# Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences

Lauren Fellers

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

# ABSTRACT

## Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences

**Lauren Fellers**

Since its establishment the Institute of Education Sciences has been creating opportunities and driving standards to generate research in education that is high quality rigorous, and relevant. This dissertation is an analysis of current practices in Goal III and Goal IV studies, in order to (1) better understand of the types of schools that agree to take part in these studies, and (2) an assess how representative these schools are in comparison to important policy relevant populations. This dissertation focuses on a subset of studies that were funded from 2005-2014 by the Department of Education, IES, under the NCER grants-funding arm. Studies included were those whose interventions were aimed at elementary students across core curriculum and ELL program areas. Study schools were compared to two main populations, the U.S population of elementary schools and Title I elementary schools, as well as these populations on a state level. The B-index, proposed by Tipton (2014) was the main value of comparison used to assess the compositional similarity, or generalizability, of study schools to these identified inference populations. The findings show that across all studies included in this analysis, participating schools were representative of the U.S. population of schools, B-index = 0.9. Comparisons were also made between this collection of schools and the respective populations at the state level. Results showed that these schools were not representative of any individual states (no B-index values were greater than 0.90). Across all included studies, schools that agreed to participate were more often located in urban areas, had higher rates of FRL students, had more minority students enrolled, and had more total students, in both district and

school, than those schools in the population of U.S. schools.  It is clear that the movement of education research is to be relevant to a larger audience.  Through this study it is clear that, across studies, we are achieving some representation in IES funded studies.   However, the finer comparisons, study samples to individual state and individual studies to these populations, show limited similarity between study schools and populations of interest to policy makers using these study findings to make decisions about their schools.

## Table of Contents

## List of Figures

## List of Tables

# Acknowledgements

It takes a village to finish a Ph.D., and I am immensely grateful for everyone who has supported and encouraged me, and especially to those who leant an ear when I needed an hour of complaining about *still* being a student.

To my advisor, Beth Tipton, you are the mentor everyone else wishes they had. I could not have made it through this program or in this field without you. You have taught me how to be a better researcher, writer, teacher, and colleague. You have allowed me opportunities that I wasn't sure I could handle and because you never doubted me I am all the better for them. I know that my doctoral experience would have been less enjoyable and overall less effective had you not agreed to be my sponsor. Your advice and guidance both academically and professional have been invaluable and I will most likely owe you for the rest of time.

I would also like to thank the rest of my committee, Bryan Keller and Jessaca Spybrook, Matt Johnson, and Peter Bergman. Your feedback and comments were enormously helpful. Showing me how to better present and focus my work ultimately leading to work that I am extremely proud of. I sincerely hope that our paths continue to cross in the future. I would also like to say a huge thank you to the network of IES grantees that were willing to participate in this project. I could not have finished this work without your help.

I am extremely lucky to have a network of friends, both near and far, who were nothing short of amazing cheerleaders and supporters during this academic pursuit. There will never be enough thanks for this group of people who have listened to me during periods of excitement and inevitable bouts of frustration during this process, and knowing exactly when words or vacation were needed. Shanna, Emily, and Sarah, thank you for being exactly yourselves, best friends don't come along that often and I am happy to call you mine. Rebekah and Seth Daggett, you

have fed me, let me be part of the family, and made sure I was still alive when I was in the thick of it, for that I will always be truly, truly grateful (and so will my mom).

Most of all I would like to thank my entire family, grandparents, aunts, uncles, and cousins, there are a lot of you. Your love and support has been everything throughout this process. Mom and Dad, I cannot say enough. You have loved me and never told me I couldn’t do anything I wanted to do which has allowed me to do so many fun and great things. It is such a great pleasure to make you proud. Lindsay, you are the best sister and friend anyone can ask for. You have helped me have fun during the grad school years which is exactly how I survived the grad school years. I love you very much. Thank you, thank you, thank you.

Finally, to the guy at Hungry Ghost Coffee who caught me during preparation for my defense, I am sorry I yelled at you for asking me to explain my dissertation topic. You were a well meaning stranger and you deserved better.

# Chapter 1

## Introduction

Often policy makers and practitioners, including teachers, principals, and families, turn to established research findings to guide decisions of education practice. For instance a district is in search of a new math curriculum and might turn to studies evaluating several candidate math programs. In each case, the district official asks: will this program increase student achievement in the schools in *my* district?

Policy makers who wish to use evidence-based practices can turn to the What Works Clearinghouse (WWC) to make decisions regarding practice from research findings. As part of the Department of Education, the WWC provides systematic reviews of the research on educational interventions across 15 main areas, resulting in reports intended for use by policy makers and practitioners. For example, the superintendent of a district, we will call it District A, might search the repository of WWC studies regarding math curriculum and from those studies find support for a specific program. The superintendent might search beyond just a program that was successful, looking also for programs that have been tested in educational settings or with students similar to those in District A. For example, District A may be small and include mostly rural schools, while the research on a well-known program is focused mostly on urban schools in large districts. Regrettably, for this superintendent, information on school context has only been recently been required in WWC reports. This is the struggle for many educators and policy makers when it comes to implementing evidence based practices. Over the last year WCC has made some information available about study participants to enable school officials to select students similar to theirs, mostly on demographic variables, however this does not indicate if those students are statistically “like” population schools of interest.

This dissertation aims to address this concern of generalizability by investigating the recruitment processes and types of schools that take part in a subset of research funded by the Institute of Education Sciences (IES).  Studies funded by IES are intended to be based upon the best research designs, allowing for estimation of the causal impact of interventions. These studies are intended to provide evidence for the WWC and thus provide evidence of best practices for policy makers and practitioners in all areas of education.   By reviewing current practices in these studies, we will be able to (1) better understand issues faced in recruitment – including barriers faced and best practices, and (2) better understand the types of schools that agree to take part in these studies.  Ultimately this study will show how these schools compare to important policy relevant populations.

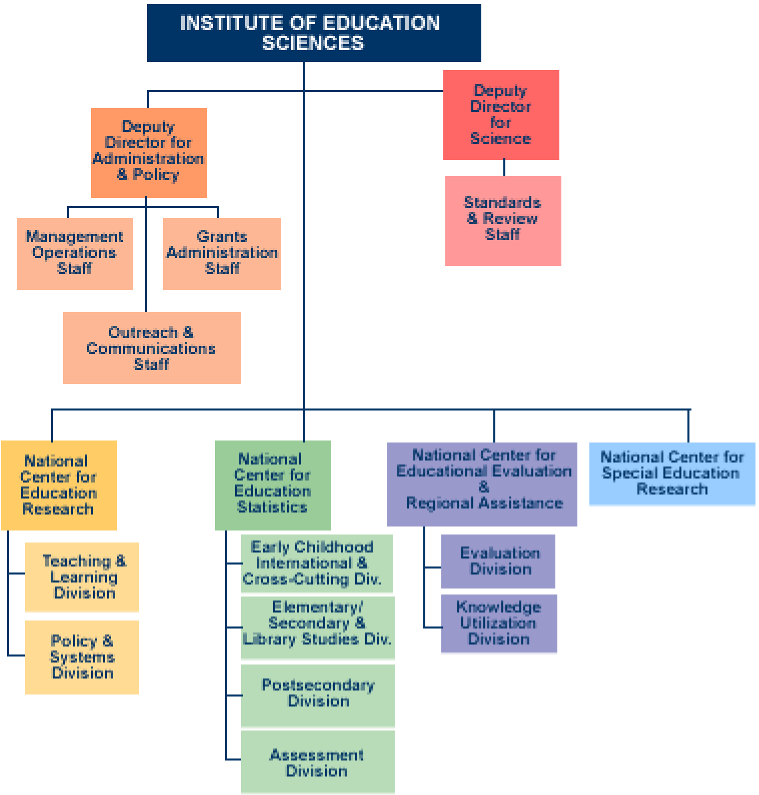
In the remainder of this chapter, I review the history of IES, its purpose, organizational structure, and research goal structure. I will then frame the issue of generalizability in four main questions asked in this study. I conclude the chapter with an overview of the dissertation.

## Institute of Education Sciences

### History

The introduction of No Child Left Behind in 2001 began a transition in the Department of Education to move toward standards and practices that required rigorous scientific inquiry to be the rule instead of the exception when informing schools what practices and programs improved student achievement.  The Education Sciences Reform Act (ESRA), passed in 2002, empowered the Department to create the Institute of Education Sciences (IES) and formally shift to the use of experimental research methods to determine effective education practices.  This created four branches within IES: the National Center for Education Research (NCER), the National Center for Education Evaluation and Regional Assistance (NCEE), the National Center for Education Statistics (NCES), and the National Center for Special Education Research (NCSER).

Figure 1.1: IES Organization Chart


INSTITUTE OF EDUCATION SCIENCES
Deputy Director for Administration & Policy
Deputy Director for Science
Standards & Review Staff
Management Operations Staff
Grants Administration Staff
Outreach & Communications Staff
National Center for Education Research
Teaching & Learning Division
Policy & Systems Division
National Center for Education Statistics
Early Childhood International & Cross-Cutting Div.
Elementary/ Secondary & Library Studies Div.
Postsecondary Division
Assessment Division
National Center for Educational Evaluation & Regional Assistance
Evaluation Division
Knowledge Utilization Division
National Center for Special Education Research

It is the mission of IES to "provide scientific evidence on which to ground education practice and policy and to share this information in formats that are useful and accessible to educators, parents, policymakers, researchers, and the public" (IES, 2015). The ESRA legislation outlines several key requirements of the functions, management, and processes of IES. Notably, it requires that all research conducted by IES must use scientifically based standards that include making claims of causality by employing random assignment experimental design whenever possible. As IES began to grant research awards from 2002 onward, these requirements could be seen in action. In a 2007 review by the Office of Management and Budget, it was reported that IES's call for high-quality, rigorous studies had transformed research within education. It created an "increased demand for scientifically based evidence of effectiveness in the field of education as a whole" (Office of Budget Management, 2013). Other government agencies like the National Science Foundation (NSF) and the education division of the National Institutes of Health (NICHD), support education related research as well with similar missions and goals, however they are beyond the feasible scope of this current work.

Created at the outset of IES, the What Works Clearinghouse (WWC) was intended to be a resource for educators and policy makers to make informed decisions about programs, practices, and curricula. The center has reviewed over 10,000 studies and created a repository of over 700 publications that provide scientific evidence of interventions that result in educational improvement for students (What Works Clearinghouse, 2015). The WWC publishes standards that are the basis for their review and acceptance into the online database. Since 2008 there have been three versions of the standards used to assess the quality of studies reviewed by WWC teams, the most recent version was released in March 2014. These standards have changed the field and reflect both IES's move to prioritize randomized control trials and the more recent move to a larger concern regarding internal validity. External validity and issues of generalizability are not mentioned as a requirement for being accepted as a study into the WWC.

### IES Goal Structure

To facilitate its mission, funding opportunities for researchers exist across all IES centers and across all topic areas including reading and writing, math and science, early childhood learning, English language learners (ELL), effective teaching/teachers, educational technology, and many others. Previously funded studies can be found on the IES website, including a brief overview of the purpose and design of each study (IES, 2015).

While grants have been awarded since 2002, the current goal structure was not introduced until 2003 for awards to be disbursed during the 2004 fiscal year. There are five main research goals that receive funding from IES. Goal I (Exploration) projects are aimed at building and informing theoretical foundations to support development of future interventions or assessment frameworks. Goal II (Development) projects focus on *developing* empirically based policies, practices, and programs. Goal III (Efficacy and Replication) projects evaluate fully developed policies, practices, and programs. Goal IV (Effectiveness; also referred to as scale-up studies) projects evaluate the impact of policies, practices, and programs at scale. Finally Goal V (Measurement) projects develop or validate measurement systems and tools. There are other funding opportunities (e.g., training, longitudinal data systems, etc.,) but most grants are awarded to studies within the goals listed above. For the purpose of this work I will focus on Goal III and Goal IV studies (details included in Chapter 3: Methods section).

Since 2002 IES has funded hundreds (956) of studies awarding millions of dollars to carry out quality studies for the improvement of student achievement under these five goals. Each year IES publishes a request for proposals (RFP) for funding across all aforementioned program areas (reading and writing, science and math, education technology, etc.) and all five goals. For each program and subsequent goal requirements and recommendations for projects are detailed for prospective grantees with the intention of aiding researchers in presenting consistent and coherent proposals. The changes seen in RFPs throughout the funding history of IES are reviewed in Chapter 2. Each government agency incurs periodic auditing and the IES is no exception. The results of the report are discussed briefly here.

### Current Progress of IES

In 2013, the Government Accountability Office (GAO) officially reviewed IES to assess its ability to support high quality research and to fulfill its mission. Findings from the review were that the Institute supports high-quality research but lacks key processes needed to fulfill other aspects of its mission. The report highlighted a specific need for more timely and relevant research. The office concluded that while steps have been taken in order to meet the needs of stakeholders, there is no formal process that allows those parties to have their feedback included in the institute's research agenda. In response to this IES hosted a meeting with 17 educational officials at the state and local levels to discuss merits of products distributed from the Regional Education Labs (REL) and the What Works Clearinghouse (WWC). The meeting was used as way for stakeholders to voice opinions about relevance, usability, and accessibility to current products from these divisions of IES.

This report documented that a practice guide, a product published by IES and included in the WWC repository for education users, regarding dropout prevention was adopted by an entire state after finding success in another area (GOA report, 2013, pg.9). This suggests that practitioners are doing just what IES and the WWC intended and using evidence to inform practices. However, the dropout study report – like many other products in the WWC – was conducted in one set of schools with very particular features and yet the results were applied broadly statewide without taking into account any information on the context of the study. This shows that district and state decision makers are utilizing available products from IES and WWC without consideration or proof that results of one study will generalize to their students and schools. Without study information regarding to whom the findings generalize, states could be implementing practices that are based on results that do not generalize to them. Without guidance for researchers on how to present to whom these findings generalize how can those practitioners hope to find studies relevant to their districts and schools?

## Statement of the Problem

This call for relevance in education research is directly linked to the issue of generalizability, which is in turn directly linked to the ability for results from studies funded by IES to be useful for people making decisions about educational practices in U.S. schools. Information regarding to whom results apply is mostly missing from reports of study findings as well as missing from the methodological standards researchers are required to uphold when proposing and completing studies. Despite this, policy makers and practitioners are called to use evidence when making decisions about best practices. This leaves a disconnect between research, which focuses only on average effects in study samples, and practice which instead asks if a program will work in *their* school or schools like theirs.

Currently there are no studies that comprehensively review how researchers are approaching generalization as part of their IES funded studies. There is also a lack of evidence for how well the current IES sponsored studies compare to the greater U.S. population of schools or to any specific sub-population of schools. A better understanding of the relationship between those schools who have participated in these studies and those who are potentially using results is necessary. Knowing the nature of schools who are likely, and conversely those who are not likely, to participate in studies could lead to better recruitment practices, better documentation of school and student contexts, and ultimately better use of study findings.

## Overview of the Study

It is the goal of this dissertation to develop an approach to investigate and assess the generalizability of findings from IES funded studies to the greater population of U.S. schools. Generalizability will be assessed using several measures to determine the compositional similarity of study samples to a specified population of interest. As an overarching goal this work aims to determine if similarity exists between *all* schools that have been included across study samples to the population of schools as a whole. Then more specifically to determine the generalizability of individual studies to specified populations of interest. Through these two goals I will develop an approach to assess generalizability of studies that can be used by researchers and practitioners.

Across its contracting and granting centers IES has funded numerous studies and evaluations over the last 12 years. To narrow the scope of this project I will only review those studies funded by grants awarded by the National Center for Education Research (NCER) arm of IES for interventions aimed at elementary school students from 2005-2014. I focus on elementary studies since they are the largest subset of funded Goal III and IV studies. I limit to studies up until 2014 since later studies have not had time to finish recruiting a sample for study participation. In Chapter 3 I explain in greater detail the exact inclusion and exclusion criteria for these studies, resulting in a final proposed subset of 25 studies.

In this study I ask four questions:

1. For this subset of studies, how similar is the composition of participating schools *overall* to those of the greater population of schools in the United States? To the population of Title I schools in the United States? And to each of these populations divided by state?  If the overall study schools are not similar to the United States or Title I schools, what is the subpopulation that is represented best by these schools in the respective populations?
2. What is the result of similarity measures across comparisons between *each* individual study's sample compared to the United States population of schools? To Title I schools in the United States? And to each of the 50 states for each of these populations?
3. What issues arose in recruitment, including barriers and strategies that may have affected the results found in Questions 1 and 2?
4. How do researchers typically report the generalizability of findings from experiments in published works (which later might inform policy)?

To answer these questions I will focus on statistical techniques and methodological innovations developed in the literature on causal generalization.  Chapter 2 provides an overview of this literature, focusing on the statistical methods and theory that will be employed in this dissertation. Also within this chapter is a review of the current emphasis IES places on generalization and external validity in its standards and funded study requirements.  In Chapter 3 I outline the methods used to answer the three questions proposed.  This includes the inclusion and exclusion criteria used to form the sample of studies used in this work, as well as the data collection procedures, statistical procedures, and estimates calculated.   Finally, Chapter 4 provides study results.

# Chapter 2: Literature Review

In this chapter I will review the broader literature on the external validity of causal effects. This includes both foundational literature related to issues of external validity and more recent literature focused on the development of statistics for making generalizations. First I will briefly review the four validity types needed for generalization of causal inferences. This review includes internal and statistical conclusion validity which has been integrated into IES policy over the last 15 years, and then external and construct validity. External validity will be reviewed in depth as it is the focus of this study.

## Overview of Validity Types

In causal inference there are four types of validity that can be addressed within a study: statistical conclusion validity, construct validity, internal validity, and external validity. Generally statistical conclusion and internal validity are regarded together as they both deal with the rigor of study design as the treatment and outcome covary. Likewise construct validity and external validity are considered together as their concern is variation across units, treatments, outcomes, or settings in an experiment.

### Statistical Conclusion and Internal Validities

Statistical conclusion validity concerns the strength of inferences made about covariation between treatment and outcome. Low statistical power, violated statistical assumptions, unreliable measures, poor implementation of treatments, and extraneous variance in experimental settings are many of the threats to statistical conclusion validity. Power analyses are now standard requirements in IES funded studies, thus avoiding problems related to statistical conclusion validity (see Spybrook 2008, Spybrook, Cullen, Lininger, 2011). These works showed a need for study designs that are methodologically sound and will lead to findings that are largely protected from violations to internal validity, which bias results. As a result of these efforts to shed light on threats to internal validity and the implications for study findings (e.g., biased estimates) IES changed standards and requirements for researchers aiming for funding showing the continual effort by the Institute to require high quality, rigorous research in education.

Internal validity concerns inferences made about the covariation between a treatment and outcome and if that observed covariance reflects a causal relationship between that treatment and outcome. This validity type is directly related to the causal connection. Cronbach (1982) noted that the aim of internal validity was not replication, and other work has shown that neither is it to make inferences to a population (Kleinbaum, Kupper and Morgenstern, 1982). The main goals for achieving internal validity are showing that treatments precede outcomes, and that other explanations of causation are not available. There are many threats that need to be addressed before a study can claim internal validity. Unambiguous temporal precedence is necessary and very easily achieved in experiments as treatments are forced to precede outcomes. An experiment should aim to remove systematic differences across conditions in participants that could be responsible for the observed effect. Similarly, researchers must ensure that simultaneously occurring events and natural changes over time are not responsible for the observed effect. Attrition is a common concern among researchers and in field studies, specifically educational research. It is an issue that plagues almost all experiments. It is imperative that researchers are able to deal with attrition to safeguard against artificial causal effects. This is also an area where standards for researchers to account for attrition have shifted in previous years which is reflected in IES study requirements.

Over the past 12 years, IES has made considerable efforts to increase the internal validity of evaluations of educational interventions. It has seen to this goal by first shifting standards to prioritize randomized control trials, and then further focusing on the internal validity and statistical conclusion validity of those trials in order to make sound causal claims.  We see this shift in requirements for proposals submitted for IES funding under Goal III and Goal IV grants (IES, 2015).  Requests for Proposals from IES began to include requirements of power analyses so as to increase the validity of statistical conclusions. Standards published by the What Works Clearinghouse in 2008, and in more detail in subsequent revisions, reflected this shift in its requirement that a study show evidence of assessing attrition and its bias in study reports, as this is a major threat to internal validity (What Works Clearinghouse, 2014).

As mentioned previously, IES publishes an annual call for proposals to conduct research across all five goals and across all program/subject areas. Requests are edited annually to reflect changes in standards set by the field of rigorous scientific research, and are intended to address the needs of stakeholders including policy makers and practitioners (IES, 2015). In its efforts to maintain rigor, IES refocused standards and placed a high premium on internal validity. This could be seen in the changes in RFPs published in 2007, and changes in proposals received from 2008-2009 forward.  In 2008 (for grants to be awarded during the 2009 fiscal year) the RFP noted in its requirements and recommendations that power analyses should be conducted. These changes reflected a need to ensure that appropriate null hypotheses could be tested resulting in sounder findings of study significance.  Requiring methods that result in unbiased estimates for study findings advanced the IES mission of high quality studies that provide dependable results which practitioners can use in making decisions about programs and curricula that directly impact student achievement. While it is clear that internal validity has been introduced and accounted for within IES study proposals, other validity types are still missing.

### Construct and External Validities

Construct validity is “the validity of inferences about the higher order constructs that represent sampling particulars” (Shadish, Cook, and Campbell, 2002, pg.65). There is a two-fold problem that arises with construct validity: first, how do we understand constructs, and second, how do we assess them? This type of validity is achieved by having clear explanations of constructs for persons, settings, treatments, and outcomes of interest; choosing carefully the instances that will match those constructs; assessing the match between the construct and the instances; and maintaining an iterative approach for constructs.

Finally, external validity, the main focus of this dissertation, refers to the ability to infer that a causal relationship will sustain over various settings, treatments, outcomes, and persons. There are two questions that arise from this type of validity: 1) Does the causal relationship hold across varied persons, settings, outcomes and treatments for those who were in the experiment and, 2) Do they hold for those persons, settings, treatments, and outcomes who were *not* in the experiment? Shadish, Cook, and Campbell conceptually equate the extent of generalizing causal relationships to statistical interactions (2002). They note that if an interaction is present between a treatment and a second variable then we are unable to say that the relationship holds across groups.

Threats to external validity are interactions of the causal relationship with units in the study, which can occur where there is an effect that is seen when some units are studied but that would not be seen if others had been studied.  Interactions of the causal relationship across treatment variations threaten generalization. Namely when effects are found with one treatment variation but not another, when treatments are combined, or when only partial treatments are implemented. When an effect is seen with one outcome observation but might not be found with other observation outcomes there is an interaction with the causal relationship and could be threatening the external validity. This also holds true for settings. Finally, when mediating variables are context dependent we have a threat to generalization (Shadish, Cook, and Campbell, 2002).

In most recent RFP versions for Goal III and Goal IV proposals researchers are required, especially for Goal IV, to include information and plans that would address generalizability (Goal III, Goal IV RFP, IES 2015). The WWC in its 2008 standards include descriptions of variations of people and settings as necessary for studies wishing to be included in the online repository of studies; however, this information is lacking from later versions of the handbook except as an undefined part of validity needs for reviewed studies. Within the handbook methods and detailed procedures are available to help researchers and reviewers address internal and statistical conclusion validity issues such as power analyses and attrition but external validity is included as an aggregate of validity in total, and no details about methods to assess it or what evidence of assessment of study generalization should be provided to readers is available from the WWC (2015).

### *External validity and RFPs*

Very recently education research as a field has made a slow turn toward external validity and its role in large scale studies. Methodologists have proposed various methods, across several other fields of study, to address issues of generalization; these innovations will be discussed later in this chapter.  Even within IES there has been a small shift from focusing solely on internal validity to a concern that covers external validity as well. In both Goal III and Goal IV studies, requirements and recommendations regarding generalization are mentioned from 2008 forward. Recall that Goal III efficacy trails aim for evaluation of interventions under "ideal, routine conditions…in authentic education settings". Goal IV effectiveness trials require the same conditions relative to a counterfactual, the major difference being that in the discussion of the context for the study, effectiveness trials are asked to describe "the heterogeneity of the sample in comparison with that of the target population".  This differs from efficacy trials where the description of the ideal or routine conditions is all that is needed (Goal III and IV RFP, IES 2015).

While it is specified in Goal IV studies, RFP details only note for Goal III efficacy trials that "results from efficacy projects have less generalizability than effectiveness projects." It should also be noted that in all years from 2009 forward a direct statement has been made that researchers should aim to meet the standards set forth by the What Works Clearinghouse (WWC).  In Goal IV recommendations, the RFP says that "scale-up evaluations require sufficient diversity in the sample of schools, classrooms, or students to ensure appropriate generalizability" (Goal IV RFP, IES 2015).  However, standards or methods of achieving generalization are not mentioned in any WWC standards from 2009 to the current edition, only that it is part of the total validity of a study.

Awards to be granted for Goal IV, effectiveness projects, in the 2011 fiscal year had a new addition to proposal requirements.  In this call for proposals researchers were asked to "detail conditions under which the intervention will be implemented".  This however focused on a need for ensuring the fidelity of implementation and did not mention anything specific to sampling procedures as they related to schools, classrooms, or students to be recruited for study participation.

The current RFP, for awards granted in 2016, states that Goal IV studies should focus on "interventions with prior evidence…to determine…beneficial impacts on student outcomes…under routine conditions." Routine conditions are characterized as those in which an intervention reflects 1) everyday practices occurring in classrooms, schools, and districts; and 2) heterogeneity of the target population. Researchers are required to include proposed routine conditions for the study as well as a proposed sample. An important change in the most recent call for proposals is asking investigators to identify their target population and include it as part of their Theory of Change. This new requirement is the first time that researchers have explicitly been asked to thoughtfully define their population of interest. This is an imperative step in planning for generalization, an admitted main goal of efficacy and effectiveness studies. IES also discusses the dissemination of results as part of the proposal process: "Effectiveness projects are to causally evaluate the impact of intervention on student outcomes. The IES considers all types of findings from these projects to be potentially useful to researchers, policy makers, and practitioners" (Goal IV RFP, IES 2015)

## In Depth Exploration of External Validity

### Shadish, Cook, Campbell, and Cronbach

Since the inception of IES the gold standard in educational research study design has been the randomized control trial (RCT). When it is feasible and ethical to complete an RCT it is the most effective way to draw a causal connection between treatment and outcome. While it is still the most desirable method of assessing interventions, a major criticism is the localized definition of study findings (Shadish, Cook, & Campbell, 2002). As noted earlier, one goal of causal inference is external validity, namely the generalization of study findings. Cook (1993) stated that generalization concerns identifying the range of possible applications of a causal connection that has been found for a specific treatment and outcomes for a specific sample of persons or settings. Cook also identifies the logical argument as an approach to address external validity, this is seen in many researcher descriptions of "diverse samples" so as to extend the generalization of study findings. Cronbach (1982) presents the idea of UTOS, which are the various Units, Treatments, Outcomes, and Settings that we hope to generalize to, and that when these entities are specified the researcher is stating what they intend to discover through their study. Cronbach also notes that when units are defined the researcher also defines settings, as these two pieces are tied together. The definition of units in these cases might then influence the settings that are of interest. Within education, researchers might want to focus their intervention on a specific type of student thereby limiting the settings that are appropriate in answering research questions. This work will focus on generalization of units, as samples and populations defined for studies, because the mission of IES and education in general prioritizes student success, so those units and their aggregate settings (schools, and districts) are the center of this study.

Cronbach, cited by Cook (1993), posits that a major problem for generalization is how to do so for a specific sample of instances to a specific population. By framing generalization as a link between a sample and population, it can be suggested that formal, well known sampling procedures are the ideal process to achieve representativeness. There are many statisticians that argue the only way to achieve a sample representative of some target population is to conduct a two-step process: simple random sampling followed by random assignment. To achieve this researchers need a well-defined population. They need the ability to conduct a simple random sample or a cluster random sample. But within the fiscal and time constraints of most educational studies, this is often not feasible. For example, Olsen, Orr, Bell, and Stuart (2013) found that in reviewing the Digest of Social Experiments, only 7 out of 273 studies had included this two-step process. Some statisticians would argue that the only method to ensure external validity of a study is to complete a random sample of the population in order to achieve representativeness. However, "even when the process is feasible, random sampling allows unbiased estimation only for this one well-defined population and does not solve the problem of generalizing to a new or different population" (Tipton, 2013, pg. 3).

Another issue in causal generalization, as introduced by Cronbach, is a lack of transfer of a causal relationship to a different population than those sampled. Different from causal connections, this causal bridge informs us how or why a causal relationship occurs, often used as the means for which a causal relationship can be transferred to varied contexts. Causal explanation is a prized finding in science, but is rarely attained in social science research. These issues lead researchers to ask three main questions; 1) What role does random selection realistically play, 2) What role does causal explanation realistically play, 3) Are there alternative theories that allow for causal generalization and do not rely on random sampling?

Campbell, and by extension Cook who summarized much of Campbell's work, proposed some less traditional, logical methods in order to address generalization. Sampling theory requires a clearly designated population and while creating the definitions to identify such a population might be somewhat simple, practical constraints still prevent researchers from utilizing random sampling (Cook, Campbell,1979; Cook, 1993). Units who are selected to be members of the sample do not always agree to participate in a study leading to mismatched target and achieved populations; for example, in a recent paper evaluating the selection of school districts in two Goal IV studies, Tipton et al (2016) found that over 95% of districts contacted refused to participate in the study. Attrition of study sites and participants is also an issue. These problems do not negate random sampling, but they do lessen the advantages. Random sampling is also a costly investment for most studies, and not fiscally feasible for many studies in social science, specifically in education.

Arguments about the best sampling methods often involve a question about trade-offs in validity, internal validity versus external validity. As recommended by Reicken and Boruch (1974) to address treatment related refusal correlations, which weaken internal validity, random assignment usually occurs after participants have agreed to participate, know the treatment options, and have agreed to be a study participant regardless of treatment assignment. These methods however lessen external validity because a higher number of possible study participants are likely to refuse to be in the study sample knowing their treatment condition might be less desirable. Cook concludes that while random sampling stems from some meaningful populations, constraints, likely financial, ethical, or political, limit the sample to less meaningful populations. Cook also mentions that at some points the goals of random assignment and random selection are at odds and that the "generality of a causal connection clearly supposes the primacy of identifying a causal connection or assessing its generality" (Cook, 1993 pg. 49). The reality of a study's budget, time, and implementation ability often limit random sampling to a circumscribed population. Because of these two conclusions Cook states that random sampling cannot be hailed as the gold standard for causal generalization despite common claims that it is. However, some statisticians would argue that the only method to ensure external validity of a study is to complete a random sample of the population in order to achieve representativeness.

There are five key principles that Cook uses to argue an alternative theory for approaching causal generalization. These principles are expounded upon in Shadish, Cook, and Campbell (2002). The first principle is proximal or surface similarity, the concept that there are apparent similarities to the target population. This idea was first proposed by Campbell and states that the sample should include units that embody the identified components of the target entity. This principle allows us to generalize an effect “with most confidence where treatment, setting, population, desired outcome and year are closest in some overall way to the original program treatment” (Campbell, 1986, pg. 75). Again, there is an emphasis on the requirement that a target population be well-defined, including what a prototypical unit might look like. Does this unit correspond to a unit like it in the population? There are some disadvantages to selecting a sample in this way. Namely that the overlap between sample and population is within a narrower range, and that explicit theoretical explanations are necessary for units to be identified as typical, which is difficult to accomplish when there are disagreements about the characterization of typical. While this rational approach to sample selection is a foundation for generalization it does not stand alone.

According to the second principle, generalization is shown when a source of variability is irrelevant to a generalization, i.e., the principle of heterogeneity of irrelevancies (Campbell, 1986; Cook, 1993; Shadish, Cook, & Campbell, 2002). In sampling persons or settings that are typical of a population often attributes associated with those prototypical cases will differ from one another in many ways. By identifying variations in persons, settings, treatments, and outcomes, we are able to say that the causal relationship generalizes across those variations. For example, this principle is applied when researchers attempt to include schools with a wide range of free and reduced lunch (FRL) status (e.g., 5 – 95%) instead of simply schools close to average.

The third principle for generalizing causal inferences is making discriminations. Much like discriminant validity in measurement theory the concept as it applies to generalization is that researchers are able to discriminate between versions of the causally-implicated construct hypothesizing that one version would change the causal relationship in either magnitude or direction.  Interpolation and extrapolation, the fourth principle, is the practice of generalizing by interpolating to unsampled values within a range of sampled persons or settings, and also by extrapolating beyond that range.  This strategy is most efficient when more levels are included in the sample, when the functional form is well-defined over the sample, and when the values being extrapolated are levels close to those in the original sample.

The final principle of generalizing causal inferences is that of the causal explanation. As noted earlier, a causal explanation is the ability to understand why or how an individual population of persons or settings is distinguished in a causal relationship (Cook, 1993). Specifically, for external validity this is the principle that a transfer of causal knowledge is able to identify "(1) which parts of the treatment (2) affect which parts of the outcome (3) through which causal mediating processes in order to accurately describe the components that need to be transferred to other situations" in order to replicated the effect (Shadish, Cook, Campbell, 2002, pg. 369).

## Statistics Literature

Kruskal and Mostellar (1979a, 1979b, 1979c, 1979d) published a total of four articles reviewing the definition and path to representative samples. As the articles progress the authors review the history of sampling and its use of the term representative sample and accompanying definition. They list that there are nine common ways in which the term is used. Six meanings or uses for the term were found in the scientific non-statistical literature and include the following: 1) General unjustified acclaim for the data, 2) Absence/presence of selective forces, 3) Miniature of the population, 4) Typical/ideal cases, 5) Coverage of the population, and 6) A vague term made to be precise. It is clear that in cases one and six it is a term that is applied but no evidence is given for its validity. Cases three and five are similar in that they account for population members in the sample in some way. The authors argue in all of these cases, that rarely is there any statistical procedure to support the claims. In the following sections a miniature of the population, where a sample is compositionally similar to a population is more common than other definitions when discussing representativeness. Definition five, coverage of the population, is discussed later in this work as a caveat needed for generalization. Three additional meanings of the term representative sample can be found within the statistical literature however they are focused on the ability to achieve good estimation or specific methods, as in simple random samples, or stratified probability sampling schemes. These are less frequent within the realm of education research as these are not generally considered feasible procedures due to the constraints of each study (Kruskal and Mostellar, 1979b).

## Generalization in Current Practice

As more and more large-scale studies are funded and carried out, the obvious next step is to develop an approach to address the generalizability and external validity of study findings. What are the best methods to plan for, adjust for, or assess generalizability? How do we approach generalization specifically in education research? There is often a discussion of the trade-off between internal and external validity, but both are achievable on some level in every study.  The previous sections detailed external validity, and noted the need for a shift in standards that focus on external validity and generalization in large-scale field studies in education.  Now I turn to a discussion of methods of addressing generalization of findings.  Four strategies emerge when discussing the generalizability of a study:  understanding the problem, better sampling methods, estimation, and assessment.  I will discuss sampling, and estimation in brief as they should be considered in the larger argument of generalization.  Methods of assessing generalization will be discussed in greater detail, as they are the main process of determining if current or completed studies are able to be generalized beyond study samples.

### Inference population

All of the methods discussed in this dissertation and in the study of external validity begin first and foremost with a well-defined inference population.  An underlying assumption of many of these methods is that comprehensive data exists on units in the population as well as the study sample. A lack of large-scale longitudinal data could be a reason that education has been slower to attend to issues of generalization. As a response to this and other issues regarding national data on schools, the National Center for Education Statistics created the Common Core of Data (CCD).   This is the nation’s single publicly available database for elementary and secondary public schools and districts.  It contains data from 1987 to the latest data year for any specific school, district, or state (generally the database is published one year behind the current school year).  The CCD is made up of information regarding fiscal and non-fiscal data, school/district location and type, and elements regarding policy and research (school size, racial/ethnic groups, students subscribing to special services, graduation rates, etc.).  This comprehensive national database has fueled many states to create their own respective databases that include this information and in many cases even more, including achievement data that might not be equivalent across states. These improvements in data collection and sharing allows for more accurate descriptions of samples and now inference populations. Without population data generalization is a more difficult process.

## Understanding the problem

Experimental designs, namely random treatment assignment, are vital components in making causal claims about treatment effects. However, there are many pitfalls for researchers when procedures and adjustments are misunderstood. Imai, King, and Stuart (2008) aim to clarify some of the challenges researchers face. To begin, note that all units in a sample have two potential outcomes, Y(1) – outcome if unit is in the treatment condition– and Y(0) – outcome if unit is in the control condition – however both cannot be observed. The difference between these two outcomes can be seen only by averaging across units within treatment and control groups on a specified outcome variable. This estimate is called the sample average treatment effect (SATE), and the population quantity is called the population average treatment effect (PATE). These can be defined respectively as:

$$SATE = \frac{1}{n} \sum_{i \in \{I_i = 1\}} TE_i, \qquad (1)$$

$$PATE = \frac{1}{N} \sum_{i=1}^{N} TE_i, \qquad (2)$$

where TE is the treatment effect for unit *i*, and $I_i$=1 if the unit is selected into the sample.

The goal here, and in the larger statistical inference literature, is to have as little bias in this population estimate as possible. The difference that exists between the two is called estimation error. Imai, King and Stuart (2008) decomposed this value to reveal two main terms:

Sample selection error and treatment imbalance. Treatment imbalance is simply the difference between treatment and control groups on observed and unobserved covariates. Sample selection error, which is the focus of this section, is of main concern. This error is the result of selecting units into a sample based on a set of observed covariates **X**. For example, say a study aims to determine the treatment effect of an intervention, let's call this intervention MATH-TUTORING. Schools are recruited into the study and then randomly assigned to treatment or control groups. The potential outcomes of these schools are now determined by the function of **X**. Differences between distributions of these covariates **X** in the population and sample are the sample selection error. If these distributions are identical this error does not exist. If schools recruited into the MATH-TUTORING study differ on district size, teacher/student ratio, or district poverty levels we risk sample selection error, and therefore risk a poor estimate of the PATE calculated from our sample. This has an impact for generalization to a larger population because this error might indicate that the only population that can be generalized to is one that is exactly like the sample. Some design choices – like the suggestion of Imai, King and Stuart (2008) to use random sampling with random assignment – reduce or eliminate this error, but this is rarely feasible for studies in education.

Often, what occurs in a typical study is that decisions regarding the eligibility of a unit or site are made with logical arguments. A unit is often eligible because of study criteria, but can also be due to budgetary concerns, personnel limitations, use of particular programs or curricula, and other similar factors. Because of these criteria recruitment and sampling are occurring in a purposive fashion, but the process is often fluid and informal, and without regard of external validity of study findings. In reality because of this logical method what is really being created is a convenience sample, and should not be defined as purposive. In some cases, generalization is addressed in its basest level because researchers aimed to recruit sites from varied geographic regions or from a cross section of socioeconomic groups. Generalization is also impacted by the want in most studies to complete recruitment with as little time and cost expended as possible.

Olsen, Orr, Bell, and Stuart (2013) expanded on the topic of informal recruitment processes and found that in the *Digest of Social Experiments* only 7 of the 280 studies used random sampling.  The authors propose a formal model for what has become a regularly used but rarely discussed or documented process of sampling among researchers. However, Olsen et al, (2012) show that there is a large cost to bias from choosing sites (i.e., districts) in this way.  They note that the increased bias comes from three contributing factors: external validity bias, internal validity bias, and the interaction between the two.  The authors note that most research designs aim to reduce internal validity bias to nearly zero so they focus on external validity bias. This bias can be written as:

$$Bias_X = \rho_{\Delta P}\sigma_{\Delta}cv_P \qquad (3)$$

This shows that external validity bias is dependent upon three factors; impact variance across target population sites ($\sigma_{\Delta}$), the coefficient of variation in inclusion probabilities across sites in the target population ($cv_P$), and the correlation between the two ($\rho_{\Delta P}$).

Importantly, this formula indicates that if the treatment impact is the same across all sites, the external validity bias will be zero.  If the probability of being selected into the study sample is the same for all sites, our external bias will be zero (the same as simple random samples). Finally, if there is no correlation between the probability of being included in the sample and treatment impacts across sites, the external validity bias will be zero.  If there is no external validity bias and internal validity bias has been removed due to study design decisions, then we expect our sample to result in a strong treatment effect estimate for the population. However, if we cannot say that these values are zero how do we achieve generalizability in our study?

**Sampling**

One approach for addressing generalization is through study design resulting in better sampling. This includes ideas of stratified sampling frames, a better assessment of recruitment, and development of strategies to increase responses and participation of units. In general, the goal is to divide the population into strata according to covariates that possibly influence treatment effects. Two approaches to create these strata are presented and should be selected depending on study context. One applies k-means cluster analysis (Tipton, 2014), and the other utilizes propensity scores (Tipton et al, 2014). The goal of both methods is to create strata (less than 10) based on a likely large set of covariates so that strata are as homogenous as possible, leading to a selection of sample units that are compositionally similar to the inference population.

Tipton (2014) proposes using cluster analysis to strategically stratify and select a sample to accomplish this. Instead of using an informal process to create a "bottom-up" generalization (where those possible generalizations, definitions of inference populations, covariate selection, and sampling plans are not well defined or documented and then adjusted for after sampling and analysis are complete) this work suggests a process of strategically sampling so as to make "top-down" generalizations. This approach does not require random sampling and can be applied broadly.

Similar to work done by Rosenbaum and Rubin (1983) in observational studies, in order to achieve bias robustness we need a sample that is balanced on many covariates and balanced on the higher moments of those covariates. The simplest way to accomplish this is stratified sampling using proportional allocation. A benefit of this approach is a reduction in bias by using many covariates to create strata. However, using more than one or two covariates to create strata (as is common in other literatures) becomes difficult, so to deal with this issue, cluster analysis methods are used. K-means partitioning methods allow a clustering of $k$ strata where units are then assigned to strata so that similarity is maximized. For recommendations of distance metrics see Tipton (2014).

Once the set of covariates and the distance metric have been selected the number of strata must be used in order to implement k-means clustering optimization algorithms. To determine the best value of k, variability between and within clusters is partitioned. Defined as the total variability within clusters ($\sigma_w^2$), and the variability between clusters ($\sigma_b^2$) (Tipton, 2013). A measure of between-cluster variability, the correlation ratio, $p_k = \sigma_{bk}^2/(\sigma_{wk}^2 + \sigma_{bk}^2)$, is calculated for each number of clusters $k$. As $p_k$ moves toward 1, variation is mostly between strata, indicating a balanced sample. Both statistical and practical constraints should be considered when determining the optimal number of strata. While a large number of strata might be ideal, in practice it might be difficult to sample if some strata are too small with regard to non-response issues in recruitment.

After the number of strata has been determined, units in the population are assigned to strata using proportional allocation. With a goal of selecting a balanced sample units can be ranked within each stratum according to their distance from the stratum average on covariates **X**. The ranked list can then be used to select units into the study sample. This cluster analysis approach to purposive sampling is simple and straightforward which is highly desirable as there is a need to complete recruitment and sampling in a timely fashion (Tipton, 2013). There are other ways to sample in order to plan for generalizations.

The approach proposed by Tipton et al. (2014), is similarly aimed at creating a sample that is representative, or as Kruskal and Mostellar defines the term, a miniature of the population (1979b). This approach uses propensity scores to create a sampling frame rather than cluster analysis. Using this method we see a more diverse sample than usual, and when the common support region (defined as the overlap of the sample and population on a set of covariates) includes the entire population there is a reduction in coverage errors (which occur when there are units in the population with no like counterpart in the sample and vice versa), leading to decreased standard errors. Achieving balance on a specific set of covariates is the focus of creating a study sample that is analogous to the population of interest for generalizations of study findings.

This method requires researchers to define three things: an inference population, a population of eligible units, and the sample size. The population of eligible units should be based on any inclusion criteria (including power analysis), or other study constraints (financial or practical). The sample size must be defined prior to recruitment. Now that these three parameters are defined and covariates $\mathbf{X}$ for balancing have been selected, we can use stratification to sample units from the population.

In this variation of stratified sampling strata are determined based on propensity scores. When there are a large number of covariates $\mathbf{X}$ stratification becomes difficult and the resulting large number of strata can lead some to be empty. The solution to this issue is to reduce the dimensionality by calculating a propensity score. Propensity scores, which are often utilized in observational studies, are used to improve balance between two groups, here the sample and population. A more in depth discussion of propensity scores, their properties and assumptions is included later in this section. The main idea, however, is similar to that of Tipton (2014), again creating *k* strata based upon the population and proportionally allocating the total sample to these strata. Now, however, not all sites in the population are eligible for recruitment; this can result in some strata with very large population proportions but very few eligible sites for inclusion into the study.

Tipton (2014) and Tipton et al. (2014) provide many benefits to researchers who wish to make generalizations to an inference population. Recruitment can be targeted for the unique units in each of the strata and resources can be allocated appropriately to strata that require more or less of the recruitment efforts (materials, time, or incentives). The primary aim of these two approaches is to help researchers plan for generalization through better sampling. These methods can also be used in conjunction with estimation and assessment methods to see how well a study's findings will generalize to the intended inference population.

## Sampling Propensity Score

In the previous section, I focused on methods for improved estimation of the PATE based on planning for generalization through design. Another approach, however, focuses instead on improved estimation of the PATE using post-hoc statistical adjustments. The first step, just as in the design-oriented approach, is to define an inference population (e.g., all elementary schools in Texas). Second, locate all schools that were included in the study in the population census of data. Third, data from these study schools are compared to the population schools on a large number of covariates. Fourth, the sample is reweighted to be compositionally similar to the population. In this section, I introduce this approach and provide an overview of different estimation strategies.

In order to create a less biased estimate of the PATE, methods from other fields were adapted for these applications. Balance between a study sample and the inference population occurs across a large set of covariates which can make comparison very difficult. Propensity scores were developed for use in observational studies where they are used to match treated subjects to a control group. In generalization, these methods were applied because it allows for a reduction in dimensionality of a reweighting problem to a single comparable dimension.

To see how propensity scores can be applied in generalization, begin by letting *s*(**X**) be the sampling propensity score for a school in the population, that is,

$$s(\boldsymbol{X}) = \Pr(Z = 1|\boldsymbol{X}), \qquad (4)$$

where the variable *Z* indicates if the school is in the sample (i.e., experiment) or not and **X** includes a set of covariates including features of the schools. Later I will explain the assumptions required for selecting these covariates in **X**. Importantly, Rosenbaum and Rubin (1983) show that propensity scores are *balancing scores*, meaning that for units with the same value of *s*(**X**), the distribution of **X** is identical. The result is that when the propensity score is correctly specified, matching units based on their propensity scores is equivalent to matching on all covariates at once.

In practice, we do not know the true propensity score for any of the units in the population. Instead they must be estimated. There are several approaches to estimation, but the simplest approach is to use logistic regression. The model for this regression is;

$$\ln\left\{\frac{s(\mathbf{X})}{(1-s(\mathbf{X}))}\right\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \beta_0 + \mathbf{X\beta}' \qquad (5)$$

where $\mathbf{X}$= ( $X_1$,..., $X_p$)′, the set of covariates, and the coefficients $\beta_0$, $(\beta_1, \dots, \beta_p) = \boldsymbol{\beta}$ are estimated from the pooled population and sample data, and regressed on the logit of the indicator variable (O'Muircheartaigh and Hedges, 2014). In many cases, it is easier to work with the estimated logits than the estimated propensity scores, since the logit scale is monotone and maintains the balancing property of propensity scores, it also tends to spread the distribution out, and appears closer to normally distributed.

The estimated propensity score can be used to estimate the PATE through a reweighting procedure. This estimate of the PATE is shown to be unbiased under a set of assumptions regarding the propensity score distributions and the covariates included in **X** (see Tipton, 2013). Along with a well-defined inference population there are three assumptions required for use of propensity scores in generalization.

> A1) **Stable Unit Treatment Value Assumption** – This must be met for all units in the population and study sample. SUTVA is met in two ways, first in relation to treatment assignment and second in relation to the sample selection process (see Tipton, 2013).

SUTVA means that the treatment status of any one unit does not affect the potential outcome of another unit. In our hypothetical example, the assignment of one school to the treatment, MATH-TUTORING, does not affect the outcomes of another school. As an extension for the sample this means that the selection of one school into the study sample does not affect the potential outcomes of students in other schools.

A2) **Strongly Ignorable Treatment Assignment** – Let Z=1 if a unit is in the sample, let W=1 if that unit is assigned to treatment, and let s(**X**) be the sampling propensity score. The treatment assignment is strongly ignorable if

$$[Y(0), Y(1)] \perp W|Z = 1, s(\boldsymbol{X}), and\ 0 < \Pr\,(W = 1|Z = 1, s(\boldsymbol{X}) < 1 \qquad (6)$$

This means that all units in the sample (i.e., the experiment) have a non-zero probability of being assigned to either treatment condition and the treatment assignment does not affect the potential outcome. Generally, random assignment aids in meeting this criterion. For this work, all studies included have utilized random assignment, therefore this assumption has been met.

A3) **Unconfounded sample selection** – The sampling process and unit treatment effects are conditionally independent, given the propensity score s(**X**),

$$\Delta = [Y(1) - Y(0)] \ \perp Z|\, s(\boldsymbol{X})\ and\ 0 < s(\boldsymbol{X}) \leq 1. \qquad (7)$$

The first part of this assumption states that the set of covariates selected includes all covariates that explain variation in the treatment impacts and differ between population and sample. The second part explains that every unit (i.e., school) must have a non-zero probability of being in the experiment. This requirement is often called the *common support* assumption and meeting it requires the comparison of propensity score distributions in the sample and population. It should be noted that ignorability can be effected by covariate selection. Sampling ignorability can fail when variables that could explain treatment effects, or variables that affect selection into a sample are omitted from the set of covariates **X.** It can also fail when there are members of the population with no comparison units in the study sample (s(**X**)=0). This assumption is often not met due to the population not containing comparison units in the sample, which makes finding subpopulations that have sample units to compare to an important step in meeting this assumption (Tipton, 2013). Covariates used can be limited by what is available for both the population and sample, as well as restrictions due to outcomes related to a specific study. Because of random assignment in most IES studies A2 is assumed to be met. This assumption, A3, becomes the focus for ensuring external validity. These assumptions originated with Rosenbaum and Rubin (1983) and were extended by Tipton (2013) for work in generalization.

### Estimation

Once the sampling propensity scores have been estimated, the next step is to use this score to develop an estimator of the PATE. There are several possible estimators in the wider propensity score literature, though in generalization two approaches are commonly employed: inverse probability weighting (IPW) and various forms of post-stratification.

One method of estimation is inverse probability weighting (IPW). This is a model based method use by Cole and Stuart (2010) to map study results onto a larger population. It is an extension of Horvitz-Thompson weighting used in survey sampling (Lohr, 2009). Using this weight, those units not selected to be in the study are given a zero weight. For those units that are selected to be in the study sample weights are determined by the marginal probability that they will be included in the study (numerator) and their probability of inclusion given a set of covariates **X** (denominator). This means that all study sample members are given real value positive weights. The conditional probabilities used for these weights are calculated using a logistic regression.

Through a simulation Cole and Stuart (2010) found that the weighted estimates were less biased than the traditional estimator (defined as the difference between the two estimates), and provided appropriate confidence intervals. The main argument however that often accompanies methods of this nature is that the selection model must be correctly specified for these estimators to be consistent (Waernbaum, 2010). Tipton (2013) also argues that this method does not reduce bias as well as other methods when selection probabilities are small, as they often are in generalization (since *n/N* is often small).

Another estimation strategy is post-stratification, also called subclassification. This estimator was first introduced by Rosenbaum and Rubin (1983) and adapted for use in generalization by Tipton (2013) and O'Muircheartaigh and Hedges (2014). Using this approach, the propensity score distribution in the population is used to stratify the sample so that strata contain an equal proportion of the population. So if there are five strata, each are defined so as to include 1/5th of the population. In this approach, a treatment effect and standard error is estimated for each stratum. The PATE is then estimated as a weighted average of the stratum-specific treatment effects, where each weight is determined by the population weight for each stratum (often equal).

There are several benefits to using the subclassification estimator for generalization. One is that it is easily understood and explained. Another is that it achieves the same or better bias reduction as other more complicated methods, and it performs better than methods like inverse probability weighting when selection probabilities are small (Tipton, 2013). There are some concerns with this approach. Tipton (2013) notes that when the propensity distribution for the population and study sample have limited overlap (small common support region) matching by stratification might be less effective in reducing bias.

Tipton (2013) shows that the reason the subclassification estimator may not perform well is because of limited overlap in the propensity score distributions, resulting in only a small number of strata (< 5) being possible, creating what is known as *coverage error*, or *under-coverage*, as it is known in survey sampling literature (Groves, 2010). There are two specific methods of dealing with this issue that are presented by Tipton (2013), truncation, and summarization of the coverage error, defined as

$$\theta = (N_0/N). \qquad (8)$$

where $N_0$ is the proportion of the population in the overlap-region and $N$ is the total population. This is the portion of the population that is included in the subpopulation. This is used to determine the distributional overlap between sample and population, indicating if ignorability has been met. Extending this we can also define:

$$\phi = (n_0/n), \qquad (9)$$

which is the proportion of the study sample that is represented in the population, where $n_0$ is the number of sample units in the common support region and $n$ is the total study sample size. These two quantities will be useful later when I discuss methods for assessing generalizability.

To address under-coverage, using fewer strata (less than 5 vs. 5 or more) to estimate the PATE is possible, however this may not result in as much bias reduction (50-70% vs. 90-95%, respectively). Truncating the population, based on estimated sampling propensity scores, is another approach to estimating the PATE. This often results in a less biased estimate but can lead to a population that is difficult to identify or understand. Tipton (2013) proposes two approaches to address this: 1) mapping units in the population (essentially visually listing the schools in the new inference population and those excluded), and 2) comparing means and standard deviations to describe differences in the two inference populations. Other approaches are explored in Tipton et al (2016).

Another concern regarding the effectiveness of the subclassification estimation strategy is that while the resulting estimator of the PATE typically has smaller bias, it also has larger sampling variance (as compared to the typical or naïve treatment effect estimator). Again, Cochran (1968) introduced this concept as it applies to observational studies and Tipton (2013) expands that for generalization. See proposition 2.2 of that paper for the definition of the expected variance inflation (EVIF) for the subclassification estimator (as compared to the naïve estimator). We can think of EVIF in a simple way: If the value is large there is little similarity between the sample and population.

Tipton's (2013) simulation showed that even in skewed distributions subclassification still resulted in significant bias reduction. It was again seen that the larger the EVIF the larger the differences between population and sample on the given covariates, further indicating coverage error issues.  To reduce these errors distributions of **X** need to be similar, reducing coverage errors (Tipton, 2013; O'Muircheartaigh and Hedges, 2014). Simulations showed that bias is reduced by 96% when $k$ = 5 strata are used and generalizations are restricted to an established subpopulation.  Not using a subpopulation, bias was still reduced by 57% -73% when $k$ = 5 strata are used.  Variance is larger in strata where $s(\boldsymbol{X}) = 0$.  These findings indicate a functional relationship to coverage errors (Tipton, 2013). This relationship will be further investigated when methods of assessing generalizability are discussed in the next section.

There are also several benefits to this approach, even when ignorability is not met.  The assumptions of covariates and their functional forms are easy to explain, and bias reduction for each covariate can be clearly assessed.   Also, with this estimator, sampling variance for the PATE can account for mismatches between population and study sample units, which indicates that the treatment effect for a particular inference population that has a large variance may not be useful for making decisions regarding the effect of an intervention.

## Assessment

The final area of current research and the main focus of this dissertation is the assessment of how generalizable a study's findings are to a specific population of interest. There are two approaches: one that requires the knowledge and use of control condition outcomes in the population and another that does not require outcomes at all. After introducing both approaches, I will focus on the latter because in looking across a wide range of studies it is impossible to find a common outcome measured in all study samples and inference populations. Importantly, in either approach the propensity score assumptions A1, A2, and A3 presented earlier, should be met.

### *Methods Requiring Outcomes*

Stuart et al. (2011) propose that one way to assess generalizability using propensity scores is to use the scores to *create* a control group that looks like the population. This approach requires that the same outcome score be available in the population and the sample. In their example, the authors employ these methods on an RCT aimed at determining the success of a school wide program, Positive Behavioral Instructional Supports (PBIS), a process of creating better systems to ultimately improve staff and student behaviors. The study measured statewide third grade reading and math proficiency scores and the percentage of students suspended throughout the year as outcomes.

The approach here is to use propensity score methods to reweight units (e.g., schools) in the control condition (W = 0) of the experimental sample so they are more similar to the population. Various estimation strategies can be used – including IPW and subclassification, as well as full matching (another version of subclassification). Since the true average outcome Y is available in the population, the reweighted average outcome from the sample control condition can be compared in order to assess the performance of the propensity score approach. In their example they show that these three methods performed equally well and so they focus on outcome comparisons using weights calculated from full matching, because they consider it the intermediate approach. The authors found that when using weights from propensity score full matching average outcomes for control schools in the study tracked closely to those of the true state mean (Stuart et al, 2011).

### *Methods Not Requiring Outcomes*

While the previous approach – based on outcomes – is ideal, in most study conditions it is not possible. This is because it requires that the same test or outcome be available for all units in the population and the sample, which very rarely occurs. The second approach to assessment therefore focuses only on the comparison of covariate distributions in the sample and population. Again, the results of this assessment hinge on the inclusion of the correct covariate set; thus, it is important for any analyst to clearly state the assumptions being made in the assessment analysis.

The assessment approach here essentially focuses on the development of statistics that summarize the similarity between the estimated propensity score distributions (and thus underlying covariates) in the sample and population. To date, four metrics have been introduced, though others from the general propensity score literature are also potentially useful.

***Propensity score difference.*** The first measure, introduced by Stuart et al (2011), is the propensity score difference between the study sample and target population, defined as:

$$\Delta_p = \frac{1}{n} \sum_{i \in \{S_i = 1\}} \widehat{p_\iota} - \frac{1}{N-n} \sum_{i \in \{S_i = 0\}} \widehat{p_\iota}. \qquad (10)$$

Here, $\Delta_p$ is the difference in average propensity scores – that is, average probabilities of being in the experiment – between those in the study and those who are not in the study sample. This metric is simple to calculate, but unfortunately it is difficult to judge. That is, while the values are between 0 and 1, it is hard to know when $\Delta_p$ is "big" or "small."

***Standardized mean difference.*** A second metric, therefore, is to scale this difference $\Delta_p$, turning it into a standardized mean difference; here the standardization is with respect to the standard deviation, $\sigma$, of the propensity scores in the population, i.e.,

$$\delta = \frac{\mu_s - \mu_p}{\sigma} = \frac{\Delta_p}{\sigma}. \qquad (11)$$

Stuart et al (2011) propose this as well, noting that doing so puts it on a scale that has clear rules of thumb. They borrow the rules of thumb from Rubin (2001) and argue that if this SMD is larger than 0.25, then the sample is dissimilar from the population. This is based on the rule of thumb from Rubin (2001) showing that when SMDs are greater than 0.25, regression adjustments do not perform well.

***P-value and random sampling.*** A third metric, also proposed by Stuart et al (2011) is based on statistical significance. Here they propose resampling the inference population, and in each random sample calculate the probability difference ($\Delta_p$) and the SMD. In this approach, it is possible to determine the probability that the value observed (the p-value) would occur under random sampling. This approach focuses on random error not on the size necessary for regression adjustment. In a similar line of research, Tipton, Hallberg, Hedges and Chan (in press) show that standardized mean differences and other proposed measures can be large simply by chance in small random samples (of the size typical in large-scale experiments). In these random samples, propensity score standardized mean differences larger than 0.25 occur by chance frequently.

***Generalizability index.*** A fourth metric, proposed by Tipton (2014) is closely related to the SMD, but differs in that instead of focusing on differences in the *averages* of the propensity score distributions, it focuses on the full distributional similarity between the population and sample. Tipton argues that this is important since problems of overlap – which make generalization particularly difficult – often are more important than average differences. Like the SMD, this index allows for a quantification of the degree of similarity between a sample and population, however this index not only focuses on the balance between the two groups but also on how effective methods of adjustment using propensity scores will be when estimating the population average treatment effect.

There are three quantities that effect the ability of treatment estimates to perform well. The common support region, coverage errors, and the degree of similarity within the common support region. By accounting for all three of these entities, which greatly effect bias and variance inflation in sampling propensity scores, this index provides a way for researchers to know when their estimates can be useful for treatment effects in the population. While there are other *visual* methods for assessing the similarity of distributions (i.e., Q-Q plots), this index allows for a simple and informative measure of the ability of samples to generalize to a population of interest. Because of these properties, this index provides a simple and complete determination of compositional similarity that will lead to good treatment effect estimates.

Tipton therefore defines a generalizability index based upon the Bhattacharya index, which is a measure of similarity; it is also called the “histogram distance.” The beta-index is defined as:

$$\beta = \int \sqrt{f_s(s) f_p(s)}\ ds. \qquad (12)$$

where $f_s$(s) and $f_p$(s) are the densities of the distributions of propensity scores (or their logits) in the sample and population. In a given sample, this can be estimated using

$$B = \sum_{j=1}^{k} \sqrt{w_{pj} w_{sj}} \qquad (13)$$

where for a given set of *k* strata, $w_{pj}$ and $w_{sj}$ correspond to the proportion of the population and sample, respectively, in the j’th bin. In the paper, Tipton proposes an approach to creating these strata that is similar to the approach found in the creation of bins for histograms.

The generalizability index has several beneficial properties. First, Tipton (2014) shows that the measure can be rewritten as,

$$\beta = \sqrt{\phi\theta} \int \sqrt{f_{s0}(s) f_{p0}(s)}\ ds_0 = \sqrt{\phi\theta}\beta_0 \qquad (14)$$

where $\phi$ and $\theta$ are defined, as in Equations 8 and 9, to represent the amount of overlap of the population and sample distributions respectively, and $\beta_0$ is the similarity between the sample and population propensity score distributions in the region of common support.

Unlike the SMD, this measure takes into account overlap, which is important since it has the largest effect on the amount of bias reduction possible using propensity score adjustment methods. This means that when large values of this index are found propensity score adjustment methods will be successful for calculating estimates.  Additionally, Tipton (2014) shows that in the event that the propensity score distributions (or their logits) are normally distributed, then β can be written instead as,

$$\beta = \exp\left(-\frac{1}{8}\delta^2\right)\sqrt{\frac{1}{\frac{1}{2}\left(\omega + \frac{1}{\omega}\right)}} \qquad (15)$$

which is a function of two parameters: the SMD δ (as defined in equation 12) and the ratio of variances ($\omega^2 = \sigma_s^2/\sigma_p^2$).

Finally, a beneficial property of this index is that it takes values between 0 and 1, with the value of 1 indicating complete overlap in the distributions and the value of 0 indicating an empty common support set. Also with large values of this index we can be confident that the experimental sample is compositionally similar to the population allowing for treatment effect estimates that are precise and close to unbiased (Tipton, 2014).

In order to determine rules of thumb for *B*, Tipton (2014) conducts a simulation study, relating values of *B* to several measures, including: 1) values of B likely to be found under random sampling (similar to the third approach given by Stuart et al, 2011); 2) values of B related to the amount of bias-reduction possible using post-stratification estimators; and 3) values of B related to the amount of variance inflation using post-stratification estimators. Tipton argues that these second and third questions matter since they – somewhat like propensity score differences and standardized mean differences, as proposed by Stuart et al (2011), – provide information on how well statistical adjustments may work for making generalizations. Based on these simulations, Tipton creates four categories, given in the table below.

Table 2.1: B-index rules of thumb

| | | |
|---|---|---|
| **Very High** | $1.00\ B \geq 0.90$ | Sample is like a random sample of population |
| **High** | $0.90 > B \geq 0.80$ | Not like the population, but reweighting is useful |
| **Medium** | $0.80 > B \geq 0.50$ | Reweighting possible but estimator will be biased/standard errors will be largely inflated |
| **Low** | $B< 0.50$ | Results from this sample will not be useful, reweighting will not correct differences between sample and population |

Using these rules, a researcher can decide if applying a reweighting strategy will lead to accurate treatment effect estimates for the population. However, if B-index values indicate that a sample has little to no generalizability to a population the researcher can use the B-index, the rules of thumb, and study data, so that generalizability can be maximized by finding the subpopulation that is best represented by the study sample. Using a process of partitioning the data in order to find the criteria that is best matched by the sample, Tipton et al (2016) were able to discover to which subpopulation study findings would generalize to well.

Tipton et al (in press) argue that based on findings from simulations, these rules of thumbs should be adjusted based on the sample size. In educational studies sampling is often conducted at the aggregate, for instance school district or school level, creating a smaller number of sampling units than the total number of participants. The author's findings suggest that the rules of thumb for the B-index discussed above are best applied when the number of covariates in the logistic regression is less than the number of units in the sample. These rules are also best applied when *n* is small. This indicates that when samples are larger B-index values of 0.90 cannot be interpreted as truly "like" a random sample of the population because large differences are more likely to exist due to chance.

***Other metrics from propensity scores in observational studies.*** In observational studies, a variety of other measures for covariate balance are also used. Several other distance metrics have been proposed to measure distributional balance such as visual methods (i.e., Q-Q plots), while others have suggested statistical distances (i.e., Levy distance, C-statistic, etc.). Also, aggregate measures of the SMD and variance ratios for *each* covariate in **X** are also used; for example, it is common to report the average- absolute SMD across all covariates in **X** (where the focus on absolute values is because direction does not matter).

Goal III (efficacy) and Goal IV (effectiveness) studies funded through IES have a particular goal of presenting significant findings that assess the impact of a program or practice that will improve the achievement of students in the United States. Using the methods presented in this chapter this work will evaluate the role of generalizability in studies previously funded under these goals in order to answer the research questions presented earlier. In Chapter 3 I will detail the sample of studies selected, data compiled for analysis, and the statistical estimates used to assess the generalizability of study findings. After detailing the methods, I will then present the findings in Chapter 4.

# Chapter 3: Methods

The goal of this dissertation is to develop an approach to and carry out an assessment of the generalizability of findings from a subset of IES funded studies. This chapter provides an overview of the process in which generalizability will be assessed in this dissertation.

The focus of this dissertation is to answer four questions:

1. For this subset of studies, how similar is the composition of participating schools *overall* to those of the greater population of schools in the United States? To the population of Title I schools in the United States? And to each of these populations divided by state? If the overall study schools are not similar to the United States or Title I schools, what is the subpopulation that is represented best by these schools in the respective populations?
2. What is the result of similarity measures across comparisons between *each* individual study's sample compared to the United States population of schools? To Title I schools in the United States? And to each of the 50 states for each of these populations?
3. What issues arose in recruitment, including barriers and strategies that may have affected the results found in Questions 1 and 2?
4. How do researchers typically report the generalizability of findings from experiments in published works (which later might inform policy)?

The main goal of Question 1 is to determine how well the schools that have taken part in the selected subset (n=25) of Goal III and IV elementary school studies funded by IES between 2005-2014 represent the population of public elementary schools currently part of the United States education system. This question also aims to describe the types of schools who are participating in studies and those who are not participating. Practitioners from schools and districts across the country are presumably the targeted users of study findings housed in the repository of research operated by the What Works Clearinghouse (WWC), thus making it important to understand to whom these results apply.

Question 1 is collective in its description of similarity between sample and population while Question 2, in contrast, focuses instead on determining which of the individual studies provide *useful* information on the treatment impact for different target populations of policy interest. This question is arguably the most important for future work.  Here useful is defined as the study sample being representative or like a miniature of the respective population, and, if not, if it is at least similar enough to the population that reweighting efforts may result in an accurate (i.e., low bias, low variance) estimate of the average treatment impact for the population.

Question 3 focuses on determining potential *causes* of these imbalances and/or unrepresentativeness and potential *solutions* for future studies. This work is qualitative and focuses on the mechanisms through which samples were targeted for and recruited into the subset studies included here.  Finally, Question 4 reviews published remarks specifically regarding generalization.  This question is also reviewed qualitatively.  Using content analysis, a systematic review of interviews (to answer Question 3) and published articles (to answer Question 4) is conducted to first determine how and why study schools are selected, and then further to determine how study samples and their comparison to populations are reported.

In this chapter, I begin by providing information on the inference populations under study – the population of public elementary schools and Title I public elementary schools, as well as a discussion of the role of states in this analysis.  I then provide a discussion of the inclusion/ exclusion criteria that led to the sample of studies included in this dissertation. Third, I discuss the data collection procedures for these studies, including the process through which data on the schools in these studies is collected, as well as information on the process of recruitment in the studies. Finally, I provide an overview of the analysis strategy used for each comparison and the statistics and results that these analyses will produce.

## Inference Populations

In order to define the inference populations – of all public elementary schools and all Title I public, elementary schools – data on schools was downloaded from the most recent Common Core of Data file (2013-2014). Each year, the federal government requires states to collect information on schools and districts using standardized measures, and then states submit these to the federal government, where they are combined (and flagged where there are potential data quality issues). Importantly, some information, such as financial data and data on small subgroups (e.g., those students enrolled in special services), is only available at the district level. Only schools listed as currently operational, and classified as regular (i.e., not alternative) are included in this comparison. Schools that were missing data on the aforementioned variables were excluded from analysis. Elementary schools, designated as primary by NCES, are defined as serving students in any grades K-5. These primary schools can include schools that have students in grades other than K-5th, but they *must* include students in at least one of these grades to be designated as Primary. Further if schools have grades other than K-5 but also include K-5, they are assigned the school level designation of their lowest grades. For example, if a school is listed as having students in K-8 grades it is given the label of primary where as a school with grades 8-12 would be labeled as middle, and a school with only 9-12 would be labeled as secondary.

It is important to know the composition of schools in the United States simply because the Department of Education and IES has tasked itself with improving the academic success of all students. The population of schools in the United States, as of the 2013-2014 school year, includes over 18,000 districts with over 96,000 operational public schools. The total number of public, non-charter, elementary schools nation-wide is 45,139 schools.

Of interest especially to IES is the improvement of schools and students with the fewest resources. A detailed picture of these schools and their characteristics that potentially make them systematically different than other schools it is necessary to separate these elementary schools from those that are not labeled as Title I schools. Title I legislation provides financial aid to schools with high percentages of students who are from low-income families. Funds are intended to be used in ways that support those students and their academic achievement. A school with at least 40% of students from low-incomes families are able to use Title I funds for school wide programs. Out of the more than 45,000 schools serving elementary school students 81% are eligible to receive Title I funds, and 65% of elementary schools use those funds for school wide programming. Table 3.1 shows the average values for elementary schools across the United States on some of the key covariates of interest. This table divides this into the two populations of interest – all schools and Title I schools.

These two national populations will be used for comparisons made when answering Question 1. Each of these populations will be further subset to only include schools in each of the 50 states. This is because questions like “Will this program work in *my* school?” are *local* questions, which have to do with the similarity of particular schools, districts, and states to those in a particular study. Thus, the question of assessment of generalizability at the federal level, comparing *all* study schools to populations of interest (Question 1) speaks to priorities for the federal government. Whereas generalizability for the study level, comparing *each* study sample to populations of interest (Question 2) speaks to priorities of local schools, districts, and states with regard to federal funding and research findings.

Table 3.1: Average values for population and Title I elementary schools

| | All Schools | Title I Schools |
|---|---|---|
| Total Students | 479 | 478 |
| % Black | 13.9% | 18.1% |
| % White | 53.5% | 44.5% |
| % Hispanic | 23.3% | 29.1% |
| % FRL | 55.8% | 70.4% |
| % ELL | 8.4% | 9.7% |
| % Urban | 27.1% | 32.1% |
| % Suburban | 36.1% | 26.2% |
| % Town/Rural | 37.5% | 41.8% |
| Total District Schools | 53 | 63 |

Note: English Language Learners (ELL), Free and Reduced Lunch (FRL), urbanicity, and race are measured as proportions of the total population

## Sample

In this section I discuss in greater detail the subset of studies included in this dissertation. As discussed earlier IES has two main funding arms, NCEE for contract work (largely won by research firms), and NCER for grants (largely won by individuals, usually associated with a university). This dissertation, for reasons of feasibility, focuses on NCER studies funded under Goals III and IV. Studies conducted through NCEE have slightly different standards as well as different funding ranges.

Within NCER, studies are separated by program area for funding purposes and include all aspects of education from subject specific curriculum interventions, to professional development, to social and emotional behavior interventions, and educational leadership or technology that support educational reform. Program areas have slightly varied proposal requirements as well as varied policy implications from findings. However, the general guidelines regarding study methods and what makes a strong proposal are the same.

Program areas of IES studies reflect the policy interests involved in many studies. Some areas are focused on teachers and other education leaders (e.g., principals, superintendents, subject coaches, Title I staff) and the interventions and policies that effect student achievement in a less direct way. Other program areas include research aimed at the improvement of technology and systems that impact student academic success. Many areas focus on the direct study of an intervention's impact on student achievement. Studies in these areas focus on curriculum and instruction for specific subjects, teacher professional development (again within subjects), as well as cognitive and social emotional interventions impacting student academic success. Researchers can submit study proposals under each of the five goals across the more than 20 program areas. All studies that have been funded are required to submit a structured abstract including the goal of the study, the proposed sample, the setting of interest (school type, leadership structure etc.), study design, analysis strategies and key measures, and details of the intervention. Table 3.2 below lists all NCER Goal III and IV studies by each of these areas.

Table 3.2: IES funded studies by program and grade level

| Program | ECE | ES | MS | HS | College | All | Other | Total |
|---|---|---|---|---|---|---|---|---|
| Cognition and Student Learning | 2 | 10 | 3 | 0 | 0 | 0 | 0 | 15 |
| Early Learning Programs and Policies | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 16 |
| Education Leadership | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 3 |
| Education Policy, Finance, and Systems | 0 | 2 | 3 | 2 | 0 | 7 | 1 | 15 |
| Education Technology | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 5 |
| Effective Teachers and Effective Teaching | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| English Language Learners | 0 | 3 | 1 | 1 | 0 | 0 | 0 | 5 |
| Improving Education Systems | 0 | 0 | 0 | 3 | 0 | 2 | 0 | 5 |
| Interventions for Struggling Adolescent and Adult Readers and Writers | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 3 |
| Mathematics and Science Education | 3 | 12 | 7 | 6 | 0 | 0 | 0 | 28 |
| Middle and High School Reform | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 3 |
| Postsecondary and Adult Education | 0 | 0 | 0 | 10 | 8 | 0 | 1 | 19 |
| Reading and Writing | 2 | 14 | 2 | 0 | 0 | 0 | 0 | 18 |
| Social and Behavioral Context for Academic Learning | 0 | 17 | 6 | 3 | 0 | 0 | 0 | 26 |
| Social and Behavioral Outcomes to Support Learning | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| Teacher Quality: Mathematics and Science | 0 | 3 | 4 | 0 | 0 | 0 | 0 | 7 |
| Teacher Quality: Reading and Writing | 1 | 4 | 2 | 1 | 0 | 0 | 0 | 8 |
| Total | 24 | 73 | 34 | 29 | 8 | 10 | 2 | |

Note: Studies that included grades from multiple grade levels are counted for both groups in this table
Key: Early Childhood Education (ECE), Elementary School (ES), Middle School (MS), High School (HS)

The focus of this dissertation is on studies conducted in elementary schools within program areas of Math, Science, Reading, Writing, and English Language Learners (ELL). ELL studies are included as it was determined they focus on interventions that directly impact student achievement in areas of reading and writing. Interventions that are aimed at teachers or administrators are not included. Furthermore, each study must have been conducted from 2005 – 2014 (to ensure that sample recruitment is complete) and take place in schools in the United States. The final proposed sample includes 25 studies. Information on these studies can be found in Table 3.3 below.

Table 3.3: Study topic by year

| | English Language Learners | Math and Science | Reading and Writing | Total |
|---|---|---|---|---|
| 2005 | - | 2 | 1 | 3 |
| 2006 | - | - | - | 0 |
| 2007 | - | 2 | 3 | 5 |
| 2008 | - | 2 | 2 | 4 |
| 2009 | - | 2 | 3 | 5 |
| 2010 | 2 | 1 | 1 | 4 |
| 2011 | 1 | 1 | 1 | 3 |
| 2012 | - | - | - | 0 |
| 2013 | - | - | - | 0 |
| 2014 | - | 1 | - | 1 |
| Total | 3 | 11 | 11 | 25 |

## Data Collection

While IES provides lists of funded studies by area, year, and focus, it does not provide lists of schools taking part in these studies or information on recruitment practices. In order to answer the questions posed in this dissertation, this information must therefore be collected. To do so, each study's principal investigator (PI) was contacted via email and phone call to collect their study information. A brief overview of the project and the goals of this work were explained, as well as the expectations of privacy for shared data. Each PI was asked to provide a complete list of schools that participated in the trial. Additionally, each PI was asked to answer several questions about the intended inference population, sampling and recruitment procedures, and issues and adjustments made for non-responses from schools or districts recruited into the study (for Question 3). These questions were asked during a phone call, or an in-person meeting with the study PI. In addition to papers shared directly by PIs, published articles were gathered from federal databases (ERIC) and from searches of the individual PI's published work.

Data is being managed on an external hard drive that is password protected and remains at one physical location.  Studies are given arbitrary study ID numbers so as to ensure anonymity, and IDs are again deidentified for reporting.  All information regarding schools that participated in studies will be kept separate from other data regarding the process of generalization.  Participants in any given study will not be identified individually or by district. Results will be aggregated so as to describe participants without revealing identifiable characteristics.  It should be noted that due to agreements with participating districts, or IES statutes this information might be more sensitive and therefore more difficult to attain.

Once identified, for each school taking part in any of the 25 studies, data on the school is located within the Common Core of Data for the recruitment year of the study. For example, if the study was funded in 2008, the school would be located in the 2007-2008 CCD year. This data will include both school-level data as well as relevant district-level variables for the district in which the school is found.

### Variable Selection for Comparison

In order to answer Question 1 and Question 2 a selection of covariates is needed.  Shadish et al (2008) and Steiner et al (2010) note that to account for selection bias in studies covariates that are highly correlated with the real selection process or the potential outcomes should be selected. These authors also note that selection of those variables with certainty is difficult.  Without empirical evidence to show which variables must be used to achieve balance or account for difference between populations and samples, it is important to review what is being done in other generalizability studies.

The covariates used in this dissertation are of interest because they have been shown to carry strong policy implications, but also because they might provide some insight as to how, until now, schools and districts have been selected to participate in large-scale research. For example, when researchers are considering both financial constraints of recruitment and the necessary number of schools that must be included in a study for strong statistical conclusion validity it might be desirable to find a district with a large number of schools. This allows for minimal cost with potential maximum benefit, but does not always provide for generalization. Although is not currently validated, the results from the initial qualitative review of interview data (used to answer Question 3) was also considered when selecting variables. The variables most frequently used to describe schools or districts targeted for recruitment as well as variables commonly used to describe eligibility criteria. These variables are also frequently seen in published articles to describe samples and populations. Variables related to district resources are less frequent in reports, however they could be strong indicators of the type of schools and districts that choose to participate in IES studies. Some of these variables are also cited in the only other large assessment of IES generalization (Stuart et al, 2016), which focuses on 11 NCEE studies.

In order to answer Questions 1 and 2, the schools in the studies will be compared to the inference populations on a large set of covariates. The variables included for comparison of study to population schools are:

- District/school size –total schools within a district
- Demographics – as measured by the number of students in each racial/ethnic group, gender group, total students participating in free and reduced lunch programs, the total number of students considered English Language Learners, these values are calculated as percentages for all analyses
- Title I – as measured by a school's Title I status, and the use of Title I funds (school wide use of those funds)

- Location – as measured by school urbanicity classification (city, suburb, town, rural).
- Resources – as measured by student to teacher ratio at the school level

The selection of covariates is limited by what is available from CCD on the population of U.S. schools, but is quite robust in its offering of variables that are available for researchers to use when discovering aspects of any one school (or the specific schools in their study) but also includes all variables used to describe study samples. The largest exception being standardized test scores which vary and often are not equivalent across states. Because of curricular differences that exist across schools, districts, and states, it is necessary to lean on variables that *are* consistent across these levels. For some studies, the goal of generalization might be to a specific state, district, or area because of these restrictions. However, for the purpose of this study, first we will look at how well studies perform to the entire population then to how well findings generalize to more specific subpopulations.

Some variables changed in name across the 10-year data span. In later data years more information is available due to an increase in the ability of states to report better data and the national call for more detailed data about schools. For this analysis, variables across all data years remain consistent. The measure of school or district urbanicity improved in later data years therefore some equating was necessary. Datasets for districts and schools were merged using unique ID variables as assigned to individual schools and districts by the U.S. Department of Education. This process is detailed in Appendix A.

The CCD requires that data reported meet a certain standard and if data is unsatisfactory, or not reported at all values are recorded as Not Available. Due to this and the need for analysis to be conducted on only complete records, population schools with missing data were removed before analysis. For schools in the study sample missing data was imputed from the closest data year possible. Two states were found to be missing data for all schools on a single variable (different variables for each state). For both states, no data was recorded for this variable among any of the records gathered from any CDD files. So that these states would not be excluded from analysis, the necessary data was imputed from each state's department of education data for the 2013-14 school year. In total, 1.1% of the total 45,624 cases were removed due to missing data leaving the final number of population schools at 45,139.

For this analysis all schools from the U.S. were retrieved from CCD database for the 2013-2014 school year. Only schools listed as operational, regular, non-charter, non-magnet, and primary (as noted above, CCD defines as having any students in K-5$^{th}$ grade) were included, all others were removed from the database. Further, schools that had fewer than 25 students or fewer than 5 teachers were also removed, as these are not typical elementary school settings. Finally, all cases with missing data on any covariates used in the propensity score logistic regression were removed. The table in Appendix A details the number of schools removed for each state during the inclusion/exclusion process.

## Methods and Statistics

### Quantitative Analysis

In order to answer Questions 1 and 2, study schools will be compared to 102 different inference populations (i.e., national and Title I, and each compared to all states). Therefore, the goal is to assess the degree of similarity between the composition of study schools (in all studies, and in individual studies) to these inference populations. To do so, I will calculate three main statistics for each analysis: 1) The average absolute standardized mean difference (ASMD) of the covariates; 2) the standardized mean difference of the logits (logit-SMD), as suggested by Stuart et al (2011); 3) The generalizability index (B-index), as suggested by Tipton (2014). The Komolgorov-Smirnov distance will also be measured as a secondary distance measure to compare the propensity score differences between the sample and population.

Answering the first part of Question 1 will result in eight total statistics, four each (the above) for the total population of schools in the United States and for the population of Title I schools. These together will be interpreted to report how representative studies funded by IES are of the schools of policy interest in the United States. Additional analyses will be then reported (e.g., from the logistic regression models) providing information on the types of schools that are *over* and *under* represented in funded research.

Answering the second part of Question 1 will require the same analyses, but repeated for each of the 50 states. This will provide information on states that have populations of schools well represented in IES funded research and those that do not. This provides information on local context and to whom results from the WWC may be most useful and where future work needs to be conducted. If the collection of studies is found not to be representative, or like a random sample, subsequent analysis will be conducted to determine which subpopulation is best represented by the schools in the study sample. This will consist of partitioning the population data on specific covariates to determine the greatest generalizability of study schools to the population. These criteria might be based on the minimum and maximum values of covariates observed in the sample, particularly when this range is much smaller than in the population (see Tipton et al, 2016 for an example).

To calculate these estimates (and means and standard deviations) all schools identified as participants across the 25 NCER studies (all who agree to share data for this study) will be located in the CCD database that includes identified relevant covariates for comparison.  If the whole sample of schools across the 25 studies is not compositionally similar to those in the inference population of all schools, I will identify specific subpopulations that are more appropriate for generalization.  This portion of the analysis will only be completed for the full sample.

For Question 2, the same analyses will be conducted as in Question 1, but separately for each of the 25 studies. To report these results, the distribution of values given will be reported for each of the inference populations with deidentified study ID numbers (e.g., scatter plots, boxplots). The goal here is to understand how much generalizability varies across studies. This is particularly important since the studies included in this dissertation cover a wide range of topics.

**Qualitative Analysis**

Unlike the first two questions, Question 3 proposes to understand the *process* of selection into studies.  While Question 3 is less formal than Questions 1 and 2, the goal is to provide feedback from researchers on what standard practice is for recruitment in Goal III and IV studies, as well as general feedback on recruitment. This is important since to date there is little information available on recruitment, yet the recruitment process has clear implications for generalization.  Question 4 aims to assess qualitatively what is being published directly about samples, how they were identified or selected, and how they compare to the population of interest.  This speaks to a large gap in both the standards required by granting bodies to address generalization concerns and also the lack of most researchers discussing it in formal or statistical ways.

To complete this portion of the analysis I will conduct a content analysis to qualitatively review interviews and published articles regarding the studies. This method, as stated in Kripendorff (2004) is a method of coding qualitative data into predefined categories in order to discern patterns and themes that emerge in the presentation of the data. This method aims to review these data systematically and objectively, and should reveal if plans for generalization, mentions of recruitment/incentive practices, or inference population goals. This analysis is completed in two phases, first for Question 3 and then for Question 4. First interviews are systematically reviewed for themes regarding the process of planning for generalization and recruitment. Second published articles are systematically reviewed to discover how generalization is reported (i.e., discussion of sample and inference populations, variables used for sample selection criteria, etc.). The results for Question 3 and 4, along with codebooks used for analysis are reported separately as the nuances within the method vary slightly. Those differences are highlighted here and again in the next chapter.

To answer Question 3, an informal interview was conducted with each PI to collect the data for analysis. During the interview, PIs was asked about why they targeted districts and schools in their study, about how recruitment went, including problems encountered, types of incentives that were used, and strategies that they found effective. Interviews are not transcribed; however detailed notes were taken during each one. The decision to not transcribe was a result of the small sample size, and with anonymity in mind each interview was not recorded or quoted directly, notes were used to complete the analysis. Each interview was coded as one unit of analysis. Because of the small number of studies within each program area (math and science, reading and writing, and English language learners), all interviews and their associated articles are discussed in aggregate with no identifying characteristics divulged.

To answer Question 4, PIs also shared published papers associated with their study, these were reviewed to analyze their approach to generalization, as well as the current status of reporting information regarding recruitment, sampling procedures, and the variables most frequently reported regarding participants. Each article found in relation to these 25 studies was coded for content regarding samples, populations, generalization, or other remarks concerning these topics, 24 total articles were reviewed. It is possible for each study to have more than one article published in relation to their findings, however these multiple articles were published across different journals. No single study has more than one article in any one journal. Each article was considered one unit of analysis, and complete paragraphs were coded into themes detailed in the codebook found in the next chapter. Paragraphs were used instead of individual sentences or statements as in the interview analysis due to the nature of published articles and their content. Articles were not coded for any other content regarding study quality (e.g., statistical power). All qualitative analyses were completed using NVivo Pro 11 (2015).

Results of these four questions will allow for an analysis of the overall approach to generalization in current and previously funded IES studies. Findings of how well study findings generalize to study defined inference populations and how well the overall sample of schools across the 25 studies will inform the direction that needs to be taken in order to better account for generalization in future IES studies. In the next chapter, I provide study results for each of these questions.

# Chapter 4: Results

In the previous chapters I have outlined the methods currently used to achieve and assess generalizability, and detailed the sample and my process for assessing the generalizability of those studies currently funded by IES. In this chapter I will first describe the success of data collection and the final study sample used for analysis of the four research questions.

## Final study sample and non-response

The time frame for collecting data for this work was limited to a 12-month time span. Twenty PIs were approached to share data, and be interviewed. The remaining five were not approached due to a lack of personal connection and the time frame allotted for data collection ending prior to any communication with these five PIs. To date, one PI, responsible for two studies, was unable to share participating study school information. Three PIs were interviewed, and have agreed to share participating school information upon IRB approval. However, after repeated follow-up these PIs did not respond to sharing their study school information, thus they are not used in the quantitative analysis. Three other PIs (responsible for 4 studies) were able to share study school data, but were unavailable for interview thus these studies are excluded for qualitative analysis. The final sample used to answer Question 1 and 2 is 15 total studies. The final sample used to analyze Question 3 is 12 PIs representing 16 studies. The final study sample used to analyze Question 4 is 24 articles representing 13 studies. The details of the final samples used to answer each question are detailed within the respective sections of results.

Recall, the four research questions of interest are as follows:

1. For this subset of studies, how similar is the composition of participating schools *overall* to those of the greater population of schools in the United States? To the population of

Title I schools in the United States? And to each of these populations divided by state? If the overall study schools are not similar to the United States or Title I schools, what is the subpopulation that is represented best by these schools in the respective populations?

2. What is the result of similarity measures across comparisons between *each* individual study's sample compared to the United States population of schools? To Title I schools in the United States? And to each of the 50 states for each of these populations?
3. What issues arose in recruitment, including barriers and strategies that may have affected the results found in Questions 1 and 2?
4. How do researchers typically report the generalizability of findings from experiments in published works (which later might inform policy)?

## Question 1

To answer Question 1, I will compare those schools in the final study sample (n=571) to those schools in the total population of U.S. elementary schools in 2013-14. Then I will use the same analysis steps to compare those schools classified as using Title I funds for school wide programs to all schools in the study sample (called simply Title I for the remainder of this study). The analysis will then be repeated for each population, all U.S. schools and Title I schools, for each state. In total, there are 102 comparisons of those schools selected to be in these IES studies and the respective populations. These 571 study schools represent 15 of the proposed 25 studies. First I will detail the results for comparisons to the population of all U.S. elementary schools, followed by the results for the comparison to all U.S. elementary schools classified as Title I.

**All Schools**

The population of United States elementary schools is defined as those schools that are public, non-charter schools that are currently operational, serve students in kindergarten through sixth grade, and have more than 25 students and 5 teachers. There are 46,195 schools that meet those criteria within the 2013-2014 school year data recorded by the National Center for Educational Statistics (NCES) in the Common Core of Data (CCD). After cases with missing data were removed the final number of schools in the population of all US schools is 45,139. In total there are 571 schools that were selected to participate across the 15 different IES studies. Within these 15 studies, schools were recruited from 19 different states from various geographic regions in the U.S.

Means of all covariates used to compare study and non-study schools are presented in the table below. The absolute standardized mean differences (ASMD) were calculated for each covariate, and values greater than 0.25 are bolded in the table. As noted previously Stuart et al (2011) showed that differences greater than 0.25 indicate large dissimilarity between a sample and population. In addition to means and standard deviations, the table also includes minimums and maximums. These are included to help explain and describe under-coverage – wherein there are subsets of the population without experimental schools like them.

Table 4.1: Covariate summary table for comparison of all studies to all schools

| | **All Schools** | | | | **Study Schools** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | SD | Min | Max | Mean | SD | ASMD |
| **School Wide Title I** | 0 | 100 | 64.9 | 47.7 | 0 | 1 | 65.8 | 47.4 | 0.019 |
| **Total Students** | 26 | 3765 | 478.42 | 220.97 | 139 | 2028 | 561.06 | 232.98 | **0.374** |
| **Urban** | 0 | 1 | 26.9 | 44.3 | 0 | 1 | 46.1 | 49.8 | **0.433** |
| **Suburban** | 0 | 1 | 35.4 | 47.8 | 0 | 1 | 35.7 | 47.9 | 0.007 |
| **Town/Rural** | 0% | 1 | 37.8 | 48.5 | 0 | 1 | 18.2 | 38.6 | **0.403** |
| **% FRL** | 0% | 100% | 55.7% | 27.6% | 0% | 100% | 58.0% | 27.2% | 0.081 |
| **% White** | 0% | 100% | 53.7% | 33.1% | 0% | 100% | 35.9% | 30.1% | **0.538** |
| **% Black** | 0% | 100% | 13.9% | 22.6% | 0% | 100% | 20.4% | 26.7% | **0.291** |
| **% Hispanic** | 0% | 100% | 23.1% | 27.2% | 0% | 100% | 33.9% | 29.1% | **0.395** |
| **% Other** | 0% | 100% | 9.3% | 12.2% | 0% | 64.5% | 9.7% | 11.6% | 0.039 |
| **% Female** | 0% | 100% | 48.4% | 3.0% | 39.9% | 59.0% | 48.3% | 2.7% | 0.019 |
| **% ELL** | 0% | 90.7% | 8.3% | 10.5% | 0% | 48.5% | 12.6% | 8.1% | **0.412** |
| **Student/Teacher Ratio** | 1.48 | 78.75 | 16.78 | 4.56 | 8.4 | 26.06 | 16.32 | 3.19 | 0.101 |
| **Total District Schools** | 1 | 983 | 52.87 | 121.46 | 3 | 914 | 115.98 | 186.26 | **0.520** |
| **Logit** | -10.71 | -1.37 | -4.73 | 0.85 | -6.46 | -0.89 | -4.02 | 0.88 | **0.829** |

Note: FRL – Free and reduced price lunch; ELL – English language learners

From this table we see that there are both school and district level variables where study and population schools have an ASMD greater than 0.25, indicating that the two groups exhibit large differences.  All values are calculated as proportions, as denoted in the table below. Study schools have more students per school than those in the population.  Study schools were also more often in urban locations and less in rural areas.  Sixty-six percent of schools in these studies are using Title I funded programs school wide, which is almost identical to the population (65%). In this analysis 46% of study schools are in urban areas, with the population containing 27% in urban areas.  Those schools in study samples were also found on average to have less white students (36%) and more black students (20%) than the population (54% and 14% respectively). Study schools had a larger percentage of students who qualify for free and reduced priced lunch (FRL).  Those schools that participated across these studies also belonged to districts that were larger in size, as measured by the total number of schools in the district.  All three variables have an ASMD greater than 0.25.

Schools that are more likely *not* to participate in these IES studies are those that have fewer total students in schools, are located in rural areas, have a higher percentage of white students and a lower percentage of minority students.  These schools also have a much lower percentage of English Language learners and lower percentage of students eligible for free and reduced price lunch.  Finally, schools not participating in studies have fewer schools in the district.

The following table shows the odds ratio of coefficients for the logistic regression used to calculate propensity scores for the comparison of overall study schools to the population of U.S elementary schools.

Table 4.2: Logistic regression for PS for all schools

| Coefficient | Odds Ratio | Log Odds | SE | 2.5% | 97.5% |
|---|---|---|---|---|---|
| **Intercept** | 0.95** | -2.35 | 0.78 | 0.02 | 0.43 |
| **Schoolwide Title I** | 0.97 | -0.03 | 0.14 | 0.74 | 1.28 |
| **Total Students** | 1.00*** | 0.00 | 0.00 | 1.00 | 1.00 |
| **% FRL** | 0.98*** | -0.02 | 0.00 | 0.97 | 0.99 |
| **Urban** | 1.59*** | 0.46 | 0.14 | 1.22 | 2.08 |
| **Suburban** | 1.04 | 0.04 | 0.14 | 0.80 | 1.37 |
| **% White** | 0.99** | -0.01 | 0.00 | 0.98 | 1.00 |
| **% Black** | 1.01** | 0.01 | 0.00 | 1.00 | 1.01 |
| **% Hispanic** | 1.01* | 0.01 | 0.00 | 1.00 | 1.02 |
| **% Female** | 1.00 | -0.00 | 0.01 | 0.97 | 1.02 |
| **Student/Teacher Ratio** | 0.91*** | -0.10 | 0.01 | 0.88 | 0.93 |
| **Total District Schools** | 1.00*** | 0.00 | 0.00 | 1.00 | 1.00 |
| **% ELL** | 1.02*** | 0.02 | 0.00 | 1.01 | 1.03 |

Note: *** p<0.001; ** p<0.01; * p<0.05

Using the logits of these probability values four values were calculated to assess the generalizability of study findings using these schools: B-Index, average difference in propensity score logits ($\Delta_{p\ logit}$), average ASMDs of covariates, and Komolgorov-Smirnov distance. For the analyses, all values are calculated using the propensity score logits. These four values will be used to determine the overall similarity between these schools and the population of all US elementary schools. The following table shows the results of these four indicators.

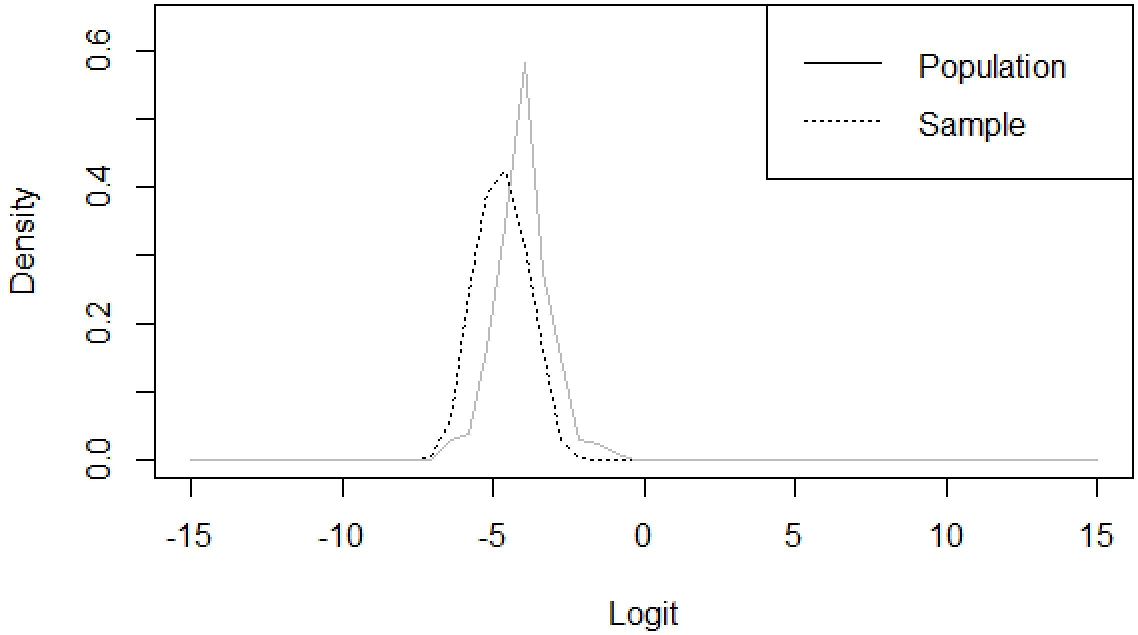
Table 4.3: Generalizability estimates of all studies to all schools

| Measure | Statistic | Rule of Thumb |
|---|---|---|
| **B-index** | 0.915 | Very High |
| **$\Delta_{p\ logit}$** | 0.829 | |
| **ASMD** | 0.291 | Greater than 0.25 |
| **KS** | 0.360*** | Significantly different at p>0.001 |

From this table, we see that according to all of these measures, the sample of schools participating in these 15 IES studies is not a representative sample of elementary schools in the United States. Recall, Tipton et al (in press) suggested that with large samples, even B-index values of 0.90 do not truly indicate representative samples. An important next step is to determine the subpopulation that *is* best represented.

The following plot is the comparison of logit densities between the population and study schools. From this it is clear that while there is a good portion of the population that is covered by schools in the studies, there are some areas where the population has schools that are not represented by those schools recruited into these 15 IES studies (for example, in the long tail). This is seen in the areas where the density for the population is not overlapped by the density of the sample. This indicates that in these areas there are members of the population with no like comparison in the study samples, meaning that estimates of the PATE would be less accurate for these members.

Figure 4.1: Comparison of logit distributions for all studies to all schools



**Title I Schools**

Continuing to answer research Question 1, the same analysis steps performed to compare study schools to all U.S. schools were applied to compare Title I schools. The population is now defined as all public (non-charter) operational elementary schools in the United States that are using Title I funds for school wide programs. Out of the 45,139 elementary schools in the United States, 29,317, use Title I funds school wide. Comparisons again were made to the 571 schools that were selected to participate in the 15 IES studies in this analysis.

The table below shows the means for covariates comparing title one schools to the study sample schools, and their absolute standardized mean differences (ASMD). Differences between the study schools and Title I schools are similar to the differences between study schools and all U.S. schools. Again, we see that both school and district level variables have an ASMD greater than 0.25, indicating that the two groups exhibit large differences. Overall study schools tended to have a smaller percentage of students who qualified for free or reduced price lunch, 58%, than the population, 71%. Study schools had higher total students per school than those in the Title I population. These schools also had a smaller percentage of white students than in the population of Title I schools. We are also able to see that schools included in the studies are more often in urban areas (50% of study schools, 32% of Title I schools) and less often in rural areas than those in the Title I population (18% of study schools, 42% of Title I schools). Looking at district variables, study schools are more often part of larger districts, as measured by total schools, total students, and average students per school than those Title I population schools. These study schools also have larger percentages of English Language Learners.

Schools that are more likely *not* to participate in these IES studies are those that again have fewer total students in schools, and have much higher percentages of students eligible for free and reduced price lunch.  Schools not participating in these studies also have a higher percentage of English language learners.  They are less frequently located in rural areas and have a higher percentage of white students.  They also have fewer total schools in the district.

Table 4.4: Covariate summary table for comparison of all studies to Title I schools

| | **Title I Schools** | | | | **Study Schools** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | SD | Min | Max | Mean | SD | ASMD |
| **Total Students** | 27 | 3765 | 476.85 | 222.19 | 139 | 2028 | 561.06 | 232.98 | **0.379** |
| **Urban** | 0% | 100% | 31.8% | 46.6% | 0% | 100% | 46.1% | 49.8% | **0.307** |
| **Suburban** | 0% | 100% | 26.0% | 43.9% | 0% | 100% | 35.7% | 47.9% | 0.222 |
| **Town/Rural** | 0% | 100% | 42.2% | 49.4% | 0% | 100% | 18.2% | 38.6% | **0.486** |
| **% FRL** | 0% | 100% | 70.7% | 19.0% | 0% | 100% | 58.0% | 27.2% | **0.669** |
| **% White** | 0% | 100% | 44.7% | 34.4% | 0% | 100% | 35.9% | 30.1% | **0.256** |
| **% Black** | 0% | 100% | 18.1% | 26.0% | 0% | 100% | 20.4% | 26.7% | 0.090 |
| **% Hispanic** | 0% | 100% | 29.0% | 30.6% | 0% | 100% | 33.9% | 29.1% | 0.161 |
| **% Other** | 0% | 100% | 8.2% | 12.1% | 0% | 64.5% | 9.7% | 11.6% | 0.128 |
| **% Female** | 0% | 100% | 48.4% | 3.0% | 39.9% | 59.0% | 48.3% | 2.7% | 0.023 |
| **% ELL** | 0% | 90.7% | 9.7% | 11.7% | 0% | 48.5% | 12.6% | 8.1% | **0.255** |
| **Student/Teacher Ratio** | 3.13 | 78.75 | 16.61 | 4.37 | 8.4 | 26.06 | 16.32 | 3.19 | 0.066 |
| **Total District Schools** | 1 | 983 | 62.45 | 140.05 | 3 | 914 | 115.98 | 186.26 | **0.382** |
| **Logit** | -9.72 | 0.77 | -4.51 | 1.01 | -6.80 | 1.32 | -3.27 | 1.26 | 1.233 |

Propensity scores were calculated using a similar logistic regression as in the comparison to all elementary schools, the main difference being the removal of Title I indicator variable. Table 4.5 shows the odds ratio of coefficients for the logistic regression used to calculate propensity scores for the comparison of overall study schools to the population of Title one elementary schools in the U.S.

Table 4.5: Logistic regression for PS for Title I schools

| **Coefficient** | **OR** | **Log Odds** | **SE** | **2.5%** | **95.7%** |
|---|---|---|---|---|---|
| **Intercept** | 1.35 | 0.30 | 083 | 0.27 | 6.68 |
| **Total Students** | 1.00*** | 0.00 | 0.00 | 1.00 | 1.00 |
| **Urban** | 2.43*** | 0.89 | 0.14 | 1.85 | 3.21 |
| **Suburban** | 1.90 | 0.64 | 0.14 | 1.45 | 2.49 |
| **% FRL** | 0.95*** | -0.05 | 0.00 | 0.95 | 0.95 |
| **% White** | 0.99*** | -0.01 | 0.00 | 0.98 | 0.99 |
| **% Black** | 1.00 | 0.00 | 0.00 | 1.00 | 1.01 |
| **% Hispanic** | 1.00 | -0.00 | 0.00 | 0.99 | 1.00 |
| **% Female** | 1.00 | -0.01 | 0.01 | 0.97 | 1.02 |
| **% ELL** | 1.03*** | 0.03 | 0.00 | 1.02 | 1.04 |
| **Student/Teacher Ratio** | 0.91*** | -0.09 | 0.01 | 0.89 | 0.93 |
| **Total District Schools** | 1.00*** | 0.00 | 0.00 | 1.00 | 1.00 |

Note: *** p<0.001; ** p<0.01; * p<0.05

The following table shows the similarity of all study schools and the population of Title I elementary schools. Results of the B-index, the Komolgorov-Smirnov test, the average difference in propensity score logits, and the average ASMD of the covariates, were measured using the logits of the logistic regression describe above.

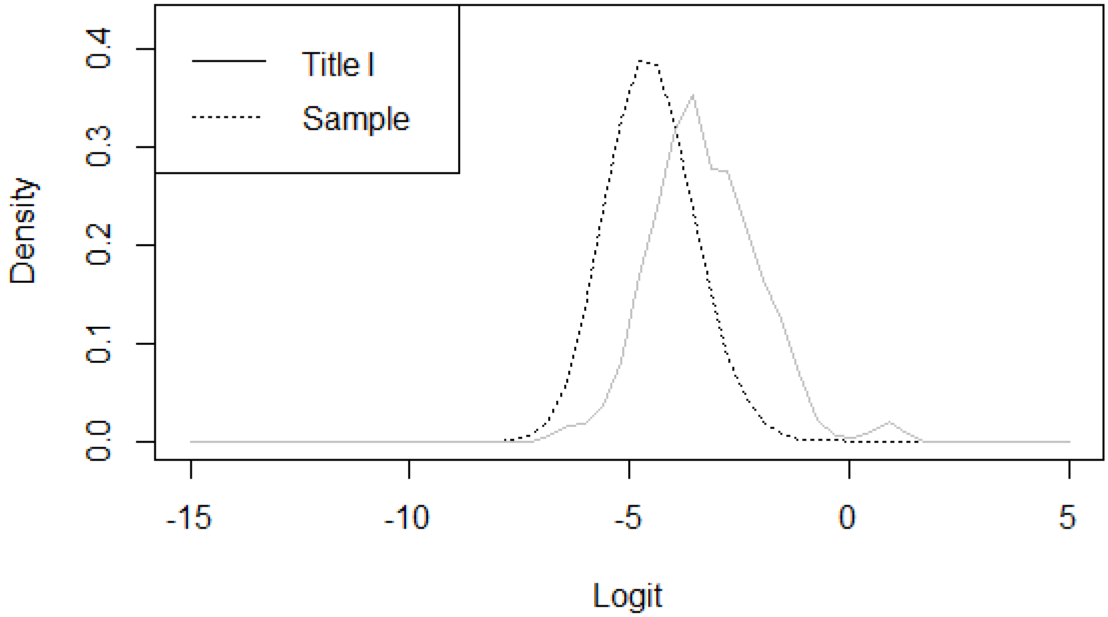
Table 4.6: Generalizability estimates of all studies to Title I schools

| **Measure** | **Statistic** | **Rule of Thumb** |
|---|---|---|
| **B-index** | 0.860 | Good |
| **$\Delta_{p\ logit}$** | 1.233 | |
| **ASMD** | 0.333 | Greater than 0.25 |
| **KS** | 0.436*** | Significantly different at p>0.001 |

As this table shows, according to all of the measures, the sample of schools in these 15 studies are not a representative sample of Title I elementary schools in the United States. An important question is if there *is* a subset of the Title I population that is well represented by the schools in IES funded studies.  This will be assessed with the full sample.

The following plot shows the distributional differences between the Title I school logits and the logits of study schools.  This plot clearly shows that there are units in the Title I population that have no like counterparts in the sample of study schools (i.e., the smaller red distribution towards the left that is not well overlapped by the green distribution. When there are members of the population with no like comparison in the study samples, estimates of the PATE would be less accurate for these members.

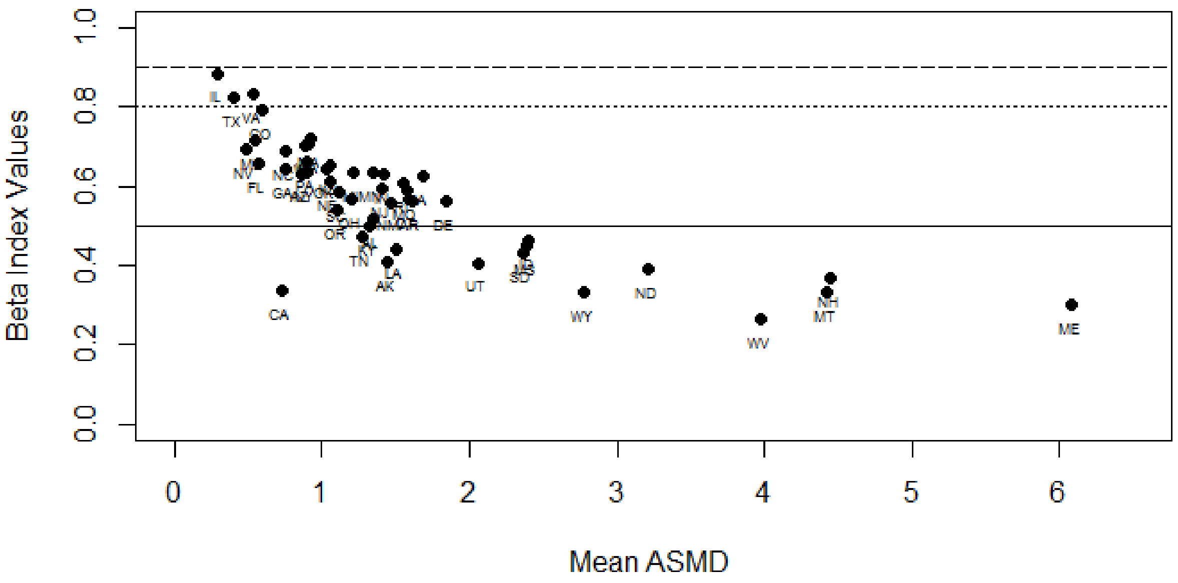
Figure 4.2: Comparison of logit distributions for all studies to Title I schools



## State Comparisons

**All schools.** In the table below, estimates are presented for the comparison of the 571 schools from the 15 studies included in this analysis to the U.S. population of all public (non-charter) elementary schools. The B-index shows that the combined study schools are not like a representative sample of any individual state (i.e., B > 0.90). However, there are three states (Illinois, Virginia, and Texas) that have values greater than 0.80. Thirty-one states have B-index values between 0.50 and 0.80. The remaining 16 states have little to no generalizability to the population of all U.S. elementary schools. Their b-index values are less than 0.50 meaning that even reweighting would not be useful for finding PATE in these states. It would be useful in these states to find a subpopulation that is better represented by these study schools. The Komolgorov-Smirnov distance between study and population propensity score logits is significant for all states, indicating that schools in the studies are significantly different than those in the populations of the respective states. The average difference in logits ranges from 1.1 to 46.7. In this figure we see that the states least represented by these 571 study schools are West Virginia, Vermont, and Hawaii. This table also shows that as B-index values get smaller the ASMD values for these states generally get larger. Tables showing the four measures of similarity for each state's population of elementary schools to the overall study schools are included in Appendix B.

Figure 4.3: Comparison of all studies to state population of all schools

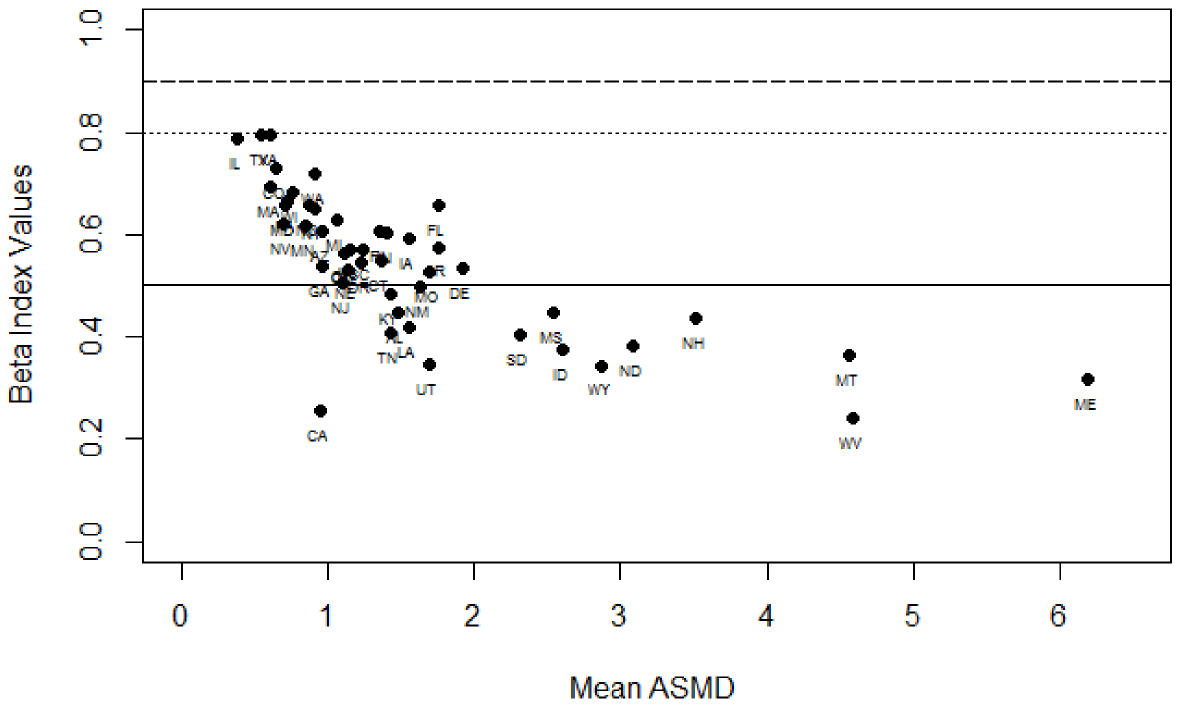


**Title I schools.** For the population of Title I schools, in the table below, estimates for the comparison of the 571 schools from the 15 studies are shown. B-index values show that study schools are not like a representative sample of any individual state. That is, no state has a B-index value greater than 0.90. Thirty-one states have a B-index value between 0.50 and 0.80. The remaining 19 states have b-index values below 0.50. It would be necessary to find a representative subpopulation in all states.

The Komolgorov-Smirnov distance between study and population propensity score logits is significant for all states, again indicating that schools in the studies are significantly different than those in the populations of the respective states, which further supports that study schools are not representative. The average difference in logits ranges from 1.6 to 28.4. Texas is the state best represented by this sample of study schools, followed by Virginia and Illinois. These are the same three states best represented by study schools in the population of all schools. Vermont and Hawaii are again two of the states least represented by this sample. Again, as B-index values decrease, ASMD values increase.

There are some sates in both populations, all schools and Title I schools, where there is little to no overlap between study schools and the respective populations. These instances will be explored further as case studies in Appendix C.

Figure 4.4: Comparison of all studies to state population of Title I schools



## Question 2

To answer Question 2, first I will compare the total population of elementary schools in the United States to the participating study schools in each of the studies included in the analysis to determine the generalizability of *each* respective study's findings.  This analysis will follow for both the population of all U.S. schools and also for the population of Title I schools.  Three of the 15 studies included are removed from the individual study analysis because they were only conducted within a single school district.  Because district level variables are utilized as part of the logistic regression to calculate propensity scores these studies do not return applicable values. The generalizability then for these three districts is such that study findings, using the logistic regression defined in Question 1, and used for comparisons in Question 2 would translate only to those in the district where studies were conducted.   Each of the remaining 12 studies are looked at individually as they compare first to the population of US public schools and then to the population of Title I schools.

The aim of Question 1 in this analysis was to determine if the overall sample of schools participating in IES studies is representative of the respective populations.  Question 2 not only answers if each study is representative of the population (U.S., Title 1, and individual states), but can also determine if reweighting the study sample will lead to useful estimates of the population average treatment effect *for each study*.  Recall the rules of thumb presented by Tipton (2014):

Table 4.7: B-index Rules of Thumb

| | | |
|---|---|---|
| **Very High** | 1.00 B ≥ 0.90 | Sample is like a random sample of population |
| **High** | 0.90 > B ≥ 0.80 | Not like the population, but reweighting is useful |
| **Medium** | 0.80 > B ≥ 0.50 | Reweighting possible but estimator will be biased/standard errors will be largely inflated |
| **Low** | B< 0.50 | Results from this sample will not be useful, reweighting will not correct differences between sample and population |

For studies that have B-index value of 0.90 or higher it can be assumed that the study sample of schools is representative of the respective population.  If they have a value of less than 0.90 that sample must either be reweighted or classified as not representative and a better represented sub-population should be found.  If studies have a value around 0.80 and higher, reweighting is possible, so that generalizations can be made from study findings to the population.  It is possible to reweight values that are 0.50 and higher, however this will often result in large variance inflation and a very small reduction in bias because of large distributional differences that are too great to be adjusted for.  Ultimately this makes estimating a population treatment effect from these non-representative samples less useful. It should be noted that three studies included schools from a single district.  In these cases, district level variables cannot be used to estimate propensity score logits and B-index values. These three studies will be removed from study by state level comparisons.

**All Schools**

The population of all schools, as defined earlier, includes all public (non-charter) elementary schools.  For this analysis each study's sample of schools is compared to this population to determine similarity.

Assessing the B-index, shown below, we see that no studies have more than medium generalizability (all values are less than 0.90). This indicates that no individual study is representative of the U.S population of schools.  The average B-index value for across these studies is 0.665. Recall the rules of thumb discussed above, nine of the 12 studies (75%) have B-index values above 0.50 meaning that these study sample schools are able to use reweighting strategies to estimate population treatment effects but would find some bias and large variance inflation. The remaining three studies (25%) are so different that even reweighting would not be useful in reducing the differences between the population and study schools. Put another way, an estimate of the PATE for all elementary schools in the United States could be calculated (using a reweighting estimator) for 25% of the studies. For these studies it would be necessary to find a subpopulation that is better represented by those study's participating schools.

A summary of all similarity measures is presented in the table below for each of the 12 studies included in this portion of the analysis. The Komolgorov-Smirnov distance is significant for all studies, meaning that the U.S. population and each respective sample exhibit large differences in the chosen covariates. The difference in average logits also support that none of the individual study's sample of schools is representative of the population of U.S. elementary schools. The average difference in the standardized mean differences between study samples and population schools is greater than 0.25 for all studies.

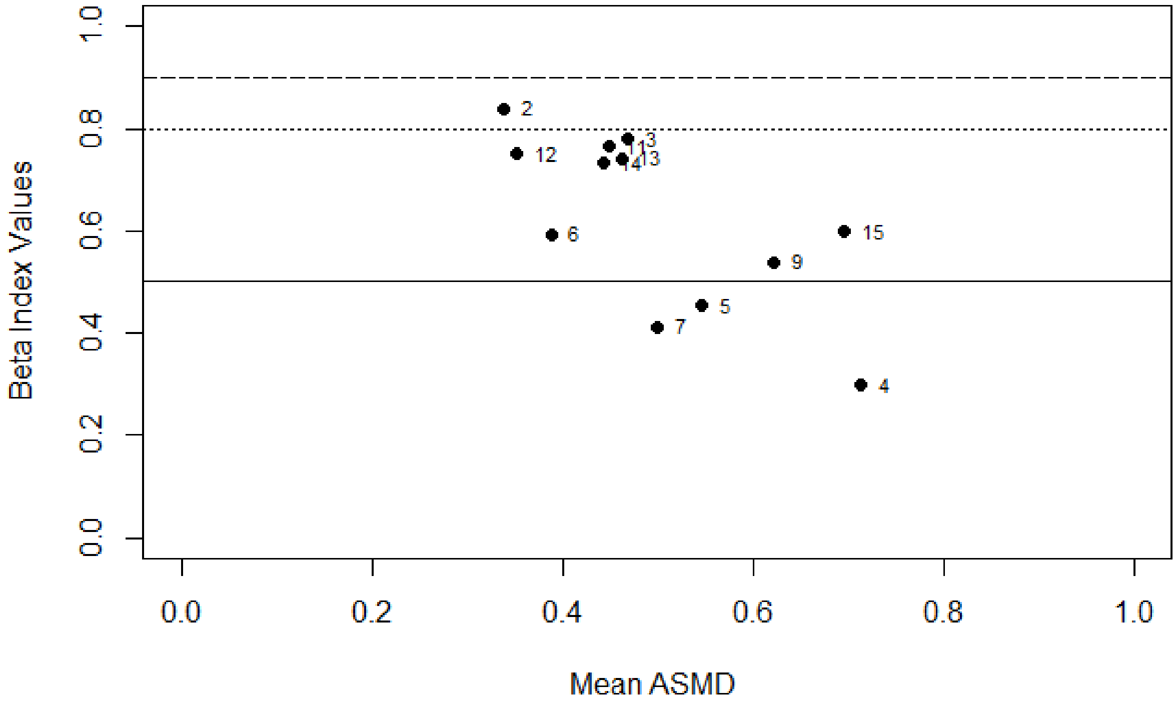
Table 4.8: Generalizability assessment for individual studies – All schools

| **Study ID** | **B-index** | **$\Delta_{p\ logit}$** | **ASMD** | **K-S test*** | **N** | **N** |
|---|---|---|---|---|---|---|
| **2** | 0.839 | 1.115 | 0.339 | 0.487 | 45139 | >100 |
| **3** | 0.780 | 1.410 | 0.469 | 0.623 | 45139 | <50 |
| **11** | 0.765 | 1.008 | 0.448 | 0.657 | 45139 | <50 |
| **12** | 0.751 | 0.792 | 0.352 | 0.566 | 45139 | <50 |
| **13** | 0.741 | 1.201 | 0.462 | 0.605 | 45139 | <75 |
| **14** | 0.732 | 1.354 | 0.443 | 0.545 | 45139 | <25 |
| **15** | 0.598 | 1.998 | 0.696 | 0.749 | 45139 | <75 |
| **6** | 0.593 | 1.316 | 0.389 | 0.718 | 45139 | <50 |
| **9** | 0.539 | 1.189 | 0.622 | 0.838 | 45139 | <50 |
| **5** | 0.453 | 0.971 | 0.547 | 0.878 | 45139 | <10 |
| **7** | 0.412 | 0.964 | 0.500 | 0.917 | 45139 | <25 |
| **4** | 0.297 | 2.077 | 0.714 | 0.916 | 45139 | <50 |

*For K-S test: all were sig. at p<0.001
Note: all study IDs, and sample sizes have been removed for confidentiality

The following figure shows a scatter plot of all 12 studies mapping the ASMD and B-index values for each study. This shows that there is a grouping of studies that while having large values of the average ASMD of covariates, they have B-index values that indicate reweighting in most cases would be useful.

Figure 4.5: Comparison of individual studies to all schools



## Title I Schools

The 12 studies included in this portion of the analysis are now compared to those schools that are eligible for Title I funds, and use those for school-wide programs. Four studies (33%) have low (B<0.50) generalizability. Reweighting for these studies would not lead to useful PATE values for this population. The average B-index value for these studies is 0.631, slightly lower than the comparison of studies to the U.S. population of all schools. As we see from the B-index values, displayed below, there are eight studies (66%) that have B-index values greater than 0.50. Two of these studies have values at 0.76, and 0.81. For these two studies, reweighting would likely be successful in creating a sample that is like the population so that estimates for PATE are reliable. The remaining six studies have larger distributional differences that, although reweighting would not remove all bias, population treatment effect estimates from these studies would be possible for the Title I population.

The Komologorov-Smirnov distance, the average difference in propensity score logits and the average standardized mean difference further support that none of the 12 studies included in this analysis are representative of the population of Title I schools. These estimates show that there are large differences between sample schools and population schools on the chosen covariates. It is necessary then to determine what subpopulations is best represented by each respective study. The summary of generalizability measures and scatter plot describing the B-index values as they relate to ASMD values for all 12 studies are presented here.

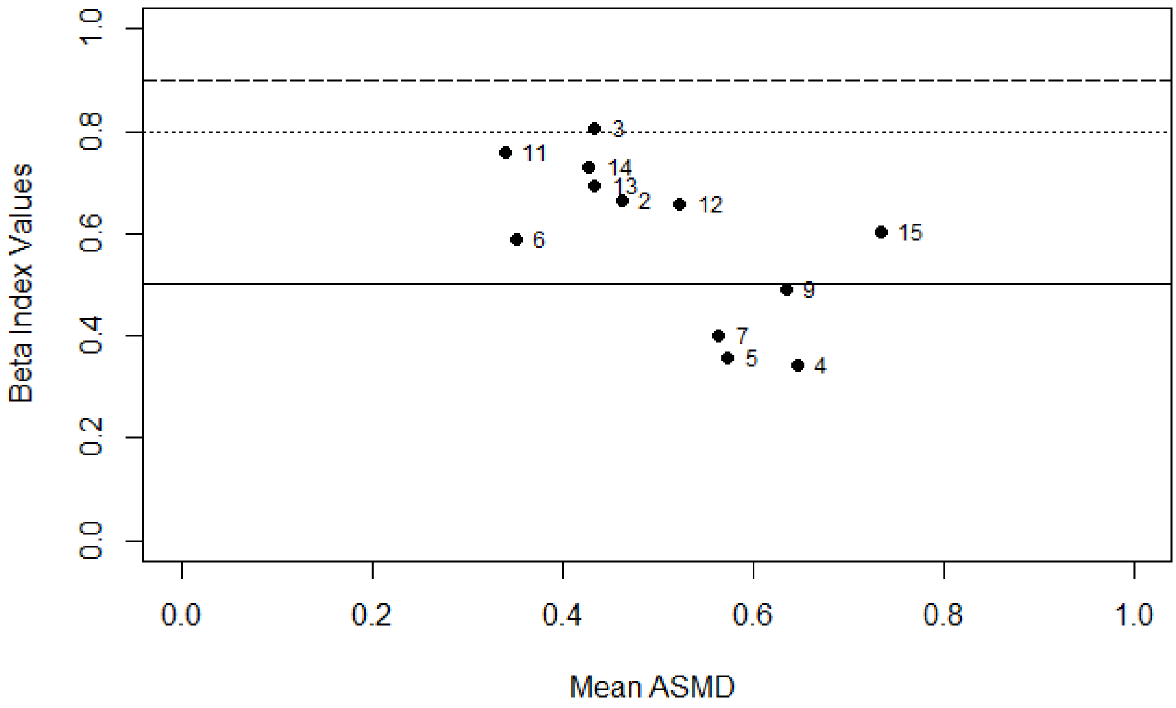
Table 4.9: Generalizability assessment for individual studies – Title I schools

| Study ID | B-index | $\Delta_{p\ logit}$ | ASMD | K-S test* | N | n |
|---|---|---|---|---|---|---|
| **11** | 0.757 | 1.741 | 0.340 | 0.567 | 29317 | <50 |
| **6** | 0.589 | 4.759 | 0.351 | 0.720 | 29317 | <50 |
| **14** | 0.731 | 3.982 | 0.428 | 0.532 | 29317 | <25 |
| **13** | 0.693 | 3.672 | 0.434 | 0.692 | 29317 | <75 |
| **3** | 0.805 | 3.608 | 0.434 | 0.576 | 29317 | <50 |
| **2** | 0.664 | 8.009 | 0.462 | 0.696 | 29317 | >100 |
| **12** | 0.658 | 2.700 | 0.524 | 0.639 | 29317 | <50 |
| **7** | 0.399 | 3.222 | 0.564 | 0.908 | 29317 | <25 |
| **5** | 0.356 | 2.125 | 0.574 | 0.962 | 29317 | <10 |
| **9** | 0.490 | 2.195 | 0.635 | 0.844 | 29317 | <50 |
| **4** | 0.343 | 2.916 | 0.647 | 0.909 | 29317 | <50 |
| **15** | 0.603 | 4.554 | 0.734 | 0.802 | 29317 | <75 |

*For K-S test: all were sig. at p<0.001
Note: all study IDs have been removed for confidentiality

In the scatter plot below we again see the comparison of study B-index values and the average ASMD of covariates for each study.

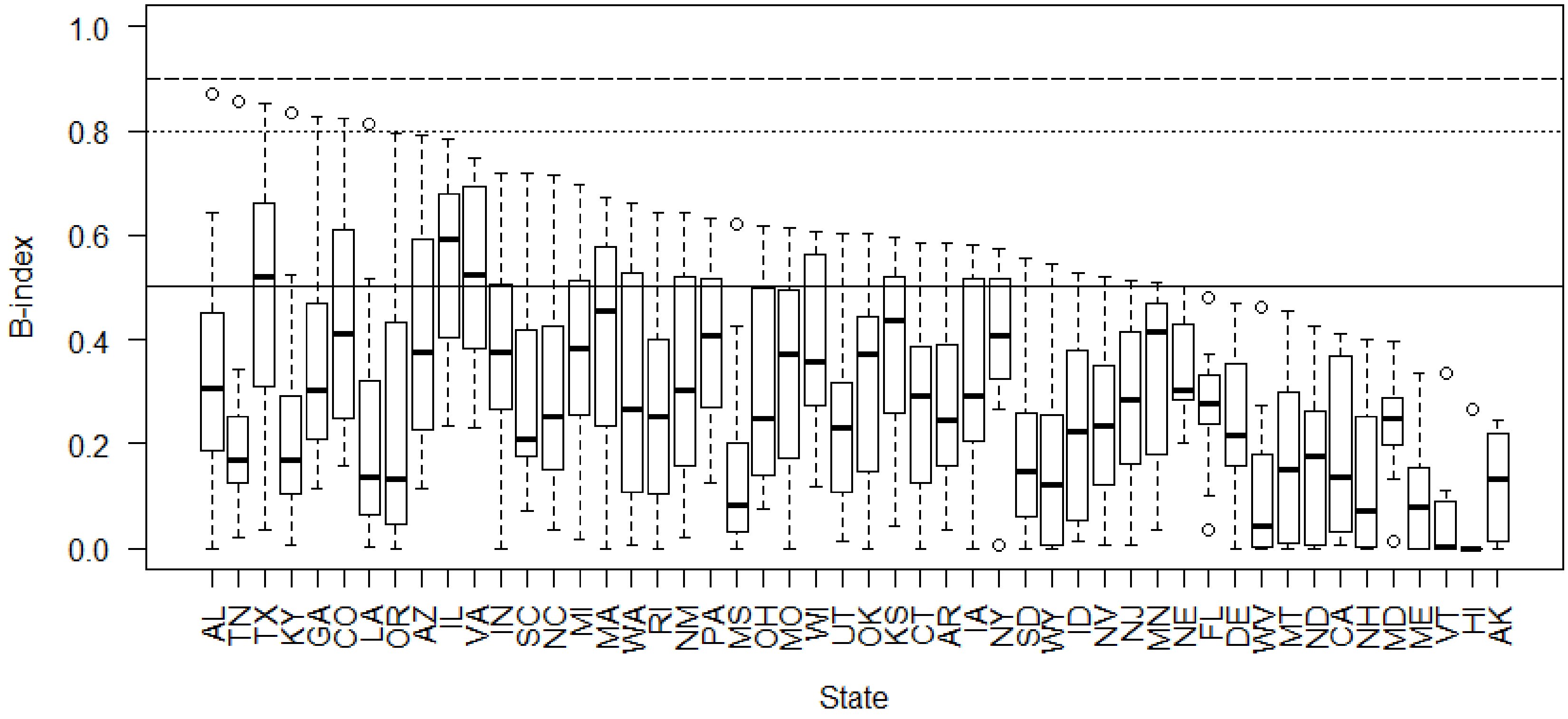
Figure 4.6: Comparison of individual studies to Title I schools



**Study by State Comparison**

**All schools.** For the comparison of each study to individual state's population of all elementary schools, the B-index will be the primary generalizability assessment estimate of interest. The box-plot below shows the distribution of B-index values for each study for each state. Only 101 out of the 600 total comparisons (12 studies by 50 individual states) have a B-index value greater than 0.50, indicating that reweighting to estimate treatment effects for the state's elementary school population is possible. Eight of those 101 comparisons, in seven unique states, have a B-index value greater than 0.80, meaning that reweighting will lead to less biased PATEs for those studies.  It should be noted that across these eight comparisons five of those states are well represented by the same study. Ten comparisons, across seven additional unique states, have values greater than 0.70 indicating that although bias will be greater, reweighting is also a useful strategy.  It is also clear that for some states there is a distribution of B-index values across the 12 studies, but for others it is mostly skewed in one direction (usually toward smaller B-index values).  States like, Alabama, Tennessee, and Texas are better represented across studies, while states such as Alaska, Hawaii, and Vermont are not.

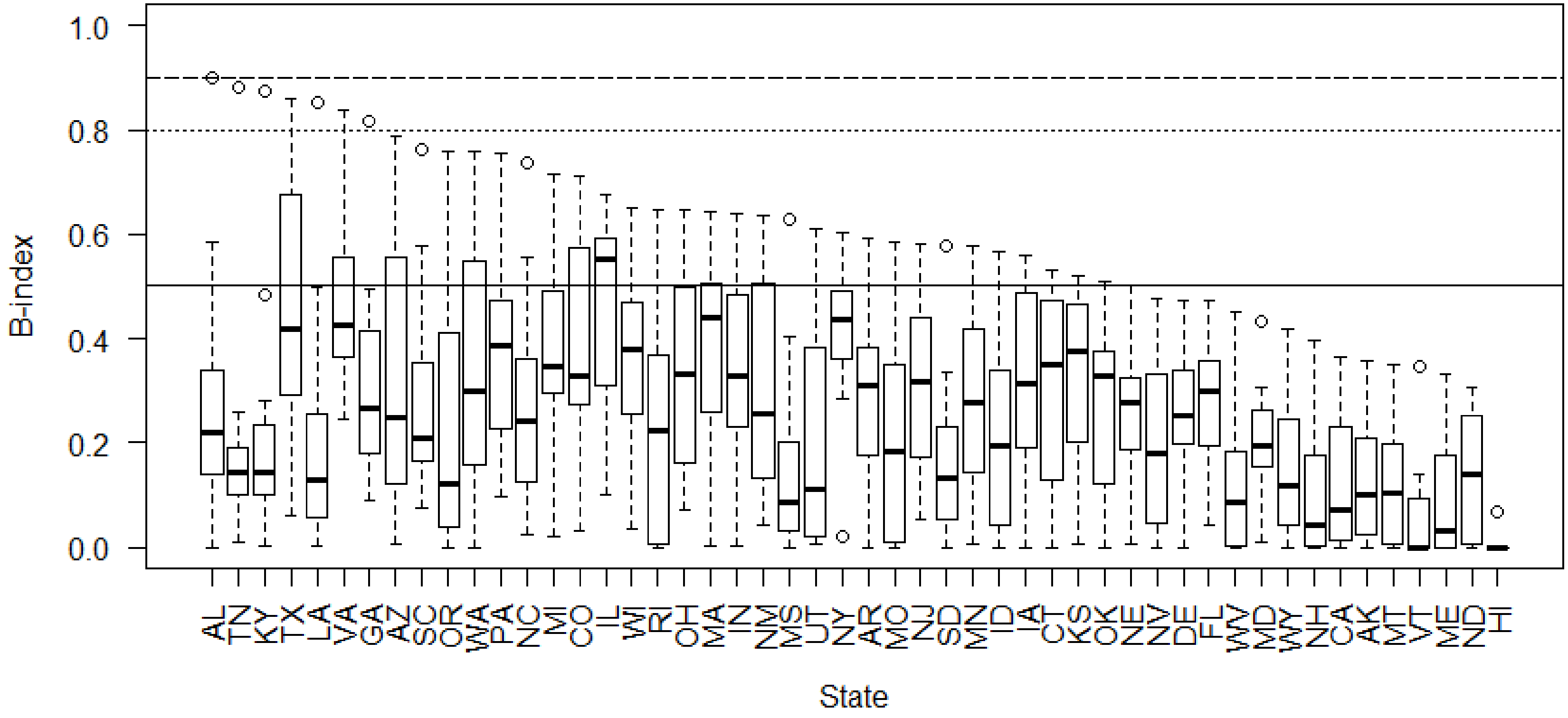
Figure 4.7: Comparison of individual studies to state population of all schools



**Title I schools.** For the comparison of each study to an individual state's population of Title I elementary schools, the B-index will be the focus of assessing generalizability. The box-plot below shows the distribution of B-index values for each study across all states.

Of the 600 comparisons made between study samples and Title I schools for each state, only 86 have B-index values greater than 0.50 (14%). This is less than the total of comparisons to the population of all elementary schools. The distribution of B-index values across studies in each state is less consistent than the population of all schools. It also appears that some of the same states are well represented (Alabama, Tennessee, and Texas) and the same are poorly represented (Hawaii and Vermont). A single state for a single study has a B-index value of 0.90. This means that this individual study is representative of the population of Title I schools in the state of Alabama. Six states had B-index values greater than 0.80 making reweighting a valuable strategy for calculating PATE (notably five of those six states are well represented by the same study). Ten comparisons, including five additional unique states, had values greater than 0.70 meaning reweighting will also likely be successful, although with slightly more bias in estimators.

Figure 4.8: Comparison of individual studies to state population of Title I schools



## Question 3

To answer this question a qualitative analysis was conducted. Coding for articles and interviews allowed for texts to be searched for several major themes regarding sampling, recruitment, and variables used to determine a school's eligibility, describe samples, and compare to populations. A codebook for both article and interview analysis and list of journals represented across all articles reviewed are provided.

From the 25 studies selected for this study 12 PIs were willing to participate in an interview.  Three PIs are responsible for more than one study within the dissertation sample. Three PIs who were able to share study school data were unavailable to be interviewed, and one was unable to share study school data but willing to answer interview questions.

Interviews, because of their directed nature, were coded for detail and statements that were notated by the interviewer.  Articles were coded systematically using Content Analysis with the unit of analysis being paragraphs within the relevant sections of published documents (abstract, research questions/purpose, methods, results, discussion/summary/conclusion). Table 4.10 and 4.11 show the codebooks used to analyze interviews and articles, along with the total count of articles or interviews and references where themes were coded.  To enhance the systematic review of articles word counts were also conducted within the sections mentioned previously for specific mentions of concepts by authors.

Table 4.10 Code book for Content analysis of interviews

| Major Theme | Definition | Extensions | Example | Sources | References |
|---|---|---|---|---|---|
| **Relationship** | Discussion of relationships between researchers and schools or districts that aided in the recruitment process | Mentions of partnerships that are produced or maintained for the sake of research projects | “Got buy in at district first; nice thing…was had a really good relationship with district superintendent and had close ties with the school” | 11 | 25 |
| **Variables** | Variables, such as proportion ELL, FRL, Ethnic/Racial make-up of a district or school that was used to determine eligibility or balance of sample between treatment and control | Geographic areas, ages of students, teacher experience, teacher’s highest degree earned, etc. | Proportion ELL, FRL, Ethnic/Racial make-up of a district or school | 9 | 17 |
| **Recruitment** | Any discussion of the recruitment process or plan | Mentions of incentives, methods, anticipated vs. actual recruitment of schools or districts | “contacted all of the schools, contacted all the principals, had 6 schools volunteer …” | 11 | 19 |
| **Eligibility** | Criteria used to determine if schools or districts were eligible to participate in a study | Variables used to select schools or districts | “[Schools with] English learners at risk for low reading outcomes” | 8 | 10 |
| **Incentives** | What was used to incentivize schools or districts to participate in a study | Any benefits provided to teachers, schools, or districts to increase the possibility of participation | “went in with a brochure; explaining how much [money] they were getting in terms of professional development, etc. that they wouldn’t get otherwise” | 1 | 1 |
| **Barriers** | Any discussion of obstacles or difficulties that arose during recruitment | Issues with research approval in a specific district, schools refusing after district agreement | “Had a very difficult time getting poorer schools to participate. Thinks that the SES level was much higher in her studies. Maybe because it was a supplement it was more appealing to middle class, maybe because it would be [riskier] for poor schools” | 9 | 20 |

*Note that references can be coded to multiple themes.

**Two PIs in the current analysis are responsible for more than one study in the sample of 25.

***Examples have been paraphrased.

The review of articles was not intended to account for any aspects of studies other than those related to recruitment, sampling, and generalization or external validity. Coding and analysis was solely focused on specific statements or remarks regarding these areas. Because of the nature of published articles regarding one study's findings across several publications, all articles collected for a single study are treated individually. It should be noted that all statements or paragraphs have the ability to be coded to more than one theme meaning that each source (interview, or study article) may contain several references (statements or paragraphs) coded within each theme.

## Content Analysis of Interviews

Interview analysis showed that 11 out of the 12 interviews noted that relationships were necessary for recruitment of schools into studies. The study that did not directly mention relationships as facilitating recruitment noted that they "stayed local" for the project, only using those schools that were geographically close by as the study required a large commitment of time to be in schools. Within the interviews there were 25 individual references to relationships and the role they play in recruitment. Ten PIs note that the relationships that aided with their recruitment of study schools was the result personal (directly with the PI or someone on the study team) relationships with *local* districts or schools and their administrators. One PI specifically mentioned that his was helpful in knowing which schools were functional which is necessary for study implementation. Five PIs attributed personal relationships with districts or schools to studies they had completed prior to the current study. One PI mentioned that partnerships are needed for school and community buy-in which aids study success. One PI mentioned that they were able to recruit study schools because of relationships that had been cultivated during presentations of other work.

When asked about their process of recruitment, most PIs noted practical constraints, such as location, time, and eligibility regarding their specific study aims (i.e., aimed at ELL students, so high proportions of those students were necessary). Most depictions of the recruitment process were similar, in that most PIs, as noted previously, had a working relationship with a district or school to start recruiting for their study sample (n=10). Recruitment was often discussed in the same conversation as eligibility criteria (high SES schools, high proportions of ELL students) of schools or students to participate in studies. It appears that the logical argument motivated most recruitment decisions.

Generally, PIs spoke about eligibility criteria and variables used for school or district selection within the same statements. For this purpose, these two coding themes are discussed together here. These criteria, whether due to study specifics (i.e., aimed at ELL students, so high proportions of those students were necessary) or for other reasons were used to guide recruitment. The variables mentioned most often were FRL (or SES as measured by FRL), ELL, and proportions of minority students (N=5, 6, and 5 respectively). Some studies (n=3) mentioned a need to focus on students that were low achieving in their study's subject area (i.e., low math or reading scores). Across some studies (n=2) PIs grouped these variables together, calling these students/schools/districts, "at-risk". Within ELL studies, PIs were cognizant of specific home language breakdowns (n=3). Geographic diversity (n=2) and district size (n=1) were also mentioned. In addition to these variables used for recruitment, individual PIs noted there were some other criteria for schools or districts to be eligible to participate in a study that were specific to their study only.

Only one PI mentioned incentives provided to encourage participation, including stipends for schools/teachers participating and professional development.

Nine of twelve PIs mentioned barriers to recruitment during their interviews.  Some PIs cite district attributes as the reason for difficulties in recruitment.  This encompasses, the dysfunctional nature of some districts (n=1) the buy-in from administrators (n=2), control of district decision making (Superintendent vs. individual principals) (n=2), difficulty in gaining research approvals(n=2).  Logistics are also listed as barriers experienced during recruitment, some (n=3) site these as timing of studies (notably if the district has just made core curriculum changes these studies are then less desirable; or if districts were in the midst of other initiatives e.g., rolling out common core standards), or the ability of the study team to physically get to sites as often as needed (n=2), changes in administration (n=1).  Attrition before and after random assignment was also cited as an issue (n=1).  One PI noted that there is a tendency to go for urban districts with more schools because of IES requirements for the number of schools in a study, makes it harder to recruit in rural district.  One PI stated that they needed more time to recruit schools.

## Question 4

### Content Analysis of Articles

The systematic review of articles of 13 studies (24 total articles) included three main coding areas.   First, sample variables: those variables used to describe the sample, an inference population of interest, or to define eligibility criteria for schools in the study.  Second, recruitment plans; any discussion of how schools were recruited, incentives used to encourage participation, or barriers that impacted recruitment practices.  Finally, generalization; items coded to this category included any discussion of generalization of study findings to other populations or subgroups, any mention of inference populations of interest, or limits to the ability of findings to be translated to other populations. Table 4.11 shows the codebook used for this analysis along with the percentages of units (paragraphs, as noted previously) coded to each theme. Each study can have more than one article published across various journals. Each article is coded independently even if reporting on the same study findings. Paragraphs within those articles were then coded, leading to a total number of references as seen in the table. It should also be noted that articles were coded only for content related to generalization, and the process of recruitment. No other aspects, including study quality, design, or significance of findings, were coded or considered for this analysis.

Table 4.11 Code Book for Content Analysis of Articles

| **Major Theme** | **Definition** | **Extensions** | **Example** | **Sources (n=24)** | **References** |
|---|---|---|---|---|---|
| **Variables - Eligibility** | Variables, such as proportion ELL, FRL, Ethnic/Racial make-up of a district or school used to describe study schools or those whom study findings might apply | Urbanicity, geographic areas, ages of students, teacher experience, teacher's highest degree earned, etc. | Schools were urban elementary schools located in specific geographic region with high percentage of ELL students | 21<br>9 | 73<br>11 |
| **Generalization – Inference population** | Any discussion of where and to whom study results apply, beyond study schools or districts | Planning recruitment with an inference population in mind, any mention of inference populations | Used a research-based model for math … with the intent to generalize the model to other subjects and other grade levels | 21 | 57 |
| **Recruitment – Barriers – Incentives** | Any discussion of the recruitment process or plan | Mentions of incentives, methods, anticipated vs. actual recruitment of schools or districts | "Contacted districts; some met the criteria as outlined by study goals | 16<br>3<br>1 | 16<br>3<br>1 |

*Note that references can be coded to multiple themes.
**Two PIs in the current analysis are responsible for more than one study in the sample.

Across the 15 studies represented in this content analysis there are several variables that are reported as a description of the sample, or used for eligibility criteria, or (in more rare cases, used to describe the population or the balance between treatment groups n=3, 1, respectively). The most frequently reported variables are socioeconomic status (as reported by the number of students eligible for Free or Reduced Price lunch program), n=30; Ethnic or racial breakdown, usually with an emphasis on the proportion of minority students, n=28; English Language Learners, n=40. Gender, district size (as indicated by urbanicity or total number of schools), and were also frequently reported, n = 23, 15, respectively. Some variables regarding teachers were also reported. Years of experience, n=6, and highest degree or certification earned, n=4, were seen the most often. Other variables seen were proportions of students registered for school services or IEP (n=14), standardized subject area test scores (n=20).

Recruitment is discussed in articles in a matter of fact way, in many cases simply indicating those schools that were ultimately selected to participate. Recruitment planning, or sampling, is usually mentioned by way of eligibility for the study, or in one case to disclose incentives provided to participants (i.e., professional development for teachers they would not otherwise receive). Issues in recruitment coded as barriers (n=3), in two cases were in reference to attrition, with one study noting that treatment assignments within schools became an issue forcing a slight change to the design. This analysis shows that eight references to recruitment are in direct mention of eligibility criteria, in five cases was stated as met by study schools with no mention of those who may not have met the criteria. In some articles there is a discussion of participants who were recruited but after employing eligibility criteria or not completing consent were not included in the final sample (n=3). These studies cited attrition, random assignment, and no consent as reasons for excluding participants after selecting them for the study. There were no mentions or descriptions across any articles about those schools or districts that were eligible and chose not to participate. In 13 of the 16 references (across 11 studies and 16 total articles) recruited participants were discussed, ten were described using previously mentioned variables, however there was no mention of how districts or schools were recruited in order for students or teachers to be included in a study's sample. In four of these references districts or schools are described as being invited, contacted, or volunteering to be in a study as the method of recruitment with no mention of any of those districts who declined participation. One study had an application process for study participation, although no details about the number of total applicants was provided. In only one case was the process of recruitment disclosed in an article, noting that three districts were contacted to determine interest, and after an informational meeting with potential schools, that the interested participants agreed to be included in a random assignment for treatment conditions.

When discussing research questions, results, and conclusions, generalization and external validity are referred to directly very few times (n=7). Six direct references to generalizing findings are in support of *not* generalizing beyond a population similar to the study sample. The final direct reference, from an effectiveness scale-up study, notes that findings could generalize to a larger population. In most articles (20 out of 24 total articles) authors mention where or to whom their findings apply. Those references within articles range in the manner in which they discuss this application of findings. Nine references were made to specific target populations (i.e., 3rd grade classrooms with low-achieving math students). Those populations included one reference to all U.S. classrooms, three reference low-income or Title I students/schools, two reference African-American students, and two reference all students within a specific grade. Eleven references were made that study findings were relevant to an unspecific population group of their participants (i.e., if a study was focused on English language learners, authors concluded that study findings could be expected for all ELL students, with no caveats respective to the sample). In other references (n=17) authors were more conservative in their discussion making sure to note the limitations of their sample as impeding the application of findings to a larger population. With five studies noting that because of these limitations findings could only be applied with certainty to populations that are like the sample. Only one reference was made that researchers were claiming direct causality of findings, while four studies made references to the moderators that they assessed as interfering with treatment effects (e.g., gender or race). Fourteen references were made across the studies that future research was needed where variations in the current sample, expanding grade levels or increasing differences in myriad variables, would garner larger more appropriate generalizations.

For reference, an exact match word frequency was run to determine how often other standards required for consideration in WWC publication, specifically attrition and power as these have received great attention in creating more reliable and rigorous educational research. The same search was also conducted for generalization. From this table it appears that while attrition and power are the most frequently reported, likely because they are required for consideration to the WWC, generalization and its counterparts are being discussed without clear or strict guidelines as to what should be reported and what evidence should be shown for those claims.

Table 4.12: Word frequency analysis

| Word | Number of Studies | Number of Total Articles | Number of Occurrences |
|---|---|---|---|
| **Attrition** | 10 | 16 | 78 |
| **Power** | 11 | 17 | 48 |
| **Population** | 8 | 10 | 22 |
| **Generalize** | 5 | 10 | 12 |
| **Generalization** | 1 | 1 | 1 |
| **External Validity** | 3 | 4 | 5 |
| **Inference population** | 0 | 0 | 0 |

# Chapter 5: Discussion

The Department of Education's Institute of Education Sciences (IES) has tasked itself with conducting high quality, rigorous research with the goal of improving achievement for all students and specifically those at high academic risk. The evolution of standards and requirements set forth by IES for researchers has shown the progress of this mission. As the field of education begins to turn its attention beyond internal validity of studies to external validity and the generalizability of study findings, so too should IES requirements and What Works Clearinghouse standards. The aim of this dissertation is to develop an approach that IES and researchers can use in order to better understand, assess, and report generalizability of study findings of past, present, and future work. In this section, I discuss the findings more generally, focusing on results that are useful for various audiences of researchers and policymakers.

## Results for IES (Questions 1 and 3)

Over the past 12 years, IES has funded 218 efficacy and effectiveness trials. Results from these studies – if studies are implemented well – are reported in the WWC. These grant funded studies are proposed by individual researchers at universities and research firms. An important question, therefore, is how well these studies represent the populations of schools that IES serves. In this dissertation, I focus on a subset of studies aimed at of elementary schools both generally and how they compare to the population of all U.S. elementary schools and those schools that are using Title I funds school wide (indicating a more at risk student group).

From this analysis, we see that the overall sample of schools (n=571 across the 15 studies included in this analysis) has a B-index value of 0.915, which while indicating some similarity, is not considered representative of the population of U.S. elementary schools. Findings were similar (though slightly worse) for the comparison of all study schools to the population of Title I schools in the United States (B-index=0.860). When comparing overall study schools to these populations in individual states B-index values indicated that this sample of 571 study schools across studies was not like a random sample of any single state. Future analyses should consider if an easily defined subpopulation can be defined for whom the schools studied *are* generalizable. For example, in Tipton et al (2016) it was found that the inference population represented best by the study sample differed from the full population in terms of measures of district size. That is, study schools tended to be in larger districts than those in the U.S. population.

Given that schools in the sample of IES studies analyzed here do not represent the populations of (Title I) elementary schools in the U.S., an important question is how schools taking part in IES funded studies *differ* from those in the population. In this dissertation, I showed that schools that *participated* in IES grant funded studies were more often in urban areas, have higher percentages of minority and lower percentages of white students, belong to larger school districts and are themselves schools (higher enrollments), and have higher percentages of English language learners (ELL). Conversely those schools *NOT* participating in IES studies tend to be smaller (in terms of enrollment), are members of smaller school districts, are more commonly in rural areas, have higher percentages of white students and lower percentages of minority students, and lower percentages of ELL students. This is true for both the population of all elementary schools and Title I schools. This finding is similar to that found in a study conducted by Bell and colleagues (2016) that reviewed the types of districts that typically participate in large-scale contract studies funded through IES (NCEE); that is, they too found that larger districts in urban areas were more often included in evaluations.

The qualitative analysis of interviews, used to understand how schools are recruited into studies (indicating why these differences might arise), showed that prior relationships with schools or districts are valuable and imperative to the success of recruitment for these studies. This was true across most studies. Thus the fact that many research universities and research firms are located in metropolitan areas may play a part in the types of schools that are available and willing to take part in studies. Interview analysis also revealed similarities in the approach to recruitment across various types of studies. One of the main similarities across these studies is the variables used to describe desirable districts or schools to be targeted as well as to determine eligibility. These include racial or ethnic variables, proportion of ELL in a school or district, geographic location, and socioeconomic status (usually measured by the proportion of students eligible for FRL). These variables are also frequently used to describe samples when researchers report study findings. This is an important finding as using these covariates to describe the inference population of interest could be largely helpful for researchers in planning for and reporting generalization.

Findings from this qualitative review can be used to guide updates of recommendations in future IES requests for proposals and study reporting standards from WWC regarding generalization. By including these common variables and knowing more about the methods of recruitment used by PIs, IES could increase the formality of recruitment, sampling, and definitions of inference populations of interest. Although recruitment, eligibility, and populations of interest are nuanced across projects, a set of guiding principles when defining populations and outlining recruitment procedures could be extremely beneficial to researchers. By making this process formal across projects, PIs could better plan for generalization.

ultimately making their study findings more useful to the practitioners, policy makers, and individual educators.

Interview analyses also showed that across studies PIs experienced similar barriers to recruitment. Several researchers noted that time and budgetary concerns were an issue when targeting districts and schools.  Having more information directly from PIs for other PIs about effective strategies to recruitment and the common issues or pitfalls would be useful not only to future funded studies but also the field as formalization of this process leads to better practices. Together these findings suggest that placing the responsibility of representing the population of all schools solely with researchers may not be the best strategy.  To better serve all schools, and researchers, IES should focus on populations of priority (i.e., those at the greatest academic risk). Providing this information, as well as the support for those aiming interventions at these groups, would help researchers better recruit a diverse set of schools.

### Results for Researchers (Questions 2 and 4)

The analysis of generalizability in individual studies is of most importance to this work. To focus this conversation about generalization, Question 2 aims to assess the generalizability of individual studies to the inference populations that IES represents.  In these analyses, the schools in *each* study were compared to the two inference populations: one of elementary schools and the other of Title I elementary schools.

Importantly, it is possible for the sample of schools in a *single* study to be representative of a population, while at the same time the *combined* sample of schools across studies might differ from the inference population. Conversely, it is also possible that the *combined* sample of schools across studies could represent a population well, while *none* of the individual studies represents the population well.

Findings from this dissertation indicate that in the sample of schools included, none of the study samples of schools were like a random sample of the U.S. population of elementary schools, or the population of Title I elementary schools.  Further no single study showed that the study sample was like a random sample of the population of U.S. elementary schools or Title I schools in any individual state.

While this is an important finding, it is not the end of the conversation. Namely, the focus of generalizability analyses is not typically limited to simply assessing if a study is like a random sample or not, but instead focuses on determining if the reweighted estimators of the population average treatment effect (PATE) are useful. Following the rules of thumb found from simulations in Tipton (2014) and restated in Table 4.6, 75% of the studies would be able to utilize reweighting strategies in order to provide estimates of the PATE for the population of all U.S. elementary schools and 66% of studies would be able to use reweighting to estimate PATE for the population of Title I schools.

Further investigation of individual study sample generalizability showed that when compared to the population of all U.S. elementary schools no single study sample was representative of a single state (i.e., no B-index value was greater than 0.90).  However, 101 of the 600 comparisons (individual studies by individual states) showed that for those state populations, reweighting would be a useful strategy for estimating PATE.  When compared to the population of Title I schools a single study was found to be representative of a single state

(B-index=0.90). Eighty-six of the 600 comparisons have B-index values greater than 0.50 meaning that for those studies reweighting could be useful to estimate PATE for those states.

The qualitative review of published articles was aimed at better understanding how researchers report generalizability and further, describe their samples or inference populations of interest. Findings from this review showed that while there are common variables across almost all articles across studies, there are not as many commonalities with regards to reporting where and to whom study findings apply.

Within the qualitative analysis of published study results, several common variables were used to describe the sample. Variables frequently used across all articles and studies were similar to those found during the interview analysis. Racial and ethnic breakdowns, proportions of students labeled as ELL, proportion of students eligible for FRL, district size, geographic location, and urbanicity were found to be the most commonly reported. When reporting these things (most authors do report most of these variables) researchers often only report percentages, or proportions, and focus only on describing the sample. In only one article did an author refer to these variables as they describe the population of interest. In addition to these some researchers report teacher level variables (years of experience and highest degree earned). These variables are also used to test differences in treatment effects and in some instances the findings for specific sub-groups within the sample (i.e., large treatment effects for African-American males) are being generalized beyond the study sample.

The qualitative review of articles also shows that some authors state that findings can be applied well beyond their study sample without any formal evidence of that generalization. For instance, if significant treatment effects were found, an author posits that those findings are relevant to all U.S. classrooms.  Qualitative analysis of interviews and articles reveals that external validity and generalization are not always a priority when planning recruitment or reporting findings for published articles.

Currently WWC only mentions generalizability (listed as external validity) as a part of the total check for validity within studies.  This lack of detail could be hindering researchers from knowing how or what to report in regards to the generalizability of their study findings.  In the CONSORT literature, flow charts are utilized to document sampling procedures that lead to the analytical study sample.  These methods could be adapted and used for the purposes of generalization of IES studies.  This might include a CONSORT style flow chart of recruitment, including details on who was recruited, who was unresponsive, those that declined participation and those that agreed to be included in the final study sample.  A chart of this type was employed in Tipton et al (2016) to track non-response data for a complete understanding of those districts and schools that ultimately participated in two effectiveness trials.  Encouraging or requiring researchers to report on this recruitment process could move the practice of recruitment to a more formal process, as well as open conversations and research agendas related to improving practice.  Additionally, this could help researchers think about and officially set a well-defined population of interest prior to recruiting schools or districts.

A benefit of encouraging researchers to report in a standardized way about recruitment and generalizability is that it could greatly improve the type of data available for the WWC, which is useful to practitioners and policy makers. Additionally, if researchers provided this information at the beginning of a study, it could be easier to ask researchers to assess and report the similarity of their sample to their desired inference population of interest.  These analyses might be similar to those conducted in Question 2 here, but could also include describing well any non-response or refusals. Detailing how and what evidence to provide in support of addressing generalizability could help the WWC further provide education stakeholders with studies that are appropriate for *their* students.

### Takeaways for Readers

The lessons learned from the quantitative and qualitative analysis of this research leave the field with some suggestions for improvement in the area of generalizability. Here I focus on three potential next steps:

1. As researchers we can be more open about recruitment practices and the most effective strategies and incentives. Further, gathering this data can help IES to better incentivize schools to participate in large-scale studies. By creating a better network of data regarding recruitment and incentives, researchers can be better prepared to target participants and expand their pool of possible study schools.
2. The IES could begin to require researchers to describe in detail inference populations of interest during the proposal stage as well as progress and final reports. This would aid researchers in their ability not only to recruit but to aim study findings to an appropriate audience. This would aid WWC in classifying studies that it admits to the online database.
3. The WWC and IES could develop standards for assessing and reporting generalizability. These standards could be similar to the methods in which standards were developed to improve internal validity in studies. This could help stakeholders better interpret and make use of study findings filed with WWC.

The field of educational research has already started a conversation regarding ways to improve the external validity of studies.  The findings from this dissertation provide feasible steps for researchers to assess the generalizability of their own studies, as well as for funding agencies to assess the combined representativeness of their funded research.  These final takeaways will aid in the movement of IES and researchers to better plan for generalizability through sampling and recruitment; better assess generalizability of their own studies in a statistical way; and finally, to report it in a formal and consistent way.

## Limitations of This Work

This study is limited by the small number of studies included in the final analysis partially due to non-response issues.  However, it does provide the necessary framework for future protocols of assessing generalizability of a study's sample to a specified population of interest.

Qualitative findings were not validated by a second rater and therefore should be interpreted and discussed with caution. These should be viewed as a beginning phase to document how researchers approach and discuss recruitment and how recruitment and generalizability are reported in published articles.  These qualitative reviews were also used to inform the covariate selection for the logistic regression to calculate propensity scores for the quantitative analysis. Based on these interviews, additional variables affecting recruitment in these models should be included, such as measures of distance from research universities or firms.

Finally, this work is only looking at the selection mechanism of schools into study samples.  The aim of this study was to examine generalization and its relationship to recruitment and sampling. Although often the logical argument many researchers are making (possibly unintentionally) is that the variables that are used for eligibility, or that are reviewed when deciding which districts or schools to target for selection, this dissertation does not test or determine which covariates might best explain variations in treatment outcomes. Knowing how treatment effects may vary between groups of students or subsets of schools or districts could further inform and formalize the process of recruitment. The investigation of treatment effect heterogeneity is a large component of this field of study and should also be considered when reporting the generalization of findings. Schochet and colleagues (2014) reviewed the literature regarding treatment effect heterogeneity and note that these differences can be grouped into two main categories, those that influence treatment effects prior to intervention and those that occur due to do the contrast between treatment and control conditions. This work notes that some possible moderators of treatment effects can be site (meaning school or district) level characteristics such as location or available resources, but that they could also be due to design decisions including the over or under sampling of a specific group. The authors suggest that as many of these factors are accounted for during the design phase as possible to reduce the influence over treatment impacts.

## Implications for Future Work

As this dissertation shows, the current process of recruiting districts and schools into experiments is typically informal and non-statistical. One of the aims of this dissertation is to provide a framework for researchers to move conversations regarding recruitment and reporting generalizability of study findings to larger populations to be more statistical and formal. The process outlined in this work can be used by researchers in the future to provide evidence of the generalizability of their study to any specified inference population of interest. Other work that investigates the generalizability of IES studies across different grade levels as well as other program areas is needed. Although there is a single study (Bell et al, 2016) that looks at the district participation in a small (n=11) number of NCEE contract awards, this area needed to be explored on a larger scale. Further exploring who typically participates in large-scale studies will help to inform future practices of recruitment. The more transparent the processes of recruitment become the better prepared future researchers can be when targeting and providing incentives to districts and schools. Other aspects of selection should also be investigated, including differences in recruitment between individual researchers and large research firms. Additionally, the role of some variables such as geographic location and the distance of participating study schools and districts to researchers.

Increasing the formality and rigor with which external validity is addressed in studies can help practitioners and decision makers better utilize studies found in IES and WWC repositories as they make choices about implementing programs and interventions in their respective districts and schools. Creating a framework for assessing and discussing generalization can also help WWC and IES shift standards to better help researchers know what evidence to report and how to discuss the generalization of study findings. While this study is focused on the larger populations that are of interest to IES as a whole, the concept can be widely applied as it is able to be mapped out to align with specific studies and the aims of individual researchers.

## References

Campbell, D.T. (1986). Relabeling internal and external validity for applied social scientists. *New Directions for Program Evaluation*, *1986*(31), 67-77.

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

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG-320 trial. *American Journal of Epidemiology*, 172, 107–115.

Cook, T.D., Campbell, D.T. (1979). *Quasi-Experimentation: Design analysis issues for field settings*. Houghton Mifflin Company

Cook, T.D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them: New directions for program evaluation* (Vol. 57, pp. 39–82). San Francisco, CA: Jossey-Bass.

Cronbach, L.J., (1982). *Designing evaluations of education and social programs*. Jossey-Bass

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., Weiner, S.S. (1980).*Toward reform of program evaluation: Aims, methods and institutional arrangements*. Jossey-Bass.

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

Government Accountability Office. (2013). *Education research: Further improvements needed to ensure relevance and assess dissemination efforts*. Report to the committee on education and the workforce, House of Representatives. GAO-14-8.

Institute of Education Sciences. (2015) *About us*. Retrieved from http://ies.ed.gov/aboutus/

Institute of Education Sciences. (2015) *Requests for applications: Education research grants*. CDFA number 84.305A. Retrieved from http://ies.ed.gov/funding/pdf/2016_84305A.pdf

Institute of Education Sciences. (2014) *Requests for applications: Education research grants*. CDFA number 84.305A. Retrieved from http://ies.ed.gov/funding/pdf/2015_84305A.pdf

Institute of Education Sciences. (2011) *Requests for applications: Education research grants*. CDFA number 84.305A. Retrieved from http://ies.ed.gov/funding/pdf/2012_84305A.pdf

Institute of Education Sciences. (2010) *Requests for applications: Education research grants*. CDFA number 84.305A. Retrieved from http://ies.ed.gov/funding/pdf/2011_84305A.pdf

Institute of Education Sciences. (2008) *Requests for applications: Education research grants*. CDFA number 84.305A. Retrieved from http://ies.ed.gov/funding/pdf/2009_84305A.pdf

Institute of Education Sciences. (2004) *Requests for applications: Education research grants*. CDFA number 84.305A. Retrieved from http://ies.ed.gov/funding/pdf/2005_84305A.pdf

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

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: principles and quantitative methods*. John Wiley & Sons.

Kruskal, W. H., & Mosteller, F. (1988). Representative sampling. *Encyclopedia of Statistical Sciences*.

Kruskal, W., & Mosteller, F. (1979). Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review/Revue Internationale de Statistique,* 111-127.

Kruskal, W., & Mosteller, F. (1979). Representative sampling, III: The current statistical literature. *International Statistical Review/Revue Internationale de Statistique*, 245-265.

Kruskal, W., & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique*, 169-195.

Lohr, S. (2009). Chapter 6: Sampling with unequal probabilities. *Sampling: Design and Analysis* (pp. 194-213). Pacific Grove, CA. Brooks/Cole Publishing Co.

National Center for Education Statistics: Common Core of Data. (2016). *Elementary/Secondary information system*. [data for public schools 2013-2014]. Retrieved from https://nces.ed.gov/ccd/elsi/tableGenerator.aspx

Olsen, R.B., Orr, L.L., Bell, S.H., & Stuart, E.A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*. Advance online publication. Doi:10.1002/pam.21600

O'Muircheartaigh, C. A, Hedges, L.V. (2014). Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(2): 195-2010.

Rhode Island Department of Education. (2015). *Report of Title I schools in Rhode Island*. [data for public schools in 2013-2014]. Retrieved from http://www.ride.ri.gov/Portals/0/ Uploads/Documents/ Students-and-Families-Great-Schools/Educational-Programming/Title1/TitleISchools-201314-2Dec15.pdf

Riecken, H. W., & Boruch, R. F. (Eds.). (1974). *Social experimentation: A method for planning and evaluating social intervention*. Elsevier.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70,41–55.

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

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. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2: 169-188.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

Shadish, W.R., Cook, T.D., Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth, Cenage Learning. Belmont, CA.

Spybrook, J. (2008). Are power analyses reported with adequate detail: Findings from the first wave of group randomized trials funded by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness, 1(3), 215-235*.

Spybrook, J., Cullen, A., & Lininger, M. (2011). An examination of the impact of changes in federal policy on the landscape of education research. *Effective Education*, 3(2).

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis, 31*(3), 298-318.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2016). Characteristics of School Districts That Participate in Rigorous National Educational Evaluations. *Journal of Research on Educational Effectiveness*, (published online 30 June 2016), 1-39.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174: 369–386.

Tipton, E. (2013). Improving generalization from experiments using propensity score subclassification: Assumptions, properties, and contexts.  Journal Educational and Behavioral Statistics, 38, 239-266.

Tipton, E. (2014a). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478-501.

Tipton, E. (2014b). Stratified sampling using cluster analysis a sample selection strategy for improved generalizations from experiments. *Evaluation review*, *37*(2), 109-139.

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114-135.

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & De Castilla, V. R. (2016). Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness*, 9(Sup1), 209-228.

Valliant, R., Dorfman, A.H., & Royall, R.M. (2000). *Finite population sampling and inference: A prediction approach (Wiley Series in Probability and Statistics)*. New York, NY: Wiley.

Waernbaum, I. (2010). Propensity score model specification for estimation of average treatment effects. *Journal of Statistical Planning and Inference, 140*(7), 1948-1956.

West Virginia Department of Education. (2016). *Source book 2014: Child nutrition data*. [data for public schools in 2013-2014]. Retrieved from http://wvde.state.wv.us/finance/ sourcebooks/2014-source-book.pdf

What Works Clearinghouse. (2015). *Who we are.* Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/ WhoWeAre

What Works Clearinghouse. (2104). *Procedures and standards handbook version 3.0*. Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/Handbooks

# Appendix A

## Inclusion/Exclusion Variables

Covariates used to employ inclusion/exclusion criteria for population schools in final database for comparison to study schools:

- School Level (to include only elementary schools – This removed the largest amount of schools
- Operational status
- Charter school
- Magnet school
- Total students (school)
- Total number of Full Time Teachers (FTE)

All other variables used in the logistic regression or for comparison are listed in table 4.1 and 4.3 in Chapter 4: Results.

The following table shows by state how many schools were included in the original data frame constructed from CCD data and how many schools remained after inclusion criteria and removal due to missing data occurred.

Table A.1 Count document of Inclusion/Exclusion and missing data procedures

| State | CCD schools* | Inclusion/ Exclusion** | Final N after missing data | Title I final N after I/E and missing data |
|---|---|---|---|---|
| **AK** | 174 | 127 | 127 | 82 |
| **AL** | 751 | 670 | 662 | 528 |
| **AR** | 546 | 498 | 498 | 449 |
| **AZ** | 1167 | 844 | 765 | 555 |
| **CA** | 5895 | 4785 | 4668 | 3065 |
| **CO** | 1052 | 871 | 871 | 341 |
| **CT** | 653 | 548 | 548 | 135 |
| **DE** | 121 | 101 | 101 | 89 |
| **FL** | 2234 | 1682 | 1606 | 1346 |
| **GA** | 1299 | 1244 | 1243 | 927 |
| **HI** | 180 | 169 | 169 | 113 |
| **IA** | 724 | 698 | 680 | 388 |
| **ID** | 366 | 289 | 286 | 241 |
| **IL** | 2478 | 2107 | 1958 | 890 |
| **IN** | 1100 | 1033 | 1023 | 730 |
| **KS** | 742 | 704 | 689 | 517 |
| **KY** | 779 | 702 | 688 | 612 |
| **LA** | 731 | 647 | 633 | 588 |
| **MA** | 1118 | 1045 | 1043 | 349 |

| | | | | |
|---|---|---|---|---|
| **MD** | 909 | 821 | 821 | 431 |
| **ME** | 358 | 312 | 303 | 230 |
| **MI** | 1727 | 1242 | 1239 | 752 |
| **MN** | 1033 | 734 | 730 | 176 |
| **MO** | 1264 | 1167 | 1158 | 835 |
| **MS** | 451 | 452 | 432 | 409 |
| **MT** | 423 | 271 | 265 | 169 |
| **NC** | 1425 | 1274 | 1217 | 1032 |
| **ND** | 266 | 230 | 230 | 78 |
| **NE** | 622 | 538 | 522 | 277 |
| **NH** | 293 | 266 | 266 | 86 |
| **NJ** | 1559 | 1459 | 1446 | 325 |
| **NM** | 476 | 417 | 413 | 375 |
| **NV** | 391 | 324 | 319 | 205 |
| **NY** | 2559 | 2274 | 1364 | 686 |
| **OH** | 1916 | 1688 | 1676 | 1145 |
| **OK** | 967 | 934 | 932 | 833 |
| **OR** | 704 | 619 | 588 | 375 |
| **PA** | 1726 | 1619 | 1615 | 927 |
| **RI** | 181 | 169 | 165 | 69 |
| **SC** | 683 | 601 | 594 | 522 |
| **SD** | 333 | 234 | 231 | 143 |
| **TN** | 1051 | 806 | 800 | 683 |
| **TX** | 4532 | 4109 | 4046 | 3393 |
| **UT** | 607 | 493 | 493 | 179 |
| **VA** | 1178 | 1135 | 1107 | 481 |
| **VT** | 213 | 199 | 192 | 135 |
| **WA** | 1245 | 1105 | 1081 | 708 |
| **WI** | 1233 | 1075 | 1074 | 511 |
| **WV** | 451 | 484 | 180 | 98 |
| **WY** | 197 | 154 | 154 | 63 |

## Variable changes across data years

Variables regarding urbanicity in 2004-2010 were separated into four groups with two sizes within each group: City (large or midsize), Suburb (fringe of large or midsize city), Town (large or small), Rural (outside or inside CBSA). Urbanicity for 2010-2014 data were separated into the same four groups but with three categories within each group: City (large, midsize, or small), Suburb (large, midsize, or small), Town (fringe, distant, or remote), Rural (fringe, distant, or remote). Only three categories were used for analysis: City (all sizes), Suburb (all sizes) and Town or Rural, collapsed into one group (all sizes).

## Appendix B

## Individual Study Comparisons

Figure B1: B-index box-plot of individual studies compared to all schools

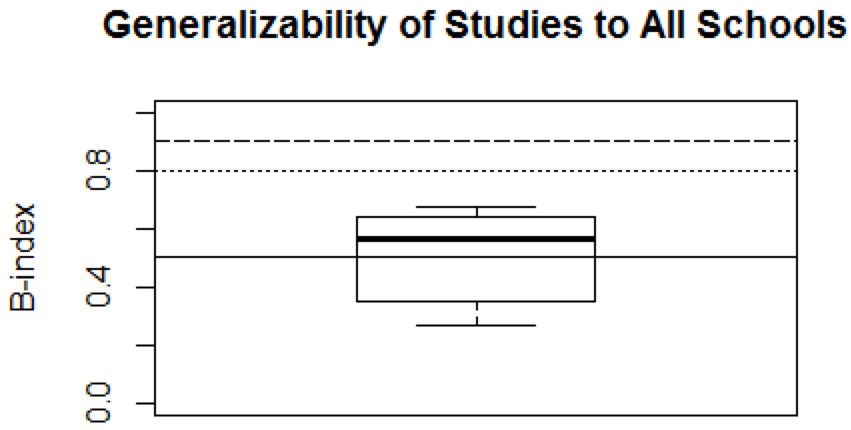


Figure B2: B-index box-plot individual studies compared to Title I schools

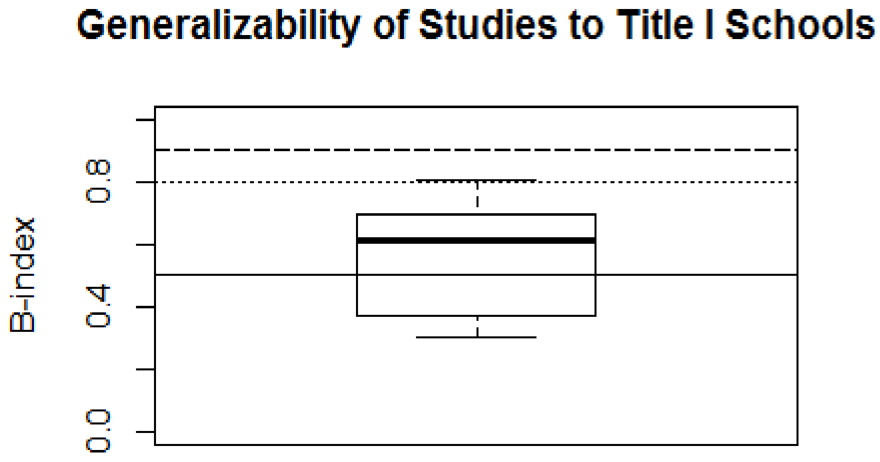


Table B.1: 5-number summary of B-Index - Studies to states, All Schools

| **Study ID** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
|---|---|---|---|---|---|---|
| **2** | 0.000 | 0.244 | 0.324 | 0.365 | 0.480 | 0.853 |
| **3** | 0.000 | 0.162 | 0.360 | 0.346 | 0.500 | 0.814 |
| **4** | 0.000 | 0.032 | 0.108 | 0.110 | 0.178 | 0.340 |
| **5** | 0.000 | 0.100 | 0.016 | 0.203 | 0.334 | 0.483 |
| **6** | 0.000 | 0.054 | 0.200 | 0.207 | 0.319 | 0.564 |
| **7** | 0.000 | 0.000 | 0.034 | 0.110 | 0.202 | 0.447 |
| **9** | 0.000 | 0.000 | 0.001 | 0.033 | 0.012 | 0.421 |
| **11** | 0.000 | 0.010 | 0.206 | 0.230 | 0.350 | 0.815 |
| **12** | 0.000 | 0.234 | 0.324 | 0.350 | 0.492 | 0.724 |
| **13** | 0.000 | 0.112 | 0.380 | 0.340 | 0.540 | 0.666 |
| **14** | 0.000 | 0.178 | 0.346 | 0.323 | 0.466 | 0.743 |
| **15** | 0.01 | 0.158 | 0.270 | 0.295 | 0.432 | 0.684 |

Table B.2 5- number summary of B-index - Studies to states Title I schools

| **Study ID** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
|---|---|---|---|---|---|---|
| **2** | 0.000 | 0.175 | 0.247 | 0.269 | 0.346 | 0.736 |
| **3** | 0.000 | 0.143 | 0.289 | 0.294 | 0.442 | 0.651 |
| **4** | 0.000 | 0.033 | 0.126 | 0.132 | 0.196 | 0.443 |
| **5** | 0.000 | 0.052 | 0.096 | 0.137 | 0.208 | 0.436 |
| **6** | 0.000 | 0.026 | 0.140 | 0.185 | 0.298 | 0.631 |
| **7** | 0.000 | 0.000 | 0.007 | 0.080 | 0.101 | 0.437 |
| **9** | 0.000 | 0.000 | 0.001 | 0.034 | 0.033 | 0.347 |
| **11** | 0.000 | 0.003 | 0.172 | 0.222 | 0.334 | 0.837 |
| **12** | 0.000 | 0.182 | 0.268 | 0.296 | 0.426 | 0.704 |
| **13** | 0.000 | 0.105 | 0.362 | 0.320 | 0.500 | 0.695 |
| **14** | 0.000 | 0.218 | 0.364 | 0.353 | 0.487 | 0.831 |
| **15** | 0.023 | 0.152 | 0.321 | 0.309 | 0.450 | 0.690 |

Table B.3: List of journals where reviewed articles are published

| Journal Title | Number of Articles Represented |
|---|---|
| American Educational Research Journal | 1 |
| AERA Conference presentation | 1 |
| Cognition and Instruction | 1 |
| Computers and Education | 1 |
| Computers in Human Behavior | 1 |
| Early Childhood Research Quarterly | 1 |
| Educational Administration Quarterly | 1 |
| Educational Evaluation and Policy Analysis; | 2 |
| Education Tech Research Development | 1 |
| The Elementary School Journal | 7 |
| Journal of Educational Psychology | 3 |
| Journal of Research on Educational Effectiveness | 4 |
| Journal for Research in Mathematics Education | 1 |
| Journal of School Psychology | 1 |
| Journal of Teacher Education | 1 |
| Learning and Instruction | 1 |
| Reading and Writing: An interdisciplinary Journal | 2 |
| Reading Research Quarterly | 1 |
| School Psychology Review | 1 |
| Scientific Studies of Reading | 1 |

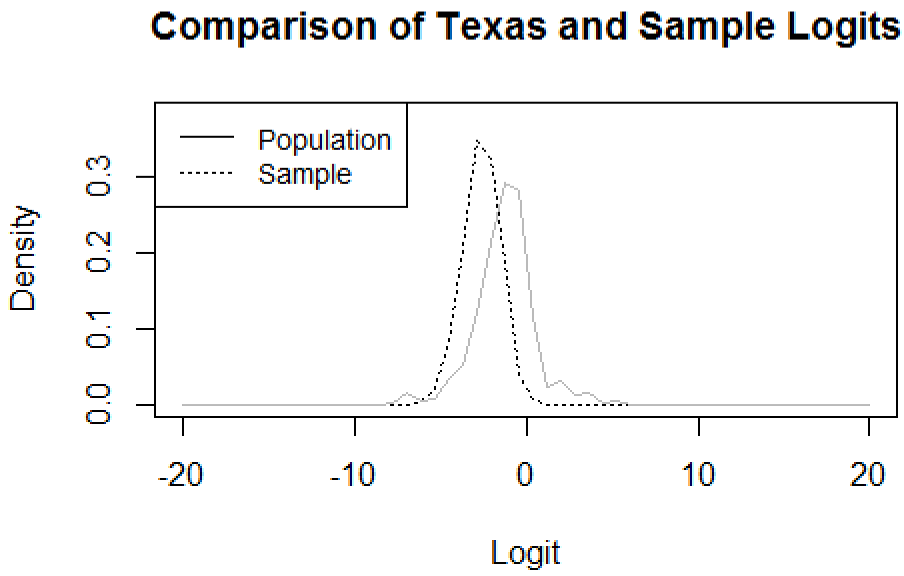
# Appendix C

## Case Studies

In this case study, three states Texas, New Mexico, and Wyoming are presented because they represent high, medium, and low generalizability index values.  First, Texas has a B-index value is 0.824.  From the following tables and figures we see that the values for the population of Texas are very close to that of the population of U.S. elementary schools.  We also see that the comparison of logit distribution is similar to that of the study schools.

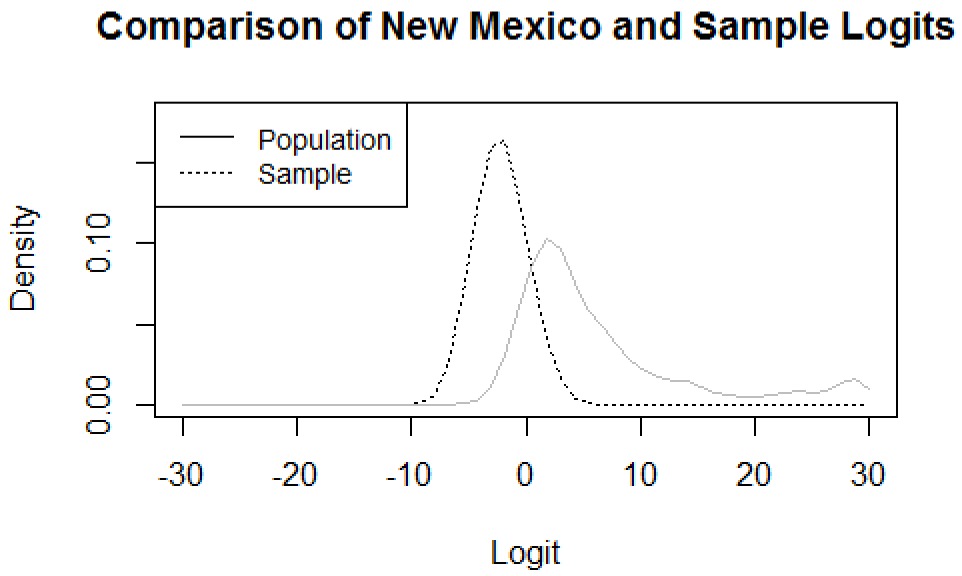
| | **All Schools** | | **Study Schools** | | |
|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | ASMD |
| **School Wide Title I** | 0.84 | 0.37 | 0.66 | 0.47 | **0.490** |
| **Total Students** | 572.65 | 206.98 | 561.06 | 232.98 | 0.056 |
| **% FRL** | 65.3% | 26.94 | 58.0% | 27.21 | **0.270** |
| **Urban** | 0.40 | 0.49 | 0.46 | 0.50 | 0.124 |
| **Suburban** | 0.29 | 0.45 | 0.36 | 0.48 | 0.148 |
| **Town/Rural** | 0.31 | 0.46 | 0.18 | 0.39 | **0.277** |
| **% White** | 30.5% | 27.20 | 35.9% | 30.06 | 0.198 |
| **% Black** | 11.1% | 15.18 | 20.4% | 26.68 | **0.615** |
| **% Hispanic** | 52.7% | 30.47 | 33.9% | 29.09 | **0.616** |
| **% Other** | 5.7% | 7.65 | 9.8% | 11.64 | **0.527** |
| **% Female** | 48.5% | 2.43 | 48.3% | 2.75 | 0.068 |
| **% ELL** | 14.9% | 11.74 | 12.6% | 8.09 | 0.194 |
| **Student/Teacher Ratio** | 15.86 | 2.35 | 16.32 | 3.19 | 0.196 |
| **Total District Schools** | 51.31 | 63.25 | 115.98 | 186.26 | **1.022** |

The logistic regression coefficients are all significant at p<0.05 or better for all coefficients except the proportion of females, and the respective proportions of ethnic groups. As you see from the figure comparing the population of Texas and the study schools there is smaller area of non-overlap.



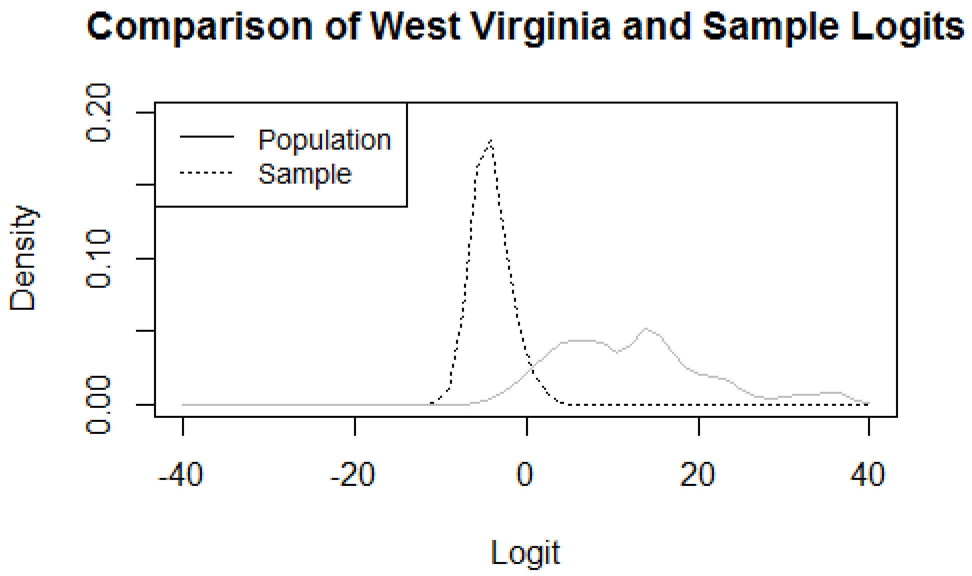
For New Mexico, the state has a B-index value of 0.560, which is considered low generalizability. We see from the figure below that there is less overlap of the propensity score logit distributions as well. The coefficient values for the logistic regression for this state are larger than those for Texas' comparison.

| | **All Schools** | | **Study Schools** | | |
|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | ASMD |
| **School Wide Title I** | 0.93 | 0.26 | 0.74 | 0.44 | **0.739** |
| **Total Students** | 0.91 | 0.29 | 0.66 | 0.47 | **0.863** |
| **% FRL** | 384.9% | 192.16 | 561.1% | 232.98 | **0.917** |
| **Urban** | 74.6 | 23.02 | 58.0 | 27.21 | **0.721** |
| **Suburban** | 0.25 | 0.44 | 0.46 | 0.50 | **0.474** |
| **Town/Rural** | 0.12 | 0.32 | 0.36 | 0.48 | **0.752** |
| **% White** | 0.6% | 0.48 | 0.18% | 0.39 | **0.926** |
| **% Black** | 24.0% | 19.75 | 35.9% | 30.1 | **0.606** |
| **% Hispanic** | 1.6% | 2.02 | 20.4% | 26.7 | **9.336** |
| **% Other** | 61.0% | 26.90 | 33.9% | 29.1 | **1.008** |
| **% Female** | 13.4% | 24.26 | 9.8% | 11.6 | 0.152 |
| **% ELL** | 48.5% | 3.48 | 48.3% | 2.7 | 0.050 |
| **Student/Teacher Ratio** | 15.3 | 10.16 | 12.6 | 8.09 | **0.263** |
| **Total District Schools** | 15.02 | 2.45 | 16.32 | 3.19 | **0.531** |



West Virginia as a state has a B-index value of 0.265, the third lowest for all 50 states. This indicates, and is supported by an almost complete lack of overlap in the plot below. This shows that there are many schools in the population of elementary schools in West Virginia that have no like counterparts in the study sample.

| | All Schools | | Study Schools | | |
|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | ASMD |
| **School Wide Title I** | 0.69 | 0.46 | 0.00 | 0.47 | 0.063 |
| **Total Students** | 318.50 | 159.66 | 139.00 | 232.98 | **1.519** |
| **% FRL** | 60.9 | 13.91 | 0.00 | 27.2 | 0.208 |
| **Urban** | 0.13 | 0.34 | 0.00 | 0.50 | **0.971** |
| **Suburban** | 0.19 | 0.39 | 0.00 | 0.48 | **0.419** |
| **Town/Rural** | 0.68 | 0.47 | 0.00 | 0.39 | **1.055** |
| **% White** | 91.7 | 10.59 | 0.00 | 30.1 | **5.271** |
| **% Black** | 3.8 | 7.02 | 0.00 | 26.7 | **2.375** |
| **% Hispanic** | 1.4 | 2.22 | 0.00 | 29.1 | **14.635** |
| **% Other** | 3.1 | 3.61 | 0.00 | 11.6 | **1.832** |
| **% Female** | 48.2 | 3.37 | 39.88 | 2.7 | 0.034 |
| **% ELL** | 0.6 | 0.74 | 0.00 | 8.1 | **16.275** |
| **Student/Teacher Ratio** | 14.57 | 2.36 | 8.40 | 3.19 | **0.740** |
| **Total District Schools** | 22.62 | 17.48 | 3.00 | 186.26 | **5.341** |



Comparison of West Virginia and Sample Logits