mg.70, o1o7.mg.80, o1o7.mg.90, o1o7.mg.100)
hist(o1o7.mg, main = "Maximum Gender Initial Difference", xlab = "Effect Size")
summary(o1o7.mg)
describe(o1o7.mg)


o1o7.mg.df <- data.frame(o1o7.mg.0, o1o7.mg.10, o1o7.mg.20, o1o7.mg.30,
         o1o7.mg.40, o1o7.mg.50, o1o7.mg.60, o1o7.mg.70, o1o7.mg.80, o1o7.mg.90, o1o7.mg.100)

str(o1o7.mg.df)
o1o7.mg.desc <- describe(o1o7.mg.df, skew=FALSE, ranges=FALSE)
xtable(o1o7.mg.desc,
      caption="Resampled Maximum Main Effects Descriptive Statistics",label="o1o7.mg")




####Gender sbi effect COMBINED
gender.id <- c(o3o5.mg ,o1o7.mg)
hist(gender.id, main ="Gender Pretest Effect", xlab = "Effect Size")
lines(-.06, 2500, type="h", col="blue")
lines(.51, 2500, type="h", col="blue")


describe(gender.id)


ks.test(gender.id, "pnorm", mean=mean(gender.id), sd=sd(gender.id))
```

```
## MINimum E.S. GENDER SELECTION BIAS INTERACTIONS
o4o6.mg.0 <- cohens.d(my.resample(o4mg.d, o6mg.d,0),o6mg.d)-ExpESTx
for (i in 1:999){
   o4o6.mg.0 <- c(o4o6.mg.0,cohens.d(my.resample(o4mg.d, o6mg.d,0),o6mg.d)-ExpESTx)
 }


o4o6.mg.10 <- cohens.d(my.resample(o4mg.d, o6mg.d, .10),o6mg.d)-ExpESTx
for (i in 1:999){
   o4o6.mg.10 <- c(o4o6.mg.10,cohens.d(my.resample(o4mg.d, o6mg.d, .10),o6mg.d)-ExpESTx)
 }


o4o6.mg.20 <- cohens.d(my.resample(o4mg.d, o6mg.d, .20),o6mg.d)-ExpESTx
for (i in 1:999){
    o4o6.mg.20 <- c(o4o6.mg.20,cohens.d(my.resample(o4mg.d, o6mg.d, .20),o6mg.d)-ExpESTx)
  }

o4o6.mg.30 <- cohens.d(my.resample(o4mg.d, o6mg.d, .30),o6mg.d)-ExpESTx
  for (i in 1:999){
     o4o6.mg.30 <- c(o4o6.mg.30,cohens.d(my.resample(o4mg.d, o6mg.d, .30),o6mg.d)-ExpESTx)
   }

o4o6.mg.40 <- cohens.d(my.resample(o4mg.d, o6mg.d, .40),o6mg.d)-ExpESTx
for (i in 1:999){
   o4o6.mg.40 <- c(o4o6.mg.40,cohens.d(my.resample(o4mg.d, o6mg.d, .40),o6mg.d)-ExpESTx)
 }

o4o6.mg.50 <- cohens.d(my.resample(o4mg.d, o6mg.d, .50),o6mg.d)-ExpESTx
for (i in 1:999){
   o4o6.mg.50 <- c(o4o6.mg.50,cohens.d(my.resample(o4mg.d, o6mg.d, .50),o6mg.d)-ExpESTx)
 }


o4o6.mg.60 <- cohens.d(my.resample(o4mg.d, o6mg.d, .60),o6mg.d)-ExpESTx
```

```
for (i in 1:999){
   o4o6.mg.60 <- c(o4o6.mg.60,cohens.d(my.resample(o4mg.d, o6mg.d, .60),o6mg.d)-ExpESTx)
 }

o4o6.mg.70 <- cohens.d(my.resample(o4mg.d, o6mg.d, .70),o6mg.d)-ExpESTx
for (i in 1:999){
   o4o6.mg.70 <- c(o4o6.mg.70,cohens.d(my.resample(o4mg.d, o6mg.d, .70),o6mg.d)-ExpESTx)
 }

o4o6.mg.80 <- cohens.d(my.resample(o4mg.d, o6mg.d, .80),o6mg.d)-ExpESTx
for (i in 1:999){
   o4o6.mg.80 <- c(o4o6.mg.80,cohens.d(my.resample(o4mg.d, o6mg.d, .80),o6mg.d)-ExpESTx)
 }

o4o6.mg.90 <- cohens.d(my.resample(o4mg.d, o6mg.d, .90),o6mg.d)-ExpESTx
for (i in 1:999){
   o4o6.mg.90 <- c(o4o6.mg.90,cohens.d(my.resample(o4mg.d, o6mg.d, .90),o6mg.d)-ExpESTx)
 }


o4o6.mg.100 <- cohens.d(my.resample(o4mg.d, o6mg.d, 1.0),o6mg.d)-ExpESTx
for (i in 1:999){
   o4o6.mg.100 <- c(o4o6.mg.100,cohens.d(my.resample(o4mg.d, o6mg.d, 1.0),o6mg.d)-ExpESTx)
 }

o4o6.mg <- c(o4o6.mg.0, o4o6.mg.10, o4o6.mg.20, o4o6.mg.30, o4o6.mg.40,
               o4o6.mg.50, o4o6.mg.60, o4o6.mg.70, o4o6.mg.80, o4o6.mg.90, o4o6.mg.100)
hist(o4o6.mg, main = "Minimum Gender Posttest", xlab = "Effect Size")
summary(o4o6.mg)
describe(o4o6.mg)


o4o6.mg.df <- data.frame(o4o6.mg.0, o4o6.mg.10, o4o6.mg.20, o4o6.mg.30, o4o6.mg.40,
                  o4o6.mg.50, o4o6.mg.60, o4o6.mg.70, o4o6.mg.80, o4o6.mg.90, o4o6.mg.100)
```

```
 str(o4o6.mg.df)
o4o6.mg.desc <- describe(o4o6.mg.df, skew=FALSE, ranges=FALSE)
xtable(o4o6.mg.desc,
      caption="Resampled Minimum Selection by Treatment Effects Descriptive Statistics",
       label="o4o6.mg")




## MAXimum E.S. GENDER SELECTION BIAS INTERACTIONS
o2o8.mg.0 <- cohens.d(my.resample(o2mg.d, o8mg.d,0),o8mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.0 <- c(o2o8.mg.0,cohens.d(my.resample(o2mg.d, o8mg.d,0),o8mg.d)-ExpESTx)
 }


o2o8.mg.10 <- cohens.d(my.resample(o2mg.d, o8mg.d, .10),o8mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.10 <- c(o2o8.mg.10,cohens.d(my.resample(o2mg.d, o8mg.d, .10),o8mg.d)-ExpESTx)
 }


o2o8.mg.20 <- cohens.d(my.resample(o2mg.d, o8mg.d, .20),o8mg.d)-ExpESTx
for (i in 1:999){
    o2o8.mg.20 <- c(o2o8.mg.20,cohens.d(my.resample(o2mg.d, o8mg.d, .20),o8mg.d)-ExpESTx)
  }

o2o8.mg.30 <- cohens.d(my.resample(o2mg.d, o8mg.d, .30),o8mg.d)-ExpESTx
  for (i in 1:999){
     o2o8.mg.30 <- c(o2o8.mg.30,cohens.d(my.resample(o2mg.d, o8mg.d, .30),o8mg.d)-ExpESTx)
   }

o2o8.mg.40 <- cohens.d(my.resample(o2mg.d, o8mg.d, .40),o8mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.40 <- c(o2o8.mg.40,cohens.d(my.resample(o2mg.d, o8mg.d, .40),o8mg.d)-ExpESTx)
```

```
 }

o2o8.mg.50 <- cohens.d(my.resample(o2mg.d, o8mg.d, .50),o2mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.50 <- c(o2o8.mg.50,cohens.d(my.resample(o2mg.d, o8mg.d, .50),o8mg.d)-ExpESTx)
 }


o2o8.mg.60 <- cohens.d(my.resample(o2mg.d, o8mg.d, .60),o8mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.60 <- c(o2o8.mg.60,cohens.d(my.resample(o2mg.d, o8mg.d, .60),o8mg.d)-ExpESTx)
 }

o2o8.mg.70 <- cohens.d(my.resample(o2mg.d, o8mg.d, .70),o8mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.70 <- c(o2o8.mg.70,cohens.d(my.resample(o2mg.d, o8mg.d, .70),o8mg.d)-ExpESTx)
 }

o2o8.mg.80 <- cohens.d(my.resample(o2mg.d, o8mg.d, .80),o8mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.80 <- c(o2o8.mg.80,cohens.d(my.resample(o2mg.d, o8mg.d, .80),o8mg.d)-ExpESTx)
 }

o2o8.mg.90 <- cohens.d(my.resample(o2mg.d, o8mg.d, .90),o8mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.90 <- c(o2o8.mg.90,cohens.d(my.resample(o2mg.d, o8mg.d, .90),o8mg.d)-ExpESTx)
 }


o2o8.mg.100 <- cohens.d(my.resample(o2mg.d, o8mg.d, 1.0),o8mg.d)-ExpESTx
for (i in 1:999){
   o2o8.mg.100 <- c(o2o8.mg.100,cohens.d(my.resample(o2mg.d, o8mg.d, 1.0),o8mg.d)-ExpESTx)
 }

o2o8.mg <- c(o2o8.mg.0, o2o8.mg.10, o2o8.mg.20, o2o8.mg.30, o2o8.mg.40,
```

```
                o2o8.mg.50, o2o8.mg.60, o2o8.mg.70, o2o8.mg.80, o2o8.mg.90, o2o8.mg.100)
hist(o2o8.mg, main = "Maximum Gender Selection Bias Interaction", xlab = "Effect Size")
describe(o2o8.mg)


o2o8.mg.df <- data.frame(o2o8.mg.0, o2o8.mg.10, o2o8.mg.20, o2o8.mg.30, o2o8.mg.40,
              o2o8.mg.50, o2o8.mg.60, o2o8.mg.70, o2o8.mg.80, o2o8.mg.90, o2o8.mg.100)

 str(o2o8.mg.df)
o2o8.mg.desc <- describe(o2o8.mg.df, skew=FALSE, ranges=FALSE)
xtable(o2o8.mg.desc,
  caption="Resampled Maximum Posttest Effects Descriptive Statistics",
     label="o2o8.mg")




####Gender Interaction effect (SBI) COMBINED
gender.es.sbi <- c(o4o6.mg, o2o8.mg)
hist(gender.es.sbi, main = "Gender Posttest Effect", xlab = "Effect Size")
lines(min.ES.post.mg, 2000, type="h", col="blue")
lines(max.ES.post.mg, 2000, type="h", col="blue")


describe(gender.es.sbi)


gender.sbi <- rbind(o4o6.mg.df, o2o8.mg.df)
str(gender.sbi)
gender.sbi.desc <- describe(gender.sbi, skew=FALSE, ranges=FALSE)
xtable(gender.sbi.desc,
       caption="Resampled Selection by Treatment Effects Descriptive Statistics",
          label="gender.sbi")
```

```
ks.test(gender.es.sbi, "pnorm", mean=mean(gender.es.sbi), sd=sd(gender.es.sbi))
```

# References

# References


Bacon, F. (2005). *The novum organon or a true guide to the interpretation of nature*. Elibron Classics. (Original work published 1855.)

Benson, K., & Hartz, A. (2000). A comparison of observational studies and randomized, controlled trials. *The New England Journal Of Medicine*, *342*(25), 1878 - 1886.

Berelson, B., & Steiner, G. (1964). *Human behavior: an inventory of scientific findings*. New York: Harcourt, Brace & World.

Berk, R. A., & Sherman, L. W. (1988). Police responses to family violence incidents: An analysis of an experimental design with incomplete randomization. *Journal of the American Statistical Association*, *83*(401), 70 - 76.

Berk, R. A., Smyth, G. K., & Sherman, L. W. (1988). When random assignment fails: Some lessons from the minneapolis spouse abuse experiment. *Journal of Quantitative Criminology*, *4*(3), 209 - 223.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297-312.

Campbell, D. T. (1969). Reforms as experiments. *American Pscyhologist*, *24*(4), 409-429.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin Company.

Chakravarti, L., & Roy. (1967). *Handbook of methods of applied statistics* (Vol. 1). John Wiley and Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

College Board. (2010). *2010 college-bound seniors total group profile report* (Tech. Rep.). College Board.

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*(4), 724750.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods*, *15*(1), 56-68.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research*, *44*(6), 828-847.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053-1062.

Donley, R. D., & Ashcraft, M. H. (1992). The methodology of testing naive beliefs in the physics classroom. *Memory & Cognition*, *20*(4), 381-391.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests [Computer software manual]. Princeton, New Jersey.

Ferriter, M., & Huband, N. (2005). Does the non-randomized controlled study have a place in the systematic review? a pilot study. *Criminal Behaviour and Mental*

*Health*, *15*(2), 111-120.

Fisher, R. A. (1935). *The design of experiments* (8th e.d, 1966 ed.). Harner Publishing Company.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *The Journal of Human Resources*, *22*(2), 194 - 227.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The ANNALS of the American Academy of Political and Social Science*, *589*, 63 - 93.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Sage Publications.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153-161.

Heckman, J. J., Hotz, V. J., & Dabos, M. (1987, August). Do we needed experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review*, *11*(4), 395 - 427.

Heckman, J. J., & Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics*, *86*(1), 30-57.

Hedges, L., Olkin, I., & Statistiker, M. (1985). *Statistical methods for meta-analysis*. Academic Press New York.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods*, *1*(2), 154-169.

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

Hume, D. (1740). *A treatise of human nature*. London, J. M. Dent & Sons Ltd.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis.* Sage Publications.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, *76*(4), 604-620.

Linacre, J. M. (2011). Winsteps rasch measurement computer program (Version 3.68.2 ed.) [Computer software manual]. Beaverton, Oregon. Available from `Winsteps.com`

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181-1209.

Meehl, P. E. (1973). Why i do not attend case conferences. In (p. 225 - 302). Minneapolis: University of Minnesota Press.

Neumann, J. V., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton University Press. Available from `http://jmvidal.cse.sc.edu/library/neumann44a.pdf`

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289-337.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, *31*(4), 463-479.

*Procedures and standards handbook (version 3.0).* (2013). Washington, DC.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Records, G. W. (2013). *Guinness world records 2013.* New York: Jim Pattison Group.

Reichardt, C. S. (2000). Research design: Donald campbell's legacy. In L. Bickman

(Ed.), (p. 89-116). Sage Publications.

Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods*, *11*(1), 118.

Reichardt, C. S. (2011). Evaluating methods for estimating program effects. *American Journal of Evaluation*, *32*(2), 246 - 272.

Reichardt, C. S., & Gollob, H. F. (1989). Ruling out threats to validity. *Evaluation Review*, *13*(1), 3-17.

Rosenbaum, P. R. (2002). *Observational studies* (Second ed.). Springer.

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer.

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.

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. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral Statistics*, *2*(1), 1 - 26.

Rubin, D. B. (2008). For objective causal inference design trumps analysis. *The Annals of Applied Statistics*, *2*(3), 808 - 840.

Rubin, D. B. (2010). Reflections stimulated by the comments of shadish (2010) and west and thoemmes (2010). *Psychological Methods*, *15*(1), 38-46.

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.

Shadish, W. R., Clark, M., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334-1356.

Shadish, W. R., & Cook, L. C., Thomas D.and Leviton. (1991). *Foundations of program evaluation: theories of practice*. Sage Publications Ltd.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology*, *64*(6), 1290-1305.

Shapiro, D. A., & Shapiro, D. (1983). Comparative therapy outcome research: Methodological implications of meta-analysis. *Journal of Consulting and Clinical Psychology*, *51*(1), 42 - 53.

Sherman, L. W., & Berk, R. A. (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review*, *49*(2), 261 - 272.

Smith, B., & Sechrest, L. (1991). Treatment of aptitude x treatment interactions. *Journal Consulting and Clinical Psychology*, *59*(2), 233 -244.

Smith, M., & Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752 - 760.

Smith, M., Glass, G., & Miller, T. (1980). *The benefits of pscyhotherapy.* Baltimore: The Johns Hopkins University Press.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250-267.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, *26*(3), 455 - 477.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659 - 706.

Wright, P. G. (1928). *The tariff on animal and vegetable oils.* New York: Macmillan.

# Curriculum Vitae

Julius Alexander Najab received his Bachelor of Arts in Psychology from University of Arizona in 2002, and a Master of Arts in Psychology from George Mason University in 2007.