12-2013

Compared to What? The Effectiveness of Synthetic Control Methods for Causal Inference in Educational Assessment

Clay Stephen Johnson
University of Arkansas, Fayetteville

Compared to What?

The Effectiveness of Synthetic Control Methods for Causal Inference in Educational Assessment

Compared to What?
The Effectiveness of Synthetic Control Methods for Causal Inference in Educational Assessment

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Educational Statistics and Research Methods

by

Clay Stephen Johnson
Harding University
Bachelor of Arts in Mathematics, 2002
Teachers College, Columbia University
Master of Arts in Mathematics Education, 2003

December 2013
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

____
Dr. Ronna C. Turner
Dissertation Director

____
Dr. Wen-Juo Lo
Committee Member

____
Dr. George S. Denny
Committee Member

____
Dr. Gary W. Ritter
Committee Member

Abstract

Synthetic control methods are an innovative matching technique first introduced within the economics and political science literature that have begun to find application in educational research as well. Synthetic controls create an aggregate-level, time-series comparison for a single treated unit of interest for causal inference with observational data. However, the strict statistical assumptions associated with matching methods for causal inference raise concerns for unobserved bias related to some data models and availability. The small but increasing set of existing synthetic controls studies with student achievement measures as the outcome of interest suggest that research is warranted into the effectiveness of this methodology in creating unbiased comparisons with necessary sensitivity to detect treatment effects typical of educational interventions. In this study I examined these concerns at an empirical level by analyzing patterns of minimum necessary effects for statistical significance across multiple data models, contrasting covariate specifications, and pools of available comparison units. Data included five years of public elementary school math and reading scores from Measures of Academic Progress (MAP) exams for approximately 35,000 unique students. Using placebo tests for statistical inference as recommended in the literature, I calculated the standardized differences necessary for both a cross-sectional and a cohort model of student progress. Results showed that the addition of demographic covariates provided no additional predictive power over matching on prior MAP achievement alone. Further, near-perfect matches across the pretreatment period were found often enough that a treated unit could not reasonably reach posttreatment effect sizes necessary for detection without also achieving a near-perfect synthetic control match. The placebo tests were sensitive to the increased difficulty of finding close matches when additional pretreatment time points were included, but overall the magnitudes of necessary effects decreased as a result.

Average z-score differences across four pools of comparison units ranged from 0.13 to 0.45 for statistical significance at 5% and from 0.10 to 0.35 for 10% significance. I offer recommendations for using synthetic controls in evaluating educational interventions with student achievement outcomes and for further research into the effectiveness of these methods in reaching conclusions based on unbiased comparisons.

Acknowledgments

I am entirely grateful to the fine folks at NWEA, and especially to John Cronin, for providing me with the Kingsbury Center Data Award that made this study possible. Thanks to Beth Tarasawa, Rebecca Moore, Nate Jensen, Don Draper, and Jed Tai for their help with the data. And thanks to Mike Dahlin, Yeow Meng Thum, Carl Hauser, and Gage Kingsbury for their input that led to my choice of topic.

I am grateful to the University of Arkansas for providing the Distinguished Doctoral Fellowship and graduate assistantship that funded my graduate study. Thank you to the university faculty and community. Thanks to Gary Ritter for his support and guidance. Thanks to Wen-Juo Lo for his advice, knowledge, and friendship. Thanks especially to Ronna Turner for all the time she spent putting up with me, building me up, talking me down, getting me through, and beating me at racquetball.

Thanks to all the friends that have stayed in touch as I went back to school, and to the new ones I made in Fayetteville. Thank you to my family—John, Judy, Sarah, Brad, Sunny, and Coral—for their endless support.

And thank you to those who unknowingly made this project possible: Arsaga’s at the Depot, Mama Carmen’s Espresso Café, Caffe Mela, Jeepers It’s Bagels, Upper Eastside Coffee Company, and Westrock, Peet’s, and Seattle’s Best coffee roasters. Finally, thanks to the late, great Miles Davis (particularly for the years 1967 to 1974, and less so for the ones that followed).

Dedication

To the memory of Dr. George Denny

Whose loss was detrimental to so many, most trivially to this paper

When the roll is called up yonder, you'll be there.

And to Trae Holzman

Who should have been here to celebrate this and many other milestones

Shine on, you crazy diamond.

Table of Contents

Compared to What? The Effectiveness of Synthetic Control Methods for Causal Inference in Educational Assessment

Synthetic control methods are an innovative technique for creating post hoc matched comparisons. First introduced by economist Alberto Abadie with applications in political science research (Abadie & Gardeazabal, 2003), a synthetic control is designed to statistically create a best weighted comparison unit in analyses where random assignment is not possible and where no single appropriate unit of comparison is available. Further, synthetic controls depend on aggregate measures of time series data. Since data of this type are often available in education research when student-level measures are not, the potential utility of synthetic controls for educational program evaluations is apparent.

As user-friendly software options for running synthetic controls have been made readily available (Abadie, Diamond, & Hainmueller, 2011; Hainmueller, Abadie, & Diamond, 2010), applications of the technique have appeared with increasing frequency within the education research literature (e.g., Bassok, Fitzpatrick, & Loeb, 2012; Belot & Vandenberghe, 2009; Fitzpatrick, 2008; Hudson, 2010; Klasik, 2013). What remains to be explored formally is the effectiveness of synthetic controls methods to produce *as if random* comparisons when applied alongside the particular challenges of latent variable measurement error and the availability of influential covariate measures within student achievement and school evaluations.

Synthetic controls are formed by optimizing the reweighted combination of comparison units from an available donor pool. The resulting weights sum to one to form the closest match for the treated unit across the time period prior to treatment. For the posttreatment period, this reweighted synthetic unit represents the potential outcome: what would have been observed for the treatment unit had it instead gone untreated. Under the strict assumptions for analyses of causal inference (e.g.,Imbens & Wooldridge, 2009; Pearl, 2000; Rubin, 2006; Shadish, 2010), the synthetic control meets the criteria for a potential outcome and can be treated *as if random*, where further analyses are performed as though treatment had been randomly assigned. The reasonableness of meeting these statistical assumptions, some of which are untestable in empirical settings, is subject to continued debate (e.g.,Gelman, 2009; Pearl, 2009; Sekhon, 2009). This study addresses this debate on empirical terms, where real academic data are analyzed to observe patterns of behavior when forming synthetic controls in attempt to determine whether student achievement measures are capable of being matched sufficiently to allow for the detection of meaningful levels of student academic progress.

Synth package for Stata (originally for R; Abadie et al., 2011) locates the best weighted combination of available comparison units using the mathematical optimization technique of constrained quadratic programming to minimize the difference between treated and synthetic control over the pretreatment period. By default the software takes into account the relative predictive power of each included covariate in relation to the outcome measure (math or reading achievement score, in this case). Since the resulting synthetic control serves as a single unit however, no variability is present for calculating traditional inferential statistics. For this reason a type of falsification exercise is recommended to compare the matching of the treated unit of interest among synthetic control matches for comparison units where no treatment was delivered (Abadie, Diamond, & Hainmueller, 2010). Called *placebo tests* in the literature, these create a unique synthetic control for every available unit in the comparison pool as a placebo stand-in for the true treatment unit. The resulting distribution of placebo synthetic units serves as an empirical, nonparametric distribution for calculating exact statistical significance.

The root mean square prediction error (RMSPE) is used as the recommended measure of goodness of fit between treated and synthetic control. This allows for the calculation of statistical significance, determined by comparison of the ratio of posttreatment RMSPE to pretreatment RMSPE, which takes into account both the goodness of match at pretreatment and the size of effect over posttreatment. In this study, I examined these distributions of ratios across 16 data specifications and two independent sets of MAP exam data from Northwest Evaluation Association to allow for comparison of minimum ratios necessary for detection of effects of a hypothetical treatment administered to a group of students at given levels of statistical significance.

Since the placebo test methods recommended by Abadie (2010) for statistical significance require the treated unit to demonstrate a ratio of post to pretreatment RMSPE larger than $(1 - \alpha) \times 100\%$ of placebo units forming the comparison donor pool, I analyzed observed aggregates of student test score units as placebos for comparison with some student group who might receive treatment. This was possible since the actual unit receiving treatment plays no part in the placebo test analyses for inference. My results offer observed cutoff levels of post/pretreatment ratios that would be required for a treated school's difference to be detected as statistically significant among the available pool of comparison schools using the synthetic controls matching method.

Chapter I: Introduction

Sir Ronald Fisher wrote in *The Design of Experiments*, “it may be said that the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged” (1935, p. 24). His argument was that it is not the ideal of perfect, error-free measurements, units, or observations that validates statistical conclusions, but that purposeful random assignment of subjects to treatment and control creates comparable groups for accurate statistical inferences. With Fisher’s publications, the standard of randomization became a fundamental element of the traditions of applied statistical methodology (Maxwell & Delaney, 2004).

Fisher began his statistical discussions of experimental treatment assignment with examples from agriculture and with a narrative about a lady tasting tea. However, randomization in educational research and other social sciences is rarely so straightforward. Often in the analysis of student achievement, the researcher has access only to nonexperimental, observational data. He or she seldom has the authority to assign treatments at the design stage. In assessing student academic performance, where the task is forming a valid counterfactual comparison group in post hoc analysis, the essential question for a meaningful outcome measure is then, *performance compared to what?*

Even as the national pool of student achievement data continues to grow, random assignment studies in educational research remain relatively rare. The U.S. Department of Education formed the What Works Clearinghouse as a database of curriculum and program evaluations in response to the standard of *scientifically based research* required by the No Child Left Behind Act of 2001. For research submitted to the What Works Clearinghouse to receive the highest rating, *meets evidence standards without reservations*, studies must be designed around randomized controlled trials (Institute of Education Sciences, 2011). But of the 6,687 studies reviewed as of the beginning of 2013, only 221 have met this criterion. Despite repeated pleas for random assignment studies in research in education (e.g., Cook, 2002)—and for reasons that may be political, ethical, or practical—alternatives to randomized experimental designs continue to be necessary for analyzing student educational outcomes.

Quasi-experimental research designs offer an expanding field of possible alternatives for forming causal inferences. Rather than relying on randomization to equally distribute all extraneous covariates between treatment and control groups, quasi-experimental designs employ some post hoc statistical adjustment to create a valid comparison for the treatment group of interest (Shadish, Cook, & Campbell, 2002). The What Works Clearinghouse online glossary states that "for a quasi-experimental design to be rigorous, the intervention and comparison groups must be similar, demonstrating baseline equivalence on observed characteristics, before the intervention is started" ("Quasi-experimental design," n.d.). The continuing challenge for educational statisticians and other nonexperimental researchers is to make the best use of observational data through rigorous quasi-experimental designs. As Cook, Shadish, and Wong stated in their comparison of within-study differences in experimental and nonexperimental outcomes, "alternatives to the experiment will always be needed, and a key issue is to identify which kinds of observational studies are most likely to generate unbiased results" (2008, p. 725).

Quasi-experimental and nonexperimental research designs serve as examples of an observational study, which is an empirical investigation "of treatments, policies, or exposures and the effects they cause, but it differs from an experiment in that the investigator cannot control the assignment of the treatments to subjects" (Rosenbaum, 2002, p. vii). When data are observational rather than collected via randomization, confounding differences between the group receiving treatment and the untreated comparison units have not been balanced away through random assignment. The consequence is selection bias; calculating unbiased effects requires post hoc statistical control. Post hoc matching methods offer one set of options, where the counterfactual for comparison is formed statistically after treatment has taken place rather than in advance of treatment through systematic assignment. Various matching methods have been introduced and applied across the quantitative research disciplines, and continual advances in statistical sophistication and computational power have furthered their popularity.

The matching method of synthetic controls was formally presented by Harvard economist Alberto Abadie with political scientists Alexis Diamond and Jens Hainmueller in 2010 and has since begun to gain popularity quickly in observational studies. Synthetic controls offer intuitive appeal in their interpretation of results and transparency in pretreatment equivalence for researchers. This has led to their recent appearance in a wide variety of applied research literature. Since the introduction of easily accessible, user-friendly, free software packages for running synthetic controls matching (Abadie et al., 2011), the method has received impressive levels of attention in the brief period of time since its introduction. Of particular interest to the purposes of educational research, synthetic controls are matched on aggregate-level rather than individual-level measures and capitalize on repeated observations over time. As educational achievement data are often made accessible as measures at the district-, school-, or grade-level while individual scores are subject to privacy law restrictions, these methods form post hoc, aggregate-level comparison groups useful for the types of research questions that are encountered in education. Consequently, economists who study educational outcomes have begun to apply synthetic controls in analyzing student achievement data as well.

The effectiveness of the synthetic control method in practice is not without question however. As with any matching method applied to observational data, synthetic controls require a strict set of statistical assumptions, some of which cannot be tested empirically. For unbiased estimation, no essential covariates can be omitted in the matching process. The issues of omitted variable biases in post hoc matching are well documented in a line of research whose foundations are often attributed to Harvard statistician Donald Rubin. In summary, these methodologists have recommended that all accessible covariates be included in forming accurate matches, and that a large enough set of covariate measures allows for adequate reduction of significant bias due to any variables that remain omitted (e.g., Rubin, 2006; Sekhon, 2008; Stuart & Rubin, 2008a). Judea Pearl at UCLA—along with other researchers who prefer to approach causal inference from theories based on structural models—has documented concern for critical violations of matching assumptions by the incorrect inclusion of covariates as well. Like-minded statisticians have theorized that the covariates commonly available to social scientists are capable of inflating rather than decreasing bias when the causal directional paths among influential variables are inappropriate (e.g., Gelman, 2009; Pearl, 2009, 2011; van der Laan & Rose, 2011). Some have expressed such extreme concern as to all but suggest that these problematic, underlying causal paths are so common in applied contexts that causal inferences via post hoc matching should be avoided in practice.

In applying synthetic controls within political science scenarios, where a multitude of control variables are available and where classical measurement error in essential economic outcomes can be arguably negligible, overt biases in resulting estimates may be evident to the analyst on sight. For example, if a political scientist forms matched comparison groups of nations whose qualities are clearly unlike the nation of interest, then the need for respecifying the matching model is evident. What remains to be demonstrated, and what I investigated in this study, is the behavior of the synthetic control method when applied within the unique context of educational achievement data. When the outcome variable of interest is a latent construct with inherent measurement error, and when matching depends on a limited set of available demographic covariates, the sensitivity and power of synthetic controls to detect differences between groups may be manifest differently than when used in economic data applications.

In this study I investigated the behavior of synthetic control matches within the context of educational assessment data to establish baseline patterns of statistical error and to address the robustness of these methods for analyzing student achievement. I used aggregated student test scores from the MAP exams alongside demographic covariates to empirically determine minimum detectible effects in terms of group academic performance. By comparing sets of placebo tests as recommended for inference with synthetic controls by Abadie et al. (2012), I investigated the method's baseline patterns of sensitivity to various model specifications and multiple sets of control unit data. In this way I established empirically predicted magnitudes of effect size for chosen Type I error rates to inform the effectiveness of synthetic controls for describing, comparing, and predicting student achievement.

Using a large-scale set of student data with MAP scores aggregated at either the school- or cohort-level as the outcome of interest, I addressed the following set of research questions:

(a) What sizes of student achievement gains are required for a treatment unit to be identified as having statistically significant differences compared to its synthetic control match?

(b) How sensitive are synthetic controls to respecification of matching covariates and data availability in an academic achievement measures context?

(c) Are patterns of improved fit in synthetic control matching consistent with what is expected from the inclusion of available academic variables?

(d) What data requirements are necessary for consistency across respecifications of synthetic controls at desirable levels of statistical power?

My analysis of the synthetic control method adds to the current body of literature in quantitative educational research by investigating the effectiveness of a quasi-experimental design from political science and economics that has promising applicability across the social sciences. As academic achievement data often are observational rather than experimental in nature, their potential use for exploring new applications of quasi-experimental methods continues to grow. Synthetic controls offer not only an interesting alternative when random assignment to treatment and control is impossible; these methods offer potential benefits to forming customized student norming groups, as when comparison to an entire pool of available student data is impossible or undesirable. A matched synthetic control group as a customized norm could offer an intuitive way to assess groups of students who are observably dissimilar to available comparison students and whose achievement and demographic data are available only at the aggregate level. Despite the initial appeal of these methods however, further formal investigation of their behavior in the context of student achievement measures is warranted.

Chapter II: Background

The synthetic control method first appeared within the economics literature with Abadie and Gardeazabal (2003) as a means for calculating the causal effect of terrorist activity on economic outcomes. Synthetic controls and other post hoc matching methods, along with all quasi-experimental designs for use in observational studies, fit within a theoretical framework of statistical models for forming causal inferences. For this reason, my discussion providing background to the method of synthetic controls begins with the foundations of the estimation of causal effects.

Causal Inference

The language of causation has begun to gain a foothold in quantitative research during the previous few decades, but causality remains a contentious subject within the field of statistics, which was founded on correlational rather than causal relationships (Pearl, 2000). It is beyond the scope of my study to summarize the centuries of philosophical debate surrounding cause and effect. Brady (2008) presented a concise summary of these philosophies as they relate to statistical methods. Several options for representing causal relationships are available, and some methodologists are critical of the limitations of alternative models (e.g., Pearl, 2009). Structural equation modeling, for example, initially gained popularity in the social sciences within psychological measurement and econometrics applications. These models offer an alternative framework for cause and effect by including the uncertainty due to measurement error into path analysis.

Across the disciplines, the most frequently cited statistical model for causal relationships, and the one underlying the current matching literature in economics, is commonly attributed to Donald Rubin (1974) as the *Rubin causal model* or the *Neyman-Rubin model of causal inference*

(Sekhon, 2008). This framework offers a way of representing treatment effects in terms of potential outcomes. Though others have framed Rubin's concerns with potential outcomes as differing from analyses of counterfactuals (e.g., Shadish, 2010), Brady defined causal inferences in counterfactual terms and posited that

> the fundamental problem of counterfactual definitions of causation is the tension between finding a suitable definition of causation that controls for confounding effects and finding a suitable way of detecting causation given the impossibility of getting perfect counterfactual worlds. (2008, p. 251)

Sekhon presented the Rubin framework as

> a nonparametric model where each unit has two potential outcomes, one if the unit is treated and the other if untreated. A causal effect is defined as the difference between the two potential outcomes, but only one of the two potential outcomes is observed. (2008, p. 273)

The reality that only one of the two outcomes present in the Rubin model ever has the possibility of being measured has sometimes been called *the fundamental problem of causal inference* (Holland, 1986). Once treatment has been administered, every unit has permanently been assigned to represent an outcome of either the treated group or the untreated group. This makes causal inference the pursuit of a sort of *Bill and Ted's excellent treatment effect*. In the film (Kroopf, Murphey, Soisson, & Herek, 1989) time travel made it possible to manipulate history and observe the alternative effects in the future (i.e. *potential outcomes*), but as with the film (whose plot's logical flaws are also beyond the scope of this study) the possibility of observing outcomes that are only measurable in an alternate reality is also confined to the realm of fiction.

As half of the outcome measures of interest are never observable, any methodology used to estimate a causal effect is often presented as an attempt to solve a problem of missing data.

Holland (1986) categorized two types of solutions to the fundamental problem of causal inference: the scientific and the statistical. The scientific solution is ubiquitous in research and typically goes unspoken, as when a piece of lab equipment is assumed to measure consistently over time due to calibration. When a lab measurement at Time 1 is compared to another at Time 2 following some research manipulation, the first measure is substituted for the missing, potential-outcome state of the treated unit had it instead gone untreated. Alternatively, Holland's (1986) discussion of the statistical solution to the fundamental problem of causal inference included any method for using mathematical expectation on observed units to replace data missing due to unobservability. He algebraically formalized the Rubin model this way: the causal effect $Y_t(u) – Y_c(u)$ is the difference between the outcome measure Y for the single unit u under its treated state and its untreated (control) state. Due to the fundamental problem of causal inference, only one of these terms is observable for any unit u, and therefore the other must be imputed. While the magnitude of this true causal effect is uncertain, the statistical solution that the Rubin model provides comes in the form of the average treatment effect T, where $T = E(Y_t) – E(Y_c)$. Given a strict set of statistical assumptions, this use of expected values to estimate the outcome measure "reveals that information on *different* units that *can be observed* can be used to gain knowledge about T" (Holland, 1986, p. 947).

The Rubin model is not a philosophy of causation but a formal statistical model that unifies all research designs for drawing statistical inference (Shadish, 2010). The traditional technique of random assignment promoted by Fisher (1935) relates to Rubin's model by creating an even distribution of observed covariates and unobservable confounds between groups to assure that the randomized control can act as a valid stand-in for the potential outcome. Nonexperimental models and quasi-experimental methods must rely on alternative statistical procedures to find a comparable counterfactual in the absence of simple randomization. The task in calculating accurate causal effects, whether in a case with random assignment or without, lies in the best choice of a counterfactual for comparison.

Observational Data and Quasi-Experiments

In nonexperimental studies, as are common in education research and in the social sciences in general, data are observational in nature rather than the results of treatment assignments under a researcher's control. This means that efforts at drawing causal inferences require post hoc statistical adjustment of the available observations in order to support the validity of their use as estimates of potential outcomes. Due to the nonexperimental nature of the assignment process that generates observational data, observational studies are built around a set of methods that are concerned with overcoming selection biases (Shadish et al., 2002). Regardless of the size of the set of available data, the lack of random assignment in the data generating process leads to bias that interferes with correct inference (Rosenbaum, 2002). Therefore the goal for appropriate analyses of observational data is to select a valid comparison group that was similar to the group of interest across the time period before treatment was administered. If this choice of comparison group can be shown to avoid violation of necessary statistical assumptions, then this group can be further analyzed as if it were a control group formed through random selection.

Cook (2008) put forward an often-ignored distinction between *quasi-experiment* and *nonexperiment*. He suggested that observational data—such as those of interest for this study—belong within the realm of nonexperiment rather than quasi-experiment. Cook admitted that the distinction is not universally acknowledged, but that a nonexperimental design implies the analyst has little or no influence over the data generating process but only access to post hoc research methods. In a quasi-experiment however, the researcher has more influence on treatment assignment and the methods for data collection, but these include some sampling alternative to random assignment (Cook et al., 2008). Comparisons of student performance using educational assessment measures are commonly referenced as *quasi-experimental* and draw on these methods of observational studies, as the available pool of data rarely represents either a whole population or a randomized sample. Experimental research is not absent in the field of education, as these type of program evaluations and curriculum studies have been cataloged and promoted in the US Department of Education’s What Works Clearinghouse (2011). But often in analyzing academic achievement data, interpretations of student progress are required whether any formal intervention has taken place or not. The classroom is like a laboratory where no student goes without some treatment, even if the treatment is *business as usual* (Jaciw & Newman, 2011).

Analysis of covariance, regression adjustment, interrupted time series, and structural equation modeling are examples of quasi-experimental methods that are currently commonplace in psychological and educational statistics, while two-stage instrumental variables regression, regression discontinuity, and difference-in-differences designs have gained popularity in economics (Shadish et al., 2002). Synthetic controls, among other matching methods such as propensity score estimation (Rosenbaum & Rubin, 1983), extended caliper matching (Stuart & Rubin, 2008b), and genetic matching (Diamond & Sekhon, 2013), offer a set of nonexperimental or quasi-experimental alternatives that have continued to grow in popularity across the social science disciplines.

Post Hoc Matching Methods

Matching methods for extracting valid causal effects from observational data have been increasing in influence across the disciplines, and the fields of economics and educational statistics are no exception (Stuart, 2010). As either an alternative to or an extension of common quasi-experimental designs, matching methods directly address the missing-data concept of Rubin's causal model by assuming that all untreated potential outcomes were unobserved. In order to select counterfactuals that can be assumed to behave *as if random*, post hoc matching imputes observations best representative of a potential outcome after either reweighting or partial deletion of available data. In the absence of the process of randomization that would have equally distributed all observed and unobserved covariates between groups, the key goal for finding viable post hoc matches is in creating covariate balance between groups over the time period before treatment.

Stuart gave a concise set of steps for implementing any matching process:

> 1. Defining "closeness:" the distance measure used to determine whether an individual is a good match for another.
> 2. Implementing a matching method, given that measure of closeness.
> 3. Assessing the quality of the resulting matched samples, and perhaps iterating with steps 1 and 2 until well-matched samples result.
> 4. Analysis of the outcome and estimation of the treatment effect, given the matching done in step 3. (2010, pp. 4–5)

For most methods of matching, each of Stuart's four steps requires that the researcher select a choice among many possible options. Each of these choices opens the researcher to criticisms concerning subjectivity and data fishing. The current methodological literature is rife with recommendations for each step of the matching process under particular conditions, such as with measures of covariate distance (e.g., Mueser, Troske, & Gorislavsky, 2007), choice of matching estimator (e.g., Zhao, 2004), testing for covariate balance (e.g., Hansen & Bowers, 2008), and assessing statistical significance of outcomes (e.g., Abadie & Imbens, 2008). Recent developments in matching methods—synthetic controls being one example—seek to improve on previous techniques by combining some of the four steps into a single process or by empirically automating the subjective choices for the researcher. I compare the beneficial features of synthetic controls among some other matching method alternatives in a section that follows.

Statistical assumptions for matching

One critical assumption underlying the Rubin causal model, and therefore all matching methods as well, is most commonly referred to as the *stable unit treatment value assumption* (SUTVA). As formally defined by Rubin (1978), SUTVA is a stricter case of the basic inferential statistical assumption of independence among observations. Rubin's model further requires a single, fixed value of the treatment effect, meaning there can be no multiple levels of outcomes caused by interaction among units. Shadish (2010) gave a first-hand account of the common difficulty of defending SUTVA in empirical applications, where the editorial review of an earlier matching article led to "several futile rounds of trying to respond" due to concerns that SUTVA was being violated. Fortunately for the current methods of my study, Imbens and Wooldridge (2009) explained that analyses involving measures of aggregated units make assumptions of non-interference more plausible in practice by essentially clustering measurements among units assumed to interact.

The second key assumption for causal inference is *strong ignorability* (Rubin, 2006). This terminology subsumes two concepts sometimes discussed as separate data assumption ideas: the assumption of unconfoundedness and the assumption of sufficient covariate overlap. The unconfoundedness assumption, sometimes equivalently called *selection on observables, exogeneity,* or *conditional independence*, means that “there are no unobserved factors that are associated both with the assignment and with the potential outcomes,” (Imbens & Wooldridge, 2009, p. 23). This assumption is statistically untestable in empirical analyses, as presented in my favorite-titled chapter section of Judea Pearl’s: “Why There Is No Statistical Test for Confounding, Why Many Think There Is, and Why They Are Almost Right” (2000, p. 182). The problem with testing statistically for violation of this assumption lies in the presence of unobservable confounders, to which Pearl further included the additional (and for some, controversial) problem of unknowable causal relationships among confounders. Overlap, or common support—the other half of the strong ignorability assumption—requires that sufficient data are available for matching the units of interest across the full distribution of all covariates (Stuart & Rubin, 2008a). Sekhon (2009) described the strong ignorability assumption as too often overlooked in the applied literature. He made a call for the explicit statement of what is being estimated on the part of the researcher by posing the causal effect in terms of the outcome of interest if the study had instead been performed as a randomized experiment.

Concerns and criticisms

Sekhon also criticized social scientists for their “fascination with the latest estimator” at the expense of addressing the foundational methodological assumptions of matching (2009, p. 487). He went on to describe his concern for the frequent disconnection between a matching model and the assumed causal effect intended to be replicated. He emphasized the importance of precise identification of the experimental causal condition being modeled during the design phase of a matching study to assure the correct interpretation of results.

In addition to concerns for violation of statistical assumptions by the omission of important covariates (i.e. omitted variable bias), concerns for bias inflation caused by the inclusion of inappropriate control variables are discussed at least as frequently in the literature. Cautions against covariates whose inclusion causes bias inflation involve matching methods that are not *equal percent bias reducing*, that is, when covariate values deviate from an ellipsoidally symmetric distribution (e.g., Rubin, 2006; Sekhon, 2008). Van der Laan and Rose called the effects of this violation *z-bias*, saying the issue is akin to choosing a covariate that more truly behaves as an instrumental variable, one that is "unrelated to the outcome but related to treatment…If the relationships between the variables are linear, bias will always be increased" (2011, p. 344). Less universally agreed upon is a concern with bias amplification due to the underlying and unknown structural relationship among the covariates selected for inclusion (e.g., Pearl, 2011). Pearl referred to this phenomenon as *M-bias*, where an observed covariate shares causal relationships with two independent variables such that controlling on the covariate reverses the true causal direction between variables and inflates bias. Pearl likes to quote Rubin's response after being confronted over potential violations of the ignorability assumption due to M-bias, to which Rubin replied that "to avoid conditioning on some observed covariates... is nonscientific ad hockery" (Pearl, 2009, p. 2). The debates continue regarding covariates' role in increasing bias, with Pearl's supporters gaining ground using structural representations and with Rubin's followers holding fast to the notation of conditional probabilities (Gelman, 2009).

In practice, the methodology of post hoc matching applications is often challenged by concerns over violation of one or all of these critical statistical assumptions. As some of these assumptions are also impossible to test statistically, arguments are necessarily posed as a theoretical argument rather than implied from empirical evidence (Fortson, Verbitsky-Savitz, Kopa, & Gleason, 2012). Beyond concerns for potential model violations, the validity of matching methods is at the mercy of the researcher, whose selection of covariates, setting of caliper sizes for measuring closeness, tolerance for unmatched units, and method for testing covariate balance—among other subjective choices—determine whether bias has been reduced to a sufficient level for estimating reliable outcomes. Although researchers have been cautioned against gaining access to outcome measures before these matching choices have been finalized, analysts using observational data are easily accused of data fishing, since subjective methodological choices might be made just as easily to maximize posttreatment impacts as to optimize pretreatment covariate balance. Synthetic controls methodology offers an alternative that seems promising in addressing several of these common concerns for violation of matching model assumptions while introducing additional factors in need of empirical investigation.

Synthetic Control Methods

The synthetic control method is a post hoc matching technique that forms a best weighted combination of comparison units measured repeatedly over time for use with a single, aggregate treatment unit. Abadie and Gardeazabal (2003) first introduced synthetic controls in an analysis of the economic effects of terrorism in the Basque region of Spain. They selected a weighted combination of nearby political regions—none of which provided a logical counterfactual on its own—to represent the economic outcomes of the Basque Country had it not been subject to terrorist activity. The method was further formalized by Abadie, Diamond, and Hainmueller (2010) in the context of creating a synthetic California for analyzing the impacts of statewide passage of tobacco regulation legislation. In this study they used data from a weighted combination of other states without similar tobacco regulations to represent the potential outcomes of California had the legislation never been enacted. Synthetic controls were presented and further extended for the broader audience of the political science community through a demonstrative analysis of the economic results of the political reunification of East and West Germany (Abadie et al., 2012). Here they selected other European nations whose economic patterns were not influenced by the 1990 German reunification to form a synthetic control group for estimating the financial impacts of the falling of the Berlin Wall.

Since their introduction in 2003, and increasingly since the availability of an early National Bureau of Economic Research working paper (Abadie, Diamond, & Hainmueller, 2007), appearances of the synthetic controls methods have skyrocketed within the economics and political science literatures. Whether or not the surge of popularity in their application is in line with Sekhon's (2009) criticism of fascination with the latest matching estimator, the increasing appearance of the technique in the applied literature is undeniable. A recent search of Google Scholar gives over 200 instances of studies implementing synthetic controls in applied research (e.g., Almer & Winkler, 2011; Coffman & Noy, 2012; Eren & Ozbeklik, 2011; Hinrichs, 2012), with new citations added to this collection on a nearly weekly basis. Several other examples of synthetic controls applications in unpublished job papers or dissertations in progress are also available via Internet search. In a presentation to the 2007 Summer Institute of the National Bureau of Economic Research, Wooldridge (2007) included the synthetic control method in his discussion of recent advances in difference-in-differences estimation. Similarly Imbens and Wooldridge treated synthetic controls as a "very interesting alternative approach to the setting with multiple control groups" within difference-in-differences methods (Imbens & Wooldridge, 2009, p. 72). The 2012 annual meeting of the American Economic Association even assigned applied researchers using synthetic controls to their own panel discussion (Hoxby, 2012). At the same time, some researchers have begun to apply the methods within educational research settings.

Synthetic controls in education research

A few notable instances of synthetic controls applied to achievement data have begun to appear, with several applications of the technique currently available on the Internet in unpublished white papers and academic working papers. An early appearance in educational research by Fitzpatrick (2008) cited the initial working-paper version of Abadie et al. (2007). Fitzpatrick performed a large-scale study of Georgia prekindergarten programs compared to synthetic controls, but only in an analysis supplementary to a difference-in-differences regression analysis, and basing statistical inference on a distribution of ratios whose denominators were essentially no different from zero. She defended her choice of synthetic controls as a secondary analysis by arguing that Georgia was noticeably different from the comparison group of all other states over the pretreatment period. She implemented placebo tests as suggested by Abadie et al. and concluded that Georgia's results were not statistically significant beyond chance, although her difference-in-differences methodology suggested otherwise. It is unclear whether her access to only two pretreatment time points for matching may have contributed to her inconsistent findings. Further, she applied synthetic control weights to a student-level analysis in a way that may be inconsistent with the design purposes intended by Abadie et al. (2010).

To examine the implementation of a new grade retention policy in French-speaking Belgium, Belot and Bandenberghe (2009) used an analysis of synthetic controls for comparison. They used scores from the international PISA exam to assess the results of the introduction of a provision for holding students back at Grade 7 or 8. They too used two time points for pretreatment matching within three periods of available PISA data. For producing inferential statistics, instead of falsification tests they used individual-level data to test for statistical significance between group means. Their choice would suggest that Abadie et al. (2010) recommended placebo tests for inference only because they had no access to individual observations. It remains unclear whether Belot and Bandenberghe's use of disaggregated inferential tests appropriately applies to their choice of synthetic controls for national-level matching.

In an extensive and detailed analysis of performance pay for teachers, Hudson (2010) used synthetic control matches for each school within its own state and across four alternate specifications. To find the effects of the nationwide Teacher Advancement Program on normalized achievement scores, she then applied difference-in-differences estimation across all participating schools using their calculated synthetic controls weights. She further examined goodness-of-fit between her participants and synthetic controls based on pretreatment performance levels, and she concluded that achievement gains in math were statistically significant and near 0.15 standard deviations in effect. However, like Belot and Bandenberghe (2009), Hudson did not make use of placebo tests for falsification inference as presented in the synthetic controls literature.

In another study of preschool programs, Bassok, Fitzpatrick, and Loeb (2012) analyzed the outcomes of states' universal-preschool policies by comparison to a synthetic Oklahoma and Georgia in contrast to a comparison to all available states and to all Southern states. They performed inferential statistics using both difference-in-differences estimation and placebo tests for synthetic controls to determine that both states experienced increases in the availability of preschool institutions after implementation of the policy.

Synthetic controls have now begun to appear in the methodologies of education research journals as well. A most recent example is Klasik (2013), whose previous working-paper version is one of the Web-accessible applications of synthetic controls mentioned above. Klasik employed synthetic controls to investigate the effects of statewide policies for adding college entrance exam requirements. He statistically formed synthetic controls for Colorado, Illinois, and Maine before running fixed effects regression using weights found using synthetic control methods. Using placebo tests to determine statistical significance as recommended by Abadie et al. (2010), Klasik concluded that significant changes in student sorting among in-state college types resulted from states' adoption of mandatory entrance exams. Further, he concluded that student performance on entrance exams was not improved as a result of these policies.

Synthetic controls as quasi-experiment

Within the literature on observational studies and quasi-experimental methods, synthetic controls have been variously presented as a comparative case study design (Abadie & Gardeazabal, 2003), an innovation in difference-in-differences estimation methods (Wooldridge, 2007), a type of interrupted time series design (Betts et al., 2010; Shadish et al., 2002), and as a systematic quantitative method for qualitative case study research (Abadie et al., 2012). Appropriate for use with repeated-measures or longitudinal data sets in the social sciences as well as in economics for panel data analyses, synthetic controls have potential for even broader research application.

With an increasing number of options for post hoc matching methods available (such as exact matching, Mahalanobis matching, propensity matching, genetic matching, etc.), and in light of criticism that researchers are too quick to adopt the latest technique unexamined (Sekhon, 2009), a summary of benefits of the synthetic control method is warranted. Although its weaknesses are the same as for other alternatives for post hoc matching, in summary, the strengths of synthetic control methods are in their systematic combination of matching and measuring covariate balance into a single step, their direct assessment of closeness of fit prior to estimating outcomes, and in their capability of handling data measured at an aggregate level.

External validity is a concern within some regression applications, where linear extrapolation beyond the measured range of available data can lead to inaccurate results within the true range of interest. Besides concern for other common violations of linear regression assumptions, Sekhon (2009) posited that regression is relied upon too often for this reason, when external validity is incorrectly given minimal attention in exchange for emphasis on internal validity. By restricting synthetic control weights to positive values no greater than $w* = 1$, this method offers protection from building counterfactuals with extreme values outside the support of the observed data. For the same reason, synthetic controls offer improvement over concerns posed by Stuart (2007) over traditional interrupted time series designs, where observed longitudinal data are used to model a prediction for an unobserved outcome. Synthetic controls avoid extrapolation over the posttreatment period by relying instead on the weighting of actual posttreatment observations. Regression discontinuity analysis, another quasi-experimental design option commonly criticized for its narrow window of generalizability, is likewise improved upon by the synthetic control method, whose pre- and post-intervention periods are not restricted by the linearity assumptions that are often problematic at the extremes for regression discontinuity.

Other problems frequently encountered in matching methods come from a discrepancy between the level of treatment, the level of matching, and the level of analysis. Abadie suggested that best practice is to perform measurement and analysis at the same level as treatment assignment in order to avoid cluster effects, where violations of statistical independence lead to biased estimates (Abadie et al., 2010). He went on to state that a frequent methodological flaw is in forming individual-level matches for use with an outcome of interest measured at an aggregate level. Additionally, data availability sometimes leads analysts to match at one level but use measurements performed at another, as when school-level rates of race or socioeconomic status are substituted for unavailable student-level demographic variables.

The availability of student-level data for education research is often limited, while aggregate data are often accessible to the public. Researchers often prefer individual-level measures for increased variability however, which enable more methodological flexibility. There are cases where aggregate data may be preferable though, as Cook et al. (2008) suggested that matching at an aggregate level can be expected to reduce bias due to violations of unit-level independence. Further, Millimet (2011) stated that classical measurement error of the type that must be considered in measuring academic achievement can be expected to be reduced when individual scores are aggregated before analysis.

Formal presentation of synthetic controls matching

The synthetic control method addresses causal inference by using the observed pretreatment measures of a single aggregated unit to find the closest weighted match from among the available pool of comparison units when no single comparison unit is an obvious choice for representing the counterfactual case. As was notated in Abadie et al. (2010), let **W** represent a vector of synthetic control weights of dimensions ($J \times 1$) that is held under two constraints: $w_j \geq 0$ for all individual weights (i.e. no weight is negative) and $w_2 + \ldots + w_{J+1} = 1$ (i.e. all weights sum to 1). Next let $\mathbf{Z}_i$ represent a ($r \times 1$) vector of observed covariates that are unaffected by the treatment, and let Y_{1t} be the observed outcome measure for the treated unit at time point t. All units from *j* = 2 through *j* = *J* + 1 serve as the donor pool of potential comparison units. Then the synthetic control method seeks to calculate each best weight w_j* such that

$$\sum_{j=2}^{J+1} w_j^* Y_{jt} = Y_{1t} \tag{1}$$

for all time periods *t* prior to introduction of the treatment, and

$$\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j = \boldsymbol{Z}_1 \tag{2}$$

for all included covariates. The resulting synthetic control weights are those that combine the units of the comparison donor pool to give an equivalent set of observed covariates and an equivalent outcome measure for every pretreatment time point. So far as weights are unavailable to result in exact equality between these treated and weighted comparison units, optimal weights are produced that solve Equations 1 and 2 approximately.

The root mean square prediction error (RMSPE) is the measure of fit that serves to assess the closeness of match between the treated unit and the synthetic control unit. For the period of time prior to treatment, it was algebraically presented in Abadie et al. (2012) as

$$RMSPE = \sqrt{\frac{1}{T_0} \sum_{t=1}^{T_0} (Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt})^2}, \tag{3}$$

which is the averaged difference between the outcome measure of the treatment and the weighted mean outcome measure of the synthetic control over all time points up to T_0, the final time point prior to treatment. As much as the RMSPE can be reduced over the pretreatment period, the researcher’s goal is to add pretreatment time points, select observable covariates, and remove weak comparison units to find the closest matched synthetic control unit.

Software packages for implementing synthetic controls

Perhaps helping to spur the popularity of synthetic controls applications in the short period of time since their introduction, multiple software packages have been made available for public access via the Internet. Abadie, Diamond, and Hainmueller (2011) gave a detailed presentation of the package available for use with R software using their original Basque Country analysis as an example. Further technical documentation is available from the inside-R Web site (Hainmueller & Diamond, 2012). Stata and MATLAB versions were also produced, and all are available online at Hainmueller's personal Web site (Hainmueller, 2011). For all analyses in this study, I used Synth package version 9.2 (Hainmueller et al., 2010) using all default options, with a minor but necessary adjustment to the original syntax as I describe in the Results section. The included *readme* file can be referenced for further documentation of the software defaults. All synthetic control matching procedures were run in Stata version 11.2.

The Synth software package minimizes the distance between the outcome measure and the observed covariates of the treated unit and the weighted synthetic control over the pretreatment period. Let $\boldsymbol{X}_1$ be a (k ×1) vector of selected pretreatment outcomes and predictors for the treated unit. $\boldsymbol{X}_0$ is a (k × J) matrix of the same pretreatment measures for the J treatment units. Then the software package solves for the matrix $\boldsymbol{W}$* of optimal weights to minimize the distance $||\boldsymbol{X}_1 – \boldsymbol{X}_0\boldsymbol{W}||_V$, which Abadie et al. defined as

$$||\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}||_{\boldsymbol{V}} = \sqrt{(\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W})'\boldsymbol{V}(\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W})} \qquad (4)$$

(2011, p. 4). $\boldsymbol{V}$ is a (k × k) symmetric, positive semidefinite matrix of weights that allows different predictor variables to have varying levels of influence on the outcome measure. By default, Synth chooses the matrix $\boldsymbol{V}$ that minimizes the mean squared prediction error over the pre-intervention period. Alternatively, Synth allows the user to choose matrix $\boldsymbol{V}$ when prior knowledge of the assumed predictive power among the variables is preferred.

The Synth package computes the matrix of w_j* weights via the mathematical optimization method of constrained quadratic programming. According to the *readme* file included in Synth for Stata (Hainmueller et al., 2010), "the constrained quadratic optimization routine is based on an algorithm that uses the interior point method to solve the constrained quadratic programming problem … implemented via a C++ plugin.” This plugin duplicates the optimization algorithms implemented using a package called *kernlab* within the R software version (Abadie et al., 2011).

Placebo Tests

To provide a method for statistical inference with synthetic controls, placebo tests have been recommended for analysis of the relative probability of finding a result as extreme as the case of interest (Abadie et al., 2010). Standard errors, as would be used in a traditional inferential statistical test, would equal zero in the case of synthetic controls due to the use of a single unit of treatment and a single comparison unit of aggregate data for estimation. But it should be noted that statistical uncertainty is still present “from ignorance about the ability of the control group to reproduce the counterfactual of how the treated would have evolved in the absence of the treatment” (Abadie et al., 2010, p. 496). The purpose of the placebo test is in analyzing this uncertainty.

Placebo tests serve as a type of permutation inference as was explored by Rosenbaum (2002). In his formulation, placebos serve as a method for detecting hidden biases by testing for systematic differences among units known to be unexposed to the treatment. They serve as a form of falsification test. In his discussion of matching methods, Sekhon discussed placebo tests as “underused robustness checks in observational studies” (2009, p. 501). Akin to the classic example of the lady tasting tea that introduced Fisher’s exact test (Fisher, 1935), these placebo tests allow exact, nonparametric statistical inference. A set of placebo tests creates a complete distribution of prediction error across the donor pool of available comparison units.

One form of placebo test recommended for use with synthetic controls is the in-space placebo, where an untreated unit from another location is substituted for the true treatment unit. In their original example, Abadie et al. (2011) used the Catalonia region to serve as an in-space placebo for the Basque Country. Other options for placebo test inference include in-time placebos, where the analysis is shifted in time to a span known to not represent the introduction of a treatment. This approach offers a falsification test for comparison, where it is assumed that no treatment effect should be detected at a time point when no treatment was introduced.

To draw statistical inference across a set of placebo tests—one for each untreated donor pool unit—Abadie et al. (2010) suggested an examination of the distribution of placebo RMSPEs. For each placebo test, the ratio of posttreatment RMSPE to pretreatment RMSPE offers a measure of how extremely the placebo group differs from its synthetic control compared to its closeness of match to its synthetic control over the pretreatment period. When this set of ratios is examined as a nonparametric distribution of donor pool cases, exact inference is possible by simply ranking the magnitude of true treatment group's ratio among its donor pool. For instance, a treatment group ratio found to be the most extreme alongside a set of 19 comparison units would suggest a 5% rate of Type I statistical error. The methods I used for examining sets of observational student achievement data capitalized on this method of using RMSPE ratios for statistical inference.

Previous investigation of synthetic controls

Betts et al. (2010) previously ran a series of falsification tests using empirical data to investigate the synthetic control method alongside other matching alternatives. Their report, published online in the form of a presentation paper, served as an initial study of potential matching methodologies for use in a large scale, in-progress evaluation of magnet schools by the

U.S. Department of Education. In comparison with whole-pool, regression-based, and propensity score matching methods, they concluded that synthetic controls performed worst with regard to false rejection rates beyond acceptable levels within the context of their particular data set for investigating magnet school performance. However, the scope of generalizability and the direct comparability among their chosen methods remains unclear.

Betts et al. (2010) used permutation methods other than the placebo tests recommended by Abadie et al. (2010), as a way of maintaining consistency for comparison across all four types of matching they investigated, with whole-pool comparison ranking the best overall. Rather than assigning placebo status to each available comparison unit in the donor pool, Betts et al. (2010) randomly assigned magnet school status to a non-magnet school, used a weighted least squares model with the non-magnet and synthetic control for inference, and repeated 1,000 times to establish rates of false positives, or Type I error. What is unclear from their analysis is how these results may have differed from or duplicated the performance of placebo tests as recommended by Abadie et al. One point that lacked clarity was their altering of the recommended constraint on the sum of synthetic control weights. They explained that

> to make this weighting scheme comparable to those implicit in our regression-based and propensity-score based matching methods, in which we chose the two closest matching comparison schools, we also set the sum of weights across the synthetic control schools equal to two." (Betts et al., 2010, p. 57)

They further discussed the use of synthetic controls in a technical appendix, but did not specify how the adjustment of the sum of weights was carried out. What is clear is that their use of *t* tests for significance of regression coefficients for each Monte Carlo run cannot be assumed to be equivalent to the nonparametric placebo test inference that was designed for use with synthetic controls.

Beyond their comparisons of Type I false rejection rates, Betts et al. (2010) further compared matching methods according to root mean square errors. In contrast to their demonstrated rejection rates, relative root mean square errors indicated that synthetic controls performed second best among the matching options. The smallest root mean square errors resulted from whole-pool comparison, which was again concluded to be the most reliable of the four methods. It should be noted that the measure of root mean square error is not equivalent to the RMSPE measure of fit discussed in Abadie et al. (2010). It is instead the "standard deviation of coefficients" (Betts et al., 2010, p. 16), and therefore a measure of consistency of the regression coefficient across their 1,000 Monte Carlo runs. In general, this finding that synthetic control matching consistently underperformed whole-pool comparison is disconcerting, as this should not be expected to be the case. Since the purpose of the Synth algorithm is to reweight to reduce the distance between the treatment and comparison, any reweighted synthetic control match should be expected to fit more closely than whole-pool comparison.

In summary the purposes of my study are the same as those of Betts et al. (2010): to examine the behavior of synthetic controls matches in the context of an available set of real data. However, I am concerned that with their study the necessity of altering suggested procedures for cross-comparison of methods came at the expense of generalizable conclusions regarding the appropriate use of synthetic controls. I suggest that the following methods I detail for this study better investigate the effectiveness of synthetic controls, by documenting the size of outcome measure differences needed for acceptable rates of Type I error, in line with the matching and inferential processes as they have been recommended for empirical practice.

Further investigation

For the purpose of directly investigating the effectiveness of creating synthetic control matches with educational assessment data, my interest is in the nonparametric inferential process provided by the recommended placebo test methods. To examine effect sizes of academic achievement measures required for traditional levels of statistical significance, I compared expected values of RMSPE, a measure of estimation error. In their discussion of estimation error as a unifying concept of correct causal inference, Imai, King, and Stuart stated,

> we focus on the most basic goal of statistical inference—the deviation of an estimate from the truth—rather than all of the various commonly used approximations to this goal, such as unbiasedness, consistency, efficiency, asymptotic distribution, admissibility, and mean-square error. (2008, p. 483)

The synthetic controls method offers direct access to this goal by allowing the researcher to access and work to reduce the RMSPE directly for the pretreatment period without regard to any measures of posttreatment outcomes, to reduce concerns of data fishing (Economist 5ae7, 2012). By comparing the patterns of RMSPE values across multiple specifications of control variables and multiple sets of observational data, I am able to directly examine the empirical levels of statistical error and effect size that can be expected in the presence of the observed and hidden biases found in academic achievement settings.

My investigation of the performance of synthetic controls further informs the use of a promising method for the evaluation of educational interventions and the potential creation of customized test norming groups. The synthetic control method's stability, sensitivity, robustness, and power to form an accurate counterfactual in the context of educational assessment data are the features of interest to this study. In summary my question is this: What size of effect is necessary for detection using synthetic controls in light of inherent levels of observed and unobserved biases resulting from the method’s implementation with academic achievement data?

Chapter III: Methods

As new methodological techniques that were developed within other disciplines continue to be applied in educational research contexts, consideration of the peculiarities specific to academic achievement data is of critical importance to drawing valid conclusions and forming meaningful interpretations. The use of synthetic controls with student test scores as outcome measures and school demographic data as covariates—as compared to analysis of variables of traditional interest in economics or political science—requires consideration of the issues surrounding specific types of statistical measurement error and the use of aggregated measures of student achievement.

To investigate the consequences of various types of measurement error on matching methods, Millimet (2011) simulated a variety of measurement error scenarios to compare using propensity scores for matching. His findings suggested that outcome measures exhibiting the classical measurement error of common concern within analyses of student achievement did not lead to bias but, as expected, reduced efficiency of estimation. Of greater influence was mean-reverting measurement error (which Millimet termed *nonclassical*), where large biases of estimates were exhibited through simulation. While social science methodologists must always be on the lookout for the predictable but meaningless effects of regression to the mean, matching methods are not immune to these concerns. As a further benefit of synthetic control methods however, much of the measurement error exhibited by individual observations can be expected to average out when aggregate measures are used, as was briefly discussed by Millimet quoting Griliches (1985).

Besides the concerns specific to the variables typical of educational research, the classroom as a research setting necessitates a unique set of considerations for analysts. Jaciw and

Newman (2011) discussed the particular issues of external validity and scalability in social science settings, with observations made on an open system where a wide range of treatments—not under researcher control—are taking place simultaneously and are interacting with one another and with time. In the interest of examining the results of forming synthetic controls with data collected under these particular empirical conditions, my analyses were carried out using real student achievement and demographic data from a single U.S. state.

Placebo Test Analyses

For the purpose of establishing baseline levels of sensitivity and power of the synthetic control method in the context of observed student achievement, I made use of many iterations of the in-place placebo tests recommended by Abadie et al. (2010). It should be noted that the implementation of a set of these placebo tests makes use of data for untreated comparison units only and no data for the treated unit of interest. For this reason I was able to calculate the values of posttreatment outcomes needed for statistically significant differences between a hypothetical treatment unit and a synthetic control formed from MAP exam observed scores and demographic covariates. As demonstrated by Abadie et al. (2010) in their California tobacco policy example, the exact test of statistical inference provided by placebo tests depends on the total number of available comparison units in the donor pool. Just as the RMSPE for the pretreatment period, as defined in Equation 3, measures closeness of match of the synthetic control, RMSPE over the posttreatment period measures the size of treatment effect. Therefore a large ratio of posttreatment RMSPE to pretreatment RMSPE ($RMSPE_{post}$ / $RMSPE_{pre}$) is an indicator not only of a large effect size between the posttreatment outcomes of the treated unit and its synthetic control but also of a close match between treated and synthetic control across the pretreatment period. When the treatment unit of interest has the single most extreme RMSPE ratio, the probability of finding as extreme a ratio by random chance is $p = 1 / (J + 1)$, where J is the number of untreated donor pool units. By extension, for statistical significance at the traditional level of 5% probability of Type I error ($\alpha = .05$) the treated unit must achieve an RMSPE ratio larger than N of the donor pool units, where the integer $N \geq 0.95\ (J + 1)$. Similarly for statistical significance at a level of 10%, $\text{RMSPE}_{\text{post}}$ / $\text{RMSPE}_{\text{pre}}$ for the treated unit must be greater than those of N control units, where the integer $N \geq 0.90\ (J + 1)$. In general, for a chosen level of Type I error α,

$$N \geq (1 - \alpha)(J + 1). \tag{5}$$

The primary purpose of my analyses is to identify and compare these ratios of $\text{RMSPE}_{\text{post}}$ / $\text{RMSPE}_{\text{pre}}$ for the *N*th-ranked unit within each set of placebo tests to find baseline levels of statistical significance for comparison with a hypothetical treatment unit.

In the interest of practical interpretation, after finding these *N*th-ranked ratios I transformed them to represent estimated gains necessary for a hypothetical unit of interest over the posttreatment period for statistical significance. To estimate a baseline level of necessary $\text{RMSPE}_{\text{post}}$ for statistical significance, I multiplied each *N*th-ranked $\text{RMSPE}_{\text{post}}$ / $\text{RMSPE}_{\text{pre}}$ ratio by a typical value of $\text{RMSPE}_{\text{pre}}$ for that set of placebo tests, which represents a measure of average closeness of synthetic control match. As shown in Equation 3, with MAP scores as an outcome measure, RMSPE is a measure of score increase per time-point of measurement. Therefore, this final measure of $\text{RMSPE}_{\text{post}}$ represents the estimated value of standardized MAP test points per test administration necessary for statistical significance at the chosen level. For final comparison across models, I transformed these estimates to represent effect sizes required for statistical significance.

Model Specifications

In this way my study examines the practical implications of concerns with observed and hidden biases inherent in educational assessment data and resulting from synthetic controls analyses. By documenting the behavior of the synthetic controls method when applied to a set of data representative of aggregate measures and demographic variables commonly available to educational researchers, I provide evidence of the robustness of these techniques when (a) matched only on the single content area pretreatment outcome scores, (b) matched additionally on a second achievement measure, (c) matched using additional available demographic covariates, (d) the treatment point is shifted in time, and (e) more pretreatment time-points are added. Table 1 and Table 2 present the combinations of each set of variable conditions manipulated to form each model specification. These combinations result in a total of 16 unique specifications compared across two student achievement research scenarios.

Table 1

List of Specifications for Cross-sectional Model Synthetic Controls

Specification	Outcome measure	Covariates	Pretreatment years
1	Math		2009 2010
2	Math	TS PM PF	2009 2010
3	Math	Reading	2009 2010
4	Math	Reading TS PM PF	2009 2010
5	Reading		2009 2010
6	Reading	TS PM PF	2009 2010
7	Reading	Math	2009 2010
8	Reading	Math TS PM PF	2009 2010
9	Math		2008 2009 2010
10	Math	TS PM PF	2008 2009 2010
11	Math	Reading	2008 2009 2010
12	Math	Reading TS PM PF	2008 2009 2010
13	Reading		2008 2009 2010
14	Reading	TS PM PF	2008 2009 2010
15	Reading	Math	2008 2009 2010
16	Reading	Math TS PM PF	2008 2009 2010

Note. All synthetic controls were matched on the outcome measure for each pretreatment year. TS = total students within school; PM = percentage of minority students (not categorized as White); PF = percentage of free or reduced-price lunch participants.

Table 2

List of Specifications for Cohort Model Synthetic Controls

Specification	Outcome measure	Covariates	Exams pre	Exams post
1	Math		5	4
2	Math	TS PM PF PE	5	4
3	Math	Reading	5	4
4	Math	Reading TS PM PF PE	5	4
5	Reading		5	4
6	Reading	TS PM PF PE	5	4
7	Reading	Math	5	4
8	Reading	Math TS PM PF PE	5	4
9	Math		6	3
10	Math	TS PM PF PE	6	3
11	Math	Reading	6	3
12	Math	Reading TS PM PF PE	6	3
13	Reading		6	3
14	Reading	TS PM PF PE	6	3
15	Reading	Math	6	3
16	Reading	Math TS PM PF PE	6	3

Note. All synthetic controls were matched on the outcome measure for each pretreatment year. TS = total students within school; PM = percentage of minority students (those not categorized as White); PF = percentage of free or reduced-price lunch participants; PE = percentage of English language learners.

Each of the described respecifications was compared across two disjoint data sets and between two different models exemplary of typical analyses of aggregated student achievement: a cross-sectional model and a student cohort model. First I computed synthetic controls at the school level to mimic a common cross-sectional, annual analysis of elementary school performance. This model, as has been seen with federal accountability during the implementation of the No Child Left Behind Act, treats data as a repeated cross-sectional or trend analysis. Each synthetic elementary school is a third-through-fifth-grade cluster whose oldest students advance and are replaced each year by a new group of third graders. This means the model replaces approximately one third of the student body as each year proceeds. The cross-sectional model was used for the spring-season MAP scores across the five years of available data, including spring 2008 through spring 2012. Second I modeled a synthetic cohort of students, where unique grade-level groups were measured thrice annually as they advanced from third through fifth grade. This analysis more closely matches a panel or longitudinal data model, as the individual students observed are consistent except in cases of transfer in or attrition out. For increased precision in this model I included the available fall and winter season MAP scores in addition to the spring-season measures used in the first analysis model. This means the cohort model was run across nine available test events from fall of 2009 to spring of 2012.

First I examined the matching behavior of synthetic controls across alternate data specifications comparable to common scenarios of data availability. As shown in Table 1 and Table 2, I began with the simplest specification of synthetic controls matched using only the pretreatment scores for the outcome measure of interest: either mean math or reading score. Then I included the secondary available subject scores—either reading or math—as an additional pretreatment covariate for improved matching. Then within each of these specifications I added the full set of available demographic covariates in attempt to optimize the closeness of synthetic match over the pretreatment period.

Next I examined each set of specifications for sensitivity to time-varying factors within the available achievement data. This was done by either including data for additional pretreatment time periods or by shifting the time point assumed to represent the treatment. For the cross-sectional model built on annual testing events, I compared the results of matching on two years of pretreatment data to the inclusion of an additional pretreatment year, so that matching is performed on three years pretreatment and outcomes are examined for two years posttreatment. For the synthetic-cohort model, where nine test events were included, I instead examined the effects of shifting the hypothetical treatment point in time. I compared the results of placing the treatment time-point between winter and spring of Grade 4 with shifting it to the point between the spring of Grade 4 and the fall of Grade 5.

Additionally, each model specification was compared across two independent sets of data to examine the robustness of the synthetic matches to the particularities of available data. This resulted in a total of 64 unique synthetic matching model specifications, each with its own unique set of placebo test matches.

Hypotheses

Each respecification and set of placebo tests was designed to address a particular hypothesis. In combination, these hypotheses address several specific aspects of synthetic controls' sensitivity to matching model respecification and availability of data.

Pretreatment measures

I designed specification conditions to compare matching on the student achievement outcome measure alone to matching with an additional measure of a different-subject test score. I hypothesized that the addition of more student measures in the pretreatment period would improve the closeness of match, as defined by reduced values of $\text{RMSPE}_{\text{pre}}$. I base this hypothesis on recommendations for matching procedures from Rubin (2006). Additionally and for the same reasons, I hypothesized that the addition of demographic variables for matching in pretreatment would improve the closeness of matches, although my results have the potential of addressing the concerns of other researchers regarding the common occurrence of covariate measures whose inclusion may adversely impact closeness of match. However, I further assumed that demographic measures of convenience should not improve matching as substantially as did the addition of achievement measures.

Pretreatment time factors

For differing model specifications involving time factors, I hypothesized that additional years of data in pretreatment would improve the closeness of synthetic control matches by reducing $\text{RMSPE}_{\text{pre}}$. This is in line with Abadie et al. (2012) who showed that the inclusion of multiple pretreatment measures over many time-points can be assumed to better reduce the effects of bias from unobservables. With regard to shifts of the treatment period in time, I hypothesized that the synthetic controls should be robust to these changes, since all analyses involved hypothetical treatment points and placebo units where no true treatment was administered.

Synthetic match level

I designed two sets of matching models—synthetic cross-sections and synthetic cohorts—to compare designs presently common in education research. Since neither model was expected to outperform the other from a theoretical perspective, I approached this comparison with the expectation that each would perform similarly as a null hypothesis. This assumes that final resulting necessary effect sizes will produce similar patterns across both data models.

Synthetic control donor pool

I further examined each matching specification concurrently between two different sets of comparison donor pools with the hypothesis that the methods were robust to data availability, and so placebo tests would perform similarly between the otherwise equivalent sets of data. This assumes that patterns of final necessary effect sizes will appear similar within each dataset.

Data

All included data represent real student test events from the Measures of Academic Progress (MAP) exams provided by Northwest Evaluation Association. My analyses were restricted to school- and cohort-level clusters with MAP scores in both math and reading. The data come from students in South Carolina public schools during school years 2007/2008 through 2011/2012.

Measures of Academic Progress (MAP)

I selected MAP test data from Northwest Evaluation Association to analyze the effectiveness of synthetic controls because of the desirable psychometric properties of MAP scores. These properties allowed me to treat MAP data as a best-case scenario for applying synthetic matching to a set of statistically sophisticated achievement data. I began with the assumption that if issues were evident with forming valid synthetic controls when using MAP scores, then the majority of current measures of student achievement could be expected to fare no better in closeness of synthetic control matching.

MAP scores are estimated using a computer adaptive testing system. Rather than administering a fixed set of test items to all students, computer-based algorithms tailor each student’s next test item based on his or her individual performance on previous items (Nunnally & Bernstein, 1994). In this way, MAP scores avoid many of the problematic floor and ceiling effects common of fixed-form tests such as with traditional state achievement exams. This also means that standard error of measurement, though never possible to eliminate, is significantly reduced for students at the upper and lower ends of performance in comparison to traditional fixed-form tests. For the purposes of this study, this means that students can be measured over time with greater precision, and therefore the data should allow more accurate matches for synthetic controls.

MAP scores also offer a consistent scale of measurement across grades. MAP items are designed by fitting the single-parameter logistic (*1PL* or *Rasch*) model, where the single parameter represents item difficulty for estimating student achievement level, theta. The MAP scores are then a simple linear transformation of the student ability parameter estimate, theta (Northwest Evaluation Association, 2011b, p. 23). MAP documentation refers to this rescaling as the *RIT* (a contraction of *Rasch unit*) scale. In this way, students' estimated values of theta are assumed to be measured consistently as they advance from grade level to grade level. This eliminates the need for linking or conversion of student test scores for comparability across grade levels as is often necessary with state-based exams, where test content and difficulty are varying across grade levels.

Data for the state of South Carolina were made available by Northwest Evaluation Association for this study. South Carolina participates in administering MAP exams at high rates since statewide policies adopted in 2005-2006 mandated statewide implementation of interim assessments. Currently approximately 97% of students in South Carolina take MAP tests (Northwest Evaluation Association, 2011a). These levels of data availability allowed me to create multiple model specifications within a set of students sharing common state-level policies and regional similarities. For the purposes of forming accurate matched comparisons, this allowed me to begin with a comparison donor pool with similarities on unobservable covariates, which is theorized to reduce unobservable bias in synthetic control matching.

Cross-sectional model data

Data for the first, cross-sectional analysis model are from two large public school districts in South Carolina. All student-, school-, and district-level identifiers were removed by Northwest Evaluation Association before data were provided to me for analysis. I formatted data for this analysis model to resemble an annual, school-level similar to *adequate yearly progress* measures under the federal No Child Left Behind Act. To form cross-sectional aggregate measures, I calculated mean test scores and demographic measures by elementary school groups across Grades 3, 4, and 5 for spring MAP test events from 2008 to 2012.

I first standardized all student-level MAP scaled scores using nationally normed mean and standard deviation values according to values published in 2011 norming documentation from Northwest Evaluation Association (2011a). Each student-level scaled score was standardized for subject by grade level by season before the resulting standardized math and reading scores were aggregated over elementary schools. This means that a math or reading score of zero indicates that a school scored at the national mean for the spring testing season. The school-level demographic measures available for use as synthetic matching covariates were (a) total number of students enrolled, (b) percentage of students identified as belonging to racial minority groups (non-White), and (c) percentage of students participating in free or reduced-price lunch (FRL) programs. Although the additional demographic measure of students identified as English language learners (ELL) was also provided, this count was only available at the district level. Since my data were analyzed across district level sets, this measure was identical for all schools across each set of data. Therefore the ELL measure had to be excluded from the cross-sectional model analyses.

Data were cleaned to retain elementary schools with complete data for math and reading and available covariates. The final sets of data used for analysis included 47 elementary schools in District #30 and 48 elementary schools in District #61. These included approximately 11,000 unique students in District #30 and 17,000 in District #61. School-level descriptives for these schools are presented in Table 3. In addition to school-level means and standard deviations, intercorrelations are given for relationships between each measure and spring math and reading MAP scores in 2011.

Table 3

Means, Standard Deviations, and Intercorrelations Among School-Level Test Scores and Demographic Measures for Cross-Sectional Model

	District #30 (*n* = 47)				District #61 (*n* = 48)			
			Correlations (2011)				Correlations (2011)	
Measure	*M*	*SD*	Math	Reading	*M*	*SD*	Math	Reading
Math								
2008	-0.1502	0.5194	.85	.88	-0.2201	0.3812	.91	.93
2009	-0.0938	0.4435	.86	.88	-0.0775	0.3902	.90	.93
2010	0.0689	0.4755	.96	.94	-0.0009	0.3555	.95	.95
2011	0.1314	0.4892	-	.96	0.0597	0.3585	-	.95
2012	0.1870	0.4961	.93	.91	0.1048	0.3516	.97	.95
Reading								
2008	-0.1123	0.5084	.87	.91	-0.1781	0.3992	.87	.94
2009	-0.1183	0.4736	.83	.89	-0.1201	0.4152	.90	.97
2010	-0.0078	0.4616	.93	.95	-0.0139	0.3870	.92	.98
2011	0.0173	0.5022	.96	-	0.0540	0.3536	.95	-
2012	0.0882	0.4750	.90	.93	0.1044	0.3194	.91	.97
Total students	460.72	221.23	.44	.41	687.06	218.28	.41	.43
Percent minority	67.09	32.57	-.86	-.88	43.60	21.26	-.51	-.60
Percent FRL	61.68	29.63	-.85	-.87	56.10	24.26	-.76	-.84

Note. Percent minority represents students not categorized as White. FRL = free or reduced-price lunch participants. All coefficients are significant at $p < .01$.

Table 3 suggests that both districts performed similarly in math and reading, with gradual score increases across the five year period. However, these increases are not large in comparison with the variability across schools as seen in the standard deviations. Schools within District #61 tend to be less variable than across District #30. Further, schools in District #61 are somewhat larger and have fewer minority students and FRL participants. In comparison with demographic measures, math and reading scores are much more highly correlated with lagged achievement measures across all available test events. It should be noted that student achievement measures like those analyzed here should be expected to be more stable over time when aggregated as compared with student-level test scores.

Figure 1 gives a graphical representation of the variability of school-level math and reading achievement measures across 47 public elementary schools in District #30. Figure 2 includes the same data for 48 schools in District #61. All scores were standardized at the national level and include mean scores for five spring testing seasons. Trend lines of overall district means are shown in bold black.

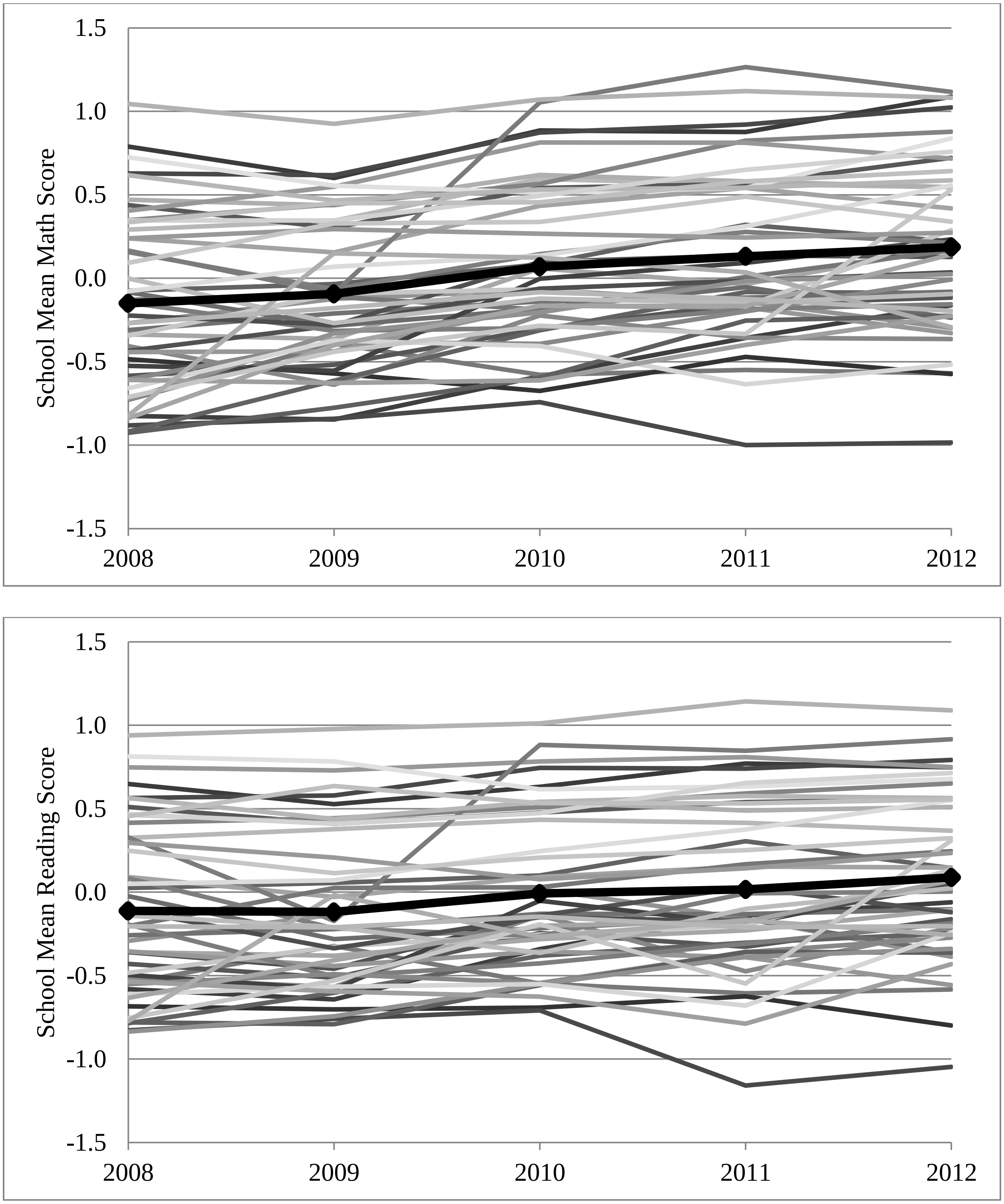


Figure 1. District #30 school achievement score trends (*n* = 47). Trend in bold black is mean score across all schools.

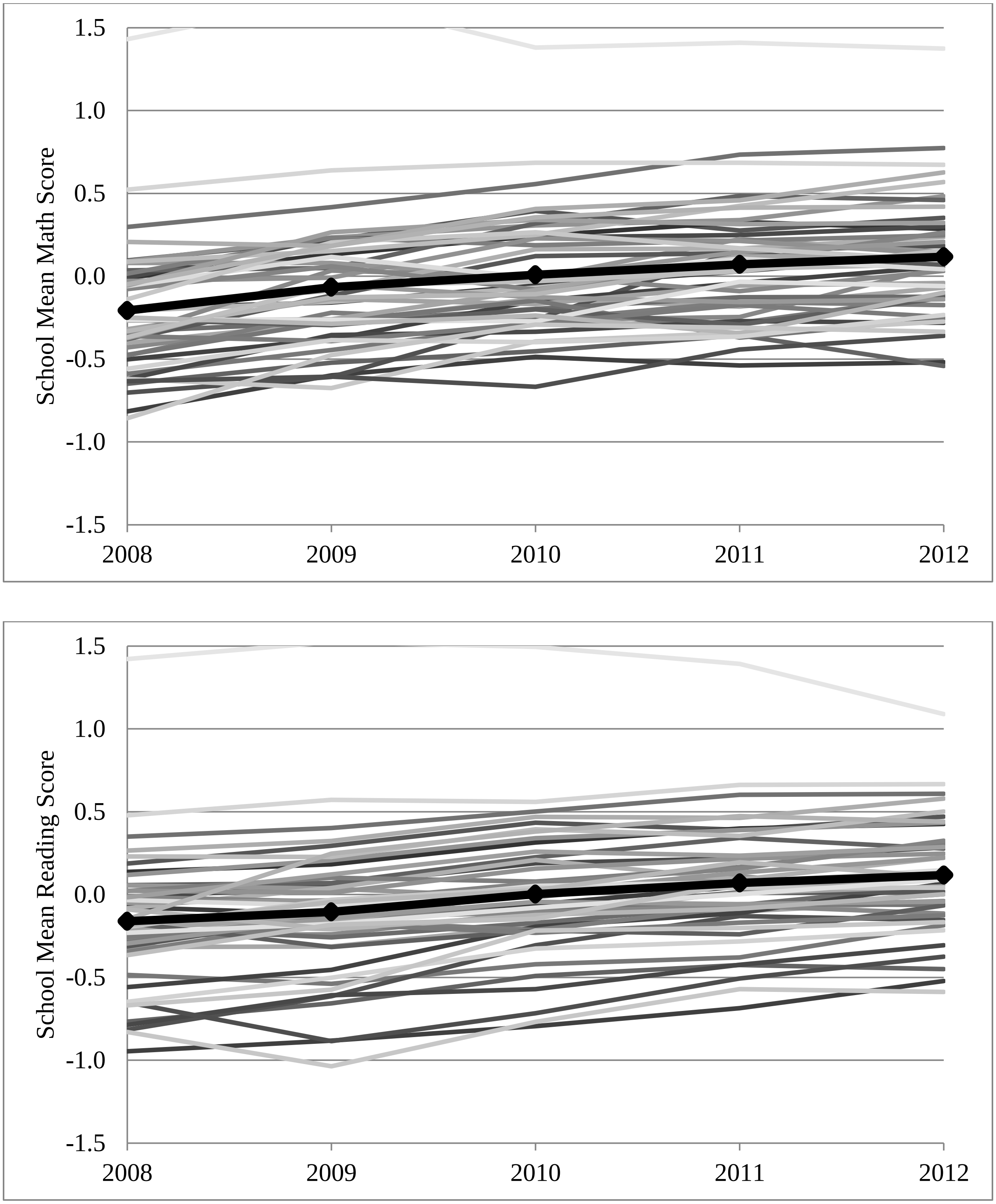


Figure 2. District #61 school achievement score trends (*n* = 48). Trend in bold black is mean score across all schools.

Cohort model data

Data for the second, cohort-level analysis model were grouped by rural school locale indicators in South Carolina. I formatted data for this model to replicate a panel data comparison, where students were measured three times annually as they rose from Grade 3 to Grade 5. All student-, school-, and district-level identifiers were removed by Northwest Evaluation Association before data were provided to me for analysis. To form cohort-level aggregate measures, I calculated mean test scores and demographic measures by grade-level groups beginning with fall of third grade in 2009 and ending with spring of fifth grade in 2012.

As with the cross-sectional data model, I first standardized all student-level MAP scaled scores using nationally normed mean and standard deviation values according to values published in 2011 norming documentation from Northwest Evaluation Association (2011a). Each student-level scaled score was standardized for subject by grade level by season before the resulting standardized math and reading scores were aggregated over elementary student cohorts. This means that a math or reading score of zero indicates that a cohort of students scored at the national mean for the same testing season: fall, winter, or spring. The available school-level demographic measures available for use as synthetic matching covariates were (a) total number of students enrolled, (b) percentage of students identified as belonging to racial minority groups (non-White), and (c) percentage of students participating in FRL programs. The (d) additional ELL measure was also included in the cohort model, although it is a measure of the proportion of English language learners by district rather than at a school level.

Based on qualitative similarity between student cohorts and data availability for forming large enough samples for analysis, two sets of cohorts were selected according to federal measures of school locale. Schools identified as Locale 41 are categorized in the federal

Common Core of Data as *rural fringe*, defined as a “rural territory that is less than or equal to 5 miles from an urbanized area, as well as rural territory that is less than or equal to 2.5 miles from an urban cluster” (National Center for Education Statistics, n.d.). Schools identified as Locale 42 are categorized as *rural distant,* a “rural territory that is more than 5 miles but less than or equal to 25 miles from an urbanized area, as well as rural territory that is more than 2.5 miles but less than or equal to 10 miles from an urban cluster” (National Center for Education Statistics, n.d.). Note that while schools included in the previous cross-sectional model tended to be located in more populous urban regions, due to the choice of test events and grade levels, a few schools appeared in both the cross-sectional and the cohort model analyses.

Data were cleaned to retain student cohorts with complete data for math and reading across fall, winter, and spring testing seasons, in addition to available covariates. Further, cohorts whose sample size of students varied over testing seasons by more than 50% were excluded from analyses, to avoid schools whose testing policies differed from others. The final sets of data used for analysis included 59 student cohorts from public schools across the state of South Carolina from *rural fringe* (Locale 41) schools (including a total of over 5,000 unique students), and 50 student cohorts from *rural distant* (Locale 42) schools (including over 3,000 unique students). Cohort-level descriptives for these student cohorts are presented in Table 4. In addition to cohort-level means and standard deviations, intercorrelations are given for relationships between each measure and fall math and reading MAP scores in 2010 (Grade 4).

Table 4

Means, Standard Deviations, and Intercorrelations Among Cohort-Level Test Scores and Demographic Measures for Cohort Model

	Locale 41 (*n* = 59)				Locale 42 (*n* = 50)			
			Correlations (F4)				Correlations (F4)	
Measure	*M*	*SD*	Math	Reading	*M*	*SD*	Math	Reading
Math								
F3	-0.1593	0.3032	.90*	.84*	-0.2681	0.3931	.91*	.81*
W3	-0.1461	0.3692	.92*	.83*	-0.1915	0.4401	.94*	.80*
S3	-0.0625	0.4091	.92*	.83*	-0.1345	0.4683	.92*	.77*
F4	-0.0912	0.3654	-	.91*	-0.1940	0.4571	-	.88*
W4	-0.0817	0.4002	.96*	.89*	-0.1549	0.4574	.93*	.78*
S4	-0.0266	0.4168	.90*	.80*	-0.0725	0.5114	.89*	.76*
F5	-0.0590	0.3916	.91*	.84*	-0.1592	0.4420	.94*	.86*
W5	-0.0540	0.4066	.85*	.79*	-0.1570	0.4510	.84*	.72*
S5	0.0652	0.4838	.80*	.74*	-0.0417	0.4946	.76*	.68*
Reading								
F3	-0.1599	0.2929	.77*	.86*	-0.2516	0.3254	.80*	.83*
W3	-0.0594	0.3379	.83*	.91*	-0.1413	0.3386	.86*	.87*
S3	-0.1023	0.3488	.86*	.91*	-0.2122	0.3702	.85*	.87*
F4	-0.1013	0.3342	.91*	-	-0.2368	0.3736	.88*	-
W4	-0.0399	0.3561	.91*	.95*	-0.1440	0.3690	.87*	.89*
S4	-0.0329	0.3441	.88*	.91*	-0.1667	0.3658	.85*	.85*
F5	-0.0330	0.3589	.82*	.89*	-0.1821	0.3522	.79*	.84*
W5	-0.0118	0.3571	.71*	.75*	-0.0795	0.3305	.68*	.70*
S5	0.0351	0.3398	.77*	.81*	-0.0758	0.3295	.74*	.74*
Total students	539.07	195.17	.20	.15	413.86	178.26	-.02	-.12
Percent minority	43.15	29.16	-.69*	-.73*	54.33	31.31	-.55*	-.49*
Percent FRL	66.87	16.36	-.59*	-.73*	74.62	14.90	-.37*	-.38*
Percent ELL	5.68	3.91	.12	.01	4.18	4.05	.05	-.03

Note. Percent minority represents students not categorized as White. Math and reading measures are grouped by season and grade level. F = fall; W = winter; S = spring; FRL = free or reduced-price lunch participants; ELL = English-language learners.

**p* < .01.

Table 4 shows that math and reading scores overall tend to fall below national averages and overall trend upward. However, in comparison with cohort-level standard deviations, these increases are not extreme. Cohorts of students in *rural distant* schools (Locale 42) tend to come from smaller schools whose proportions of minority students and FRL participants are somewhat higher than in *rural fringe* (Locale 41) schools. As with the cross-sectional model data, math and reading achievement tend to be much more highly correlated with lagged test scores than with any of the available demographic measures.

Figure *3* gives a graphical representation of the variability of cohort-level math and reading achievement measures across 59 public elementary schools identified in Locale 41 (rural fringe). Figure 4 gives the same for cohorts labeled Locale 42 (rural remote). All scores were standardized using national norms and include mean scores for nine testing seasons, three times annually for Grades 3 through 5. Trend lines of overall cohort means are shown in bold black.

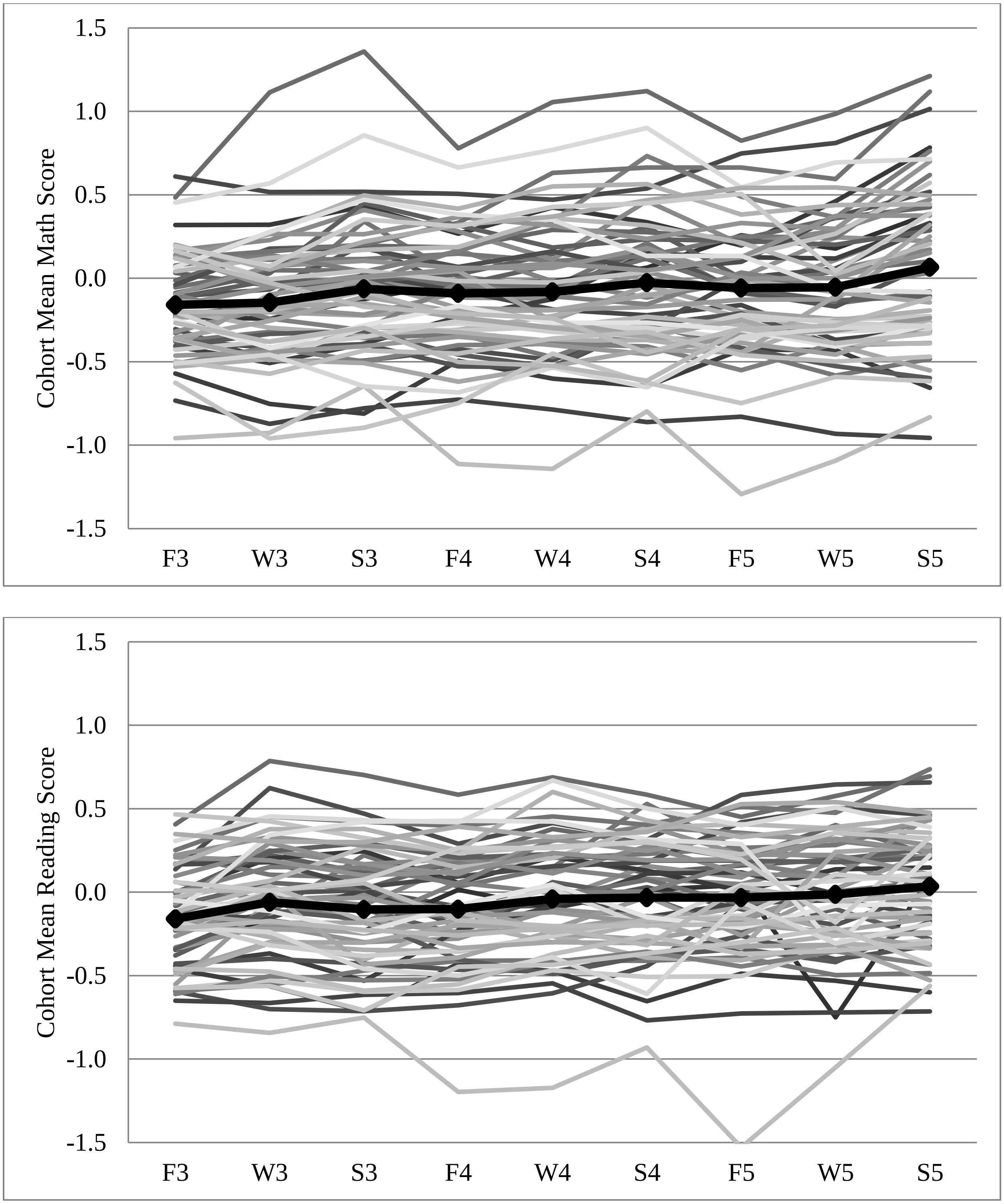
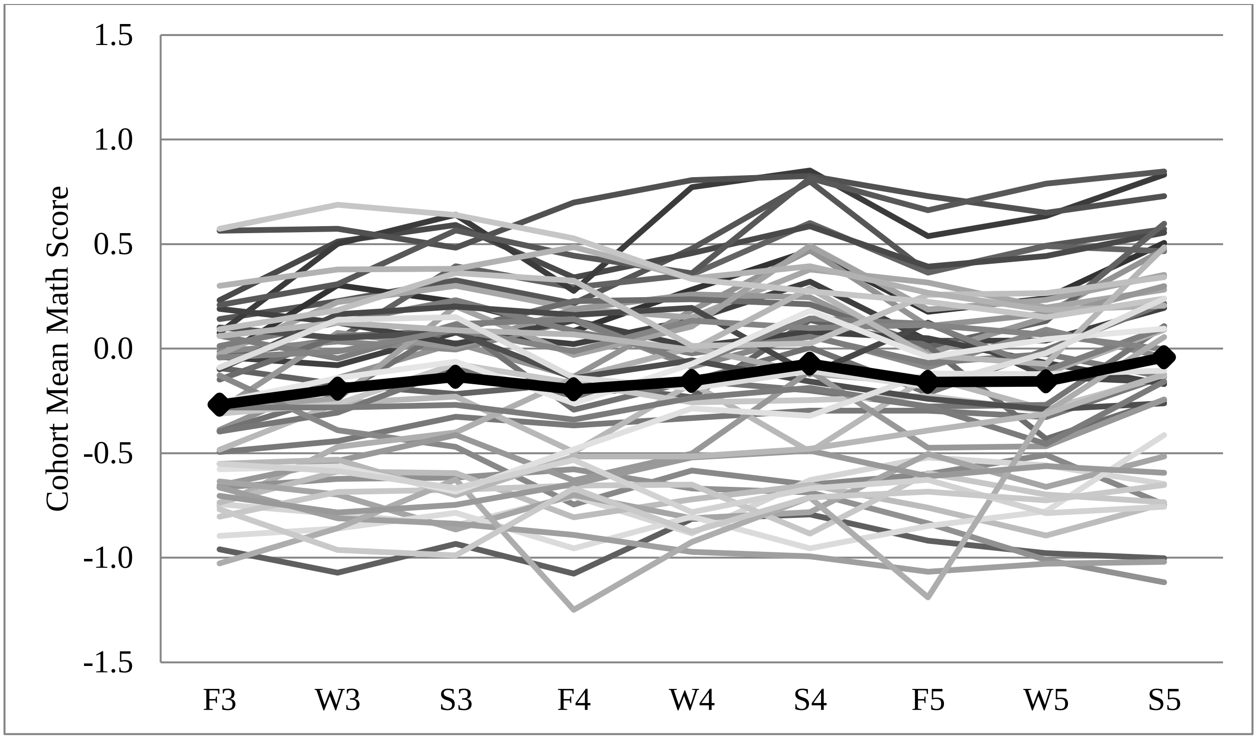


Figure 3. Cohort-level achievement score trends for schools in Locale 41(n = 59). Trend in bold black is mean score across all cohorts. The horizontal axis marks nine testing seasons from third through fifth grades. F = fall; W = winter; S = spring.



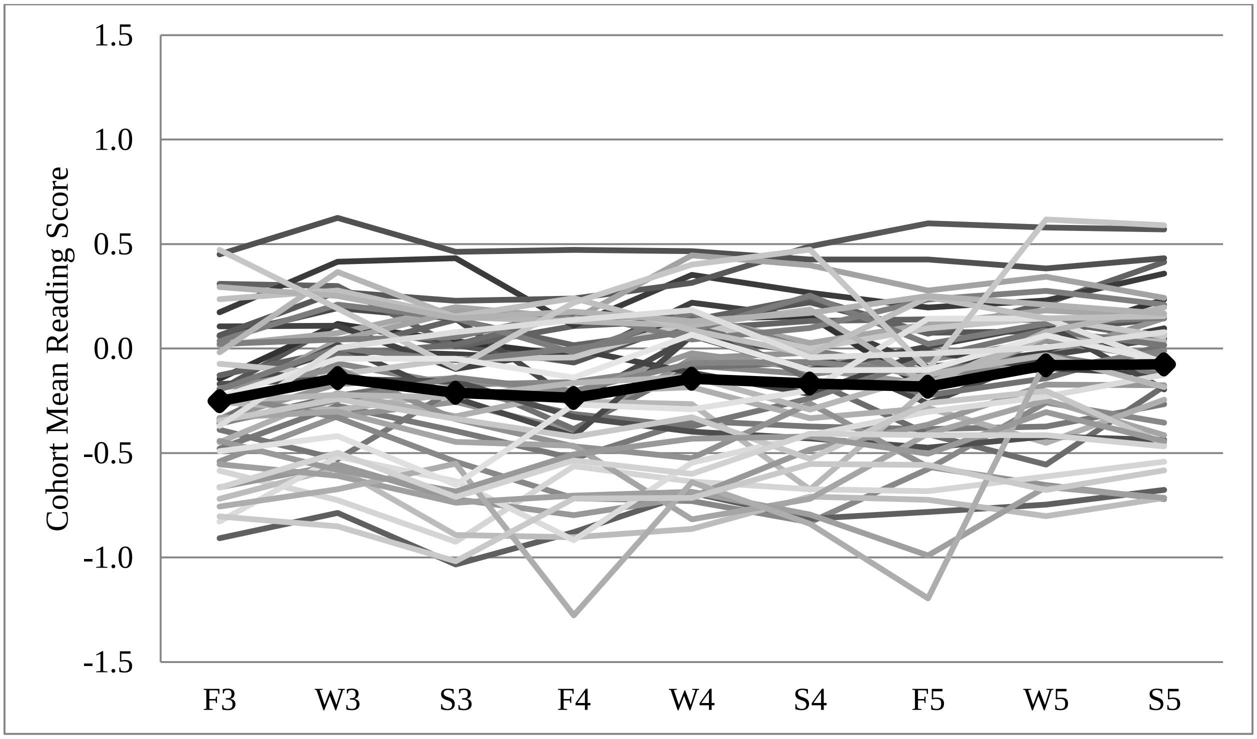


Figure 4. Cohort-level achievement score trends for schools in Locale 42 (*n* = 50). Trend in bold black is mean score across all cohorts. The horizontal axis marks nine testing seasons from third through fifth grades. F = fall; W = winter; S = spring.

Chapter IV: Results

All synthetic controls weights were calculated using default settings for Synth version 9.2 for Stata. These were run for 16 data specifications across the four distinct sets of data containing 204 comparison units for a total of 3,264 placebo test runs.

Analyses

The Synth package allows the user to specify the number of desired significant figures for use in the optimization process. If no value for significant figures is entered, the program defaults to eight digits. This value determines calculation precision via the number of unrounded decimal places that are displayed by default in the minimized $\text{RMSPE}_{\text{pre}}$ values. However, the Stata syntax includes a rounding procedure that affects any output values beyond this point. Once the RMSPE is displayed, the matrix of optimized weights is automatically rounded to three decimal values. All other output values that are displayed or saved are calculated from this rounded matrix rather than to the eight significant figures requested during the optimization procedure.

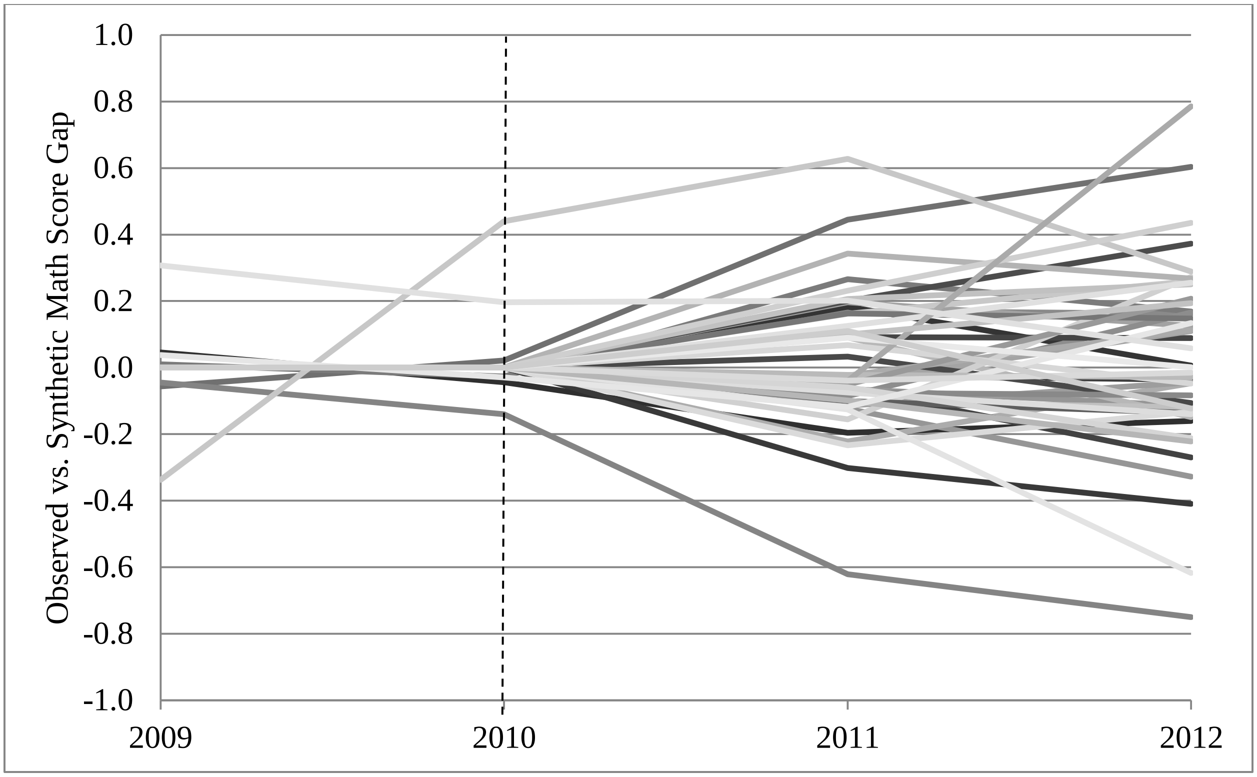
Based on my analyses, I found that this rounding discrepancy sometimes did and sometimes did not affect the final results. In the case of synthetic weightings that resulted in many units with a value of zero, the rounding of the matrix values had no effect. However, in the case of weightings where most units received some very small share of the total weight, the built-in rounding procedure caused rounding error to compound. When a small RMSPE was calculated from the weighted outcome values as shown in Equation 3, it sometimes differed from the displayed, unrounded value of RMSPE by as many as six decimal places – a factor of 1,000,000. In order to make use of the $\text{RMSPE}_{\text{post}}$ and posttreatment-to-pretreatment ratio values not calculated by Stata, it was necessary for me to adjust the original syntax of Synth to remove the matrix rounding command. Once this was done, all calculated RMSPE values matched those automatically displayed by Stata.

Since the original raw, student-level MAP scores and the published national norming values provided four significant figures, five digits were maintained throughout analyses. All values based on synthetic matching weights were then rounded to four figures for final reporting.

Placebo Test Gaps

Figure 5 through Figure 12 display the gaps between the observed values of each comparison unit and their synthetic controls. Gap values near zero indicate close matches between the unit and its reweighted control. The set of gap trends for all units within a comparison pool form a set of placebo tests. The vertical scale in each figure is marked in standardized math or reading score units. Pretreatment time points used in the matching process begin at the left side of each figure, with a vertical dashed line marking the final pretreatment point for each specification. All time points right of the vertical dashed line are posttreatment gaps between synthetic and observed. Cross-sectional model results for District #30 and #61 are followed by those for cohort models with Locale 41 and Locale 42.

The description of each data specification was given previously in Table 1 and Table 2. The synthetic weights produced by Synth showed no influence from the addition of any covariates. Therefore the resulting values and graphics for Specifications 2, 3, and 4 were identical to those for Specification 1. Likewise Specifications 6, 7, and 8 duplicated Specification 5, and so on. As a result only findings for Specifications 1, 5, 9, and 13 are reported.



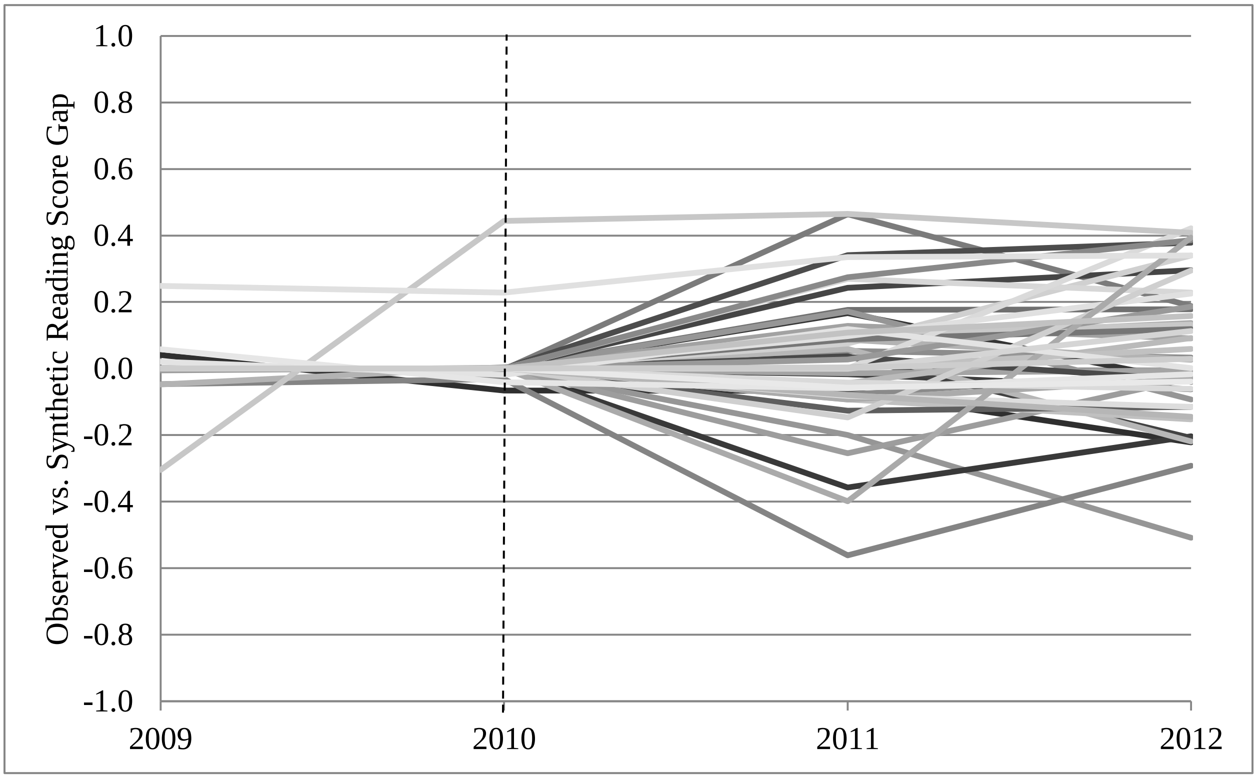
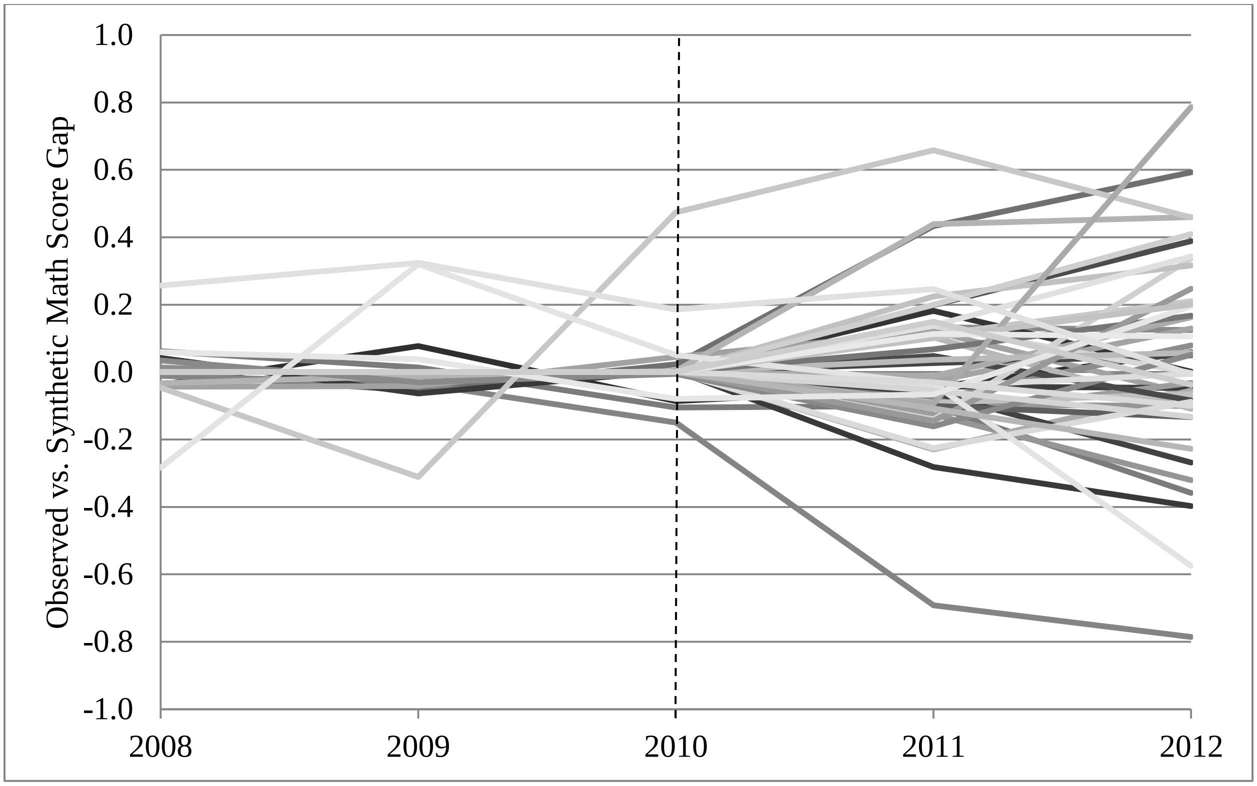


Figure 5. District #30 placebo test gaps for Specification 1 (above) and Specification 5 (below).



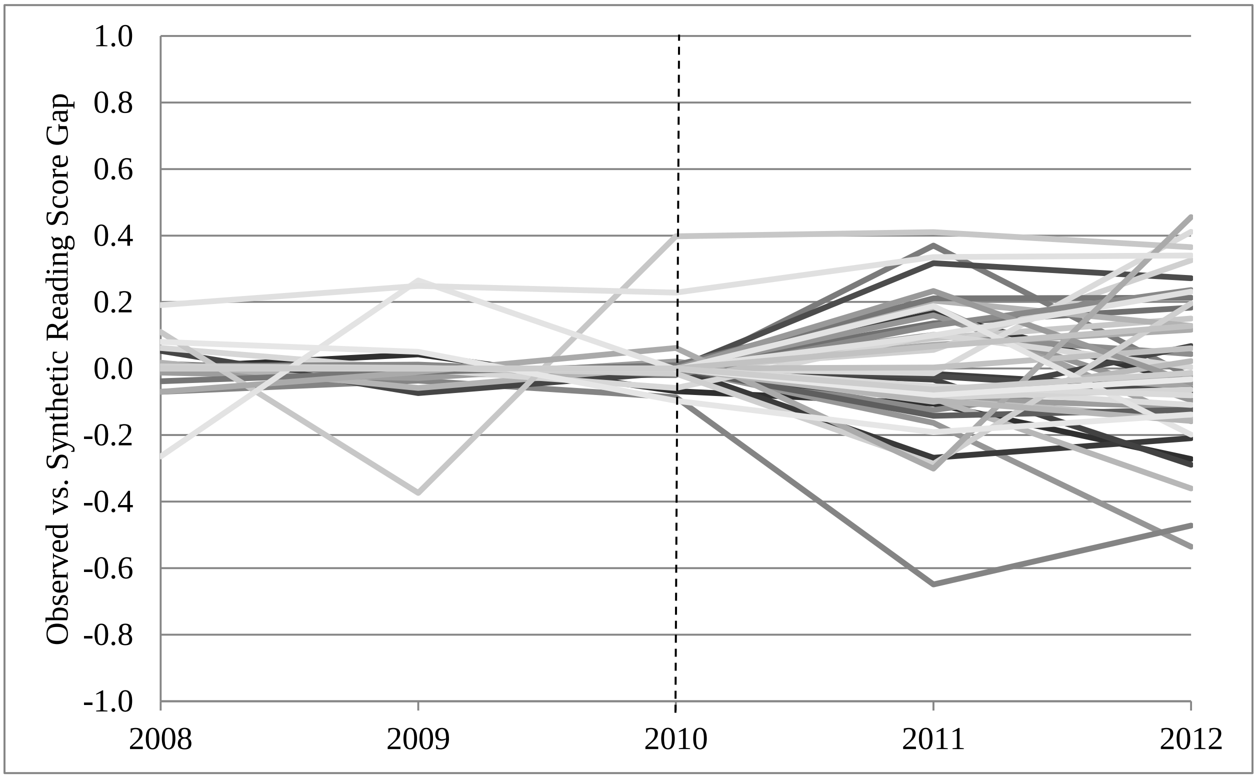
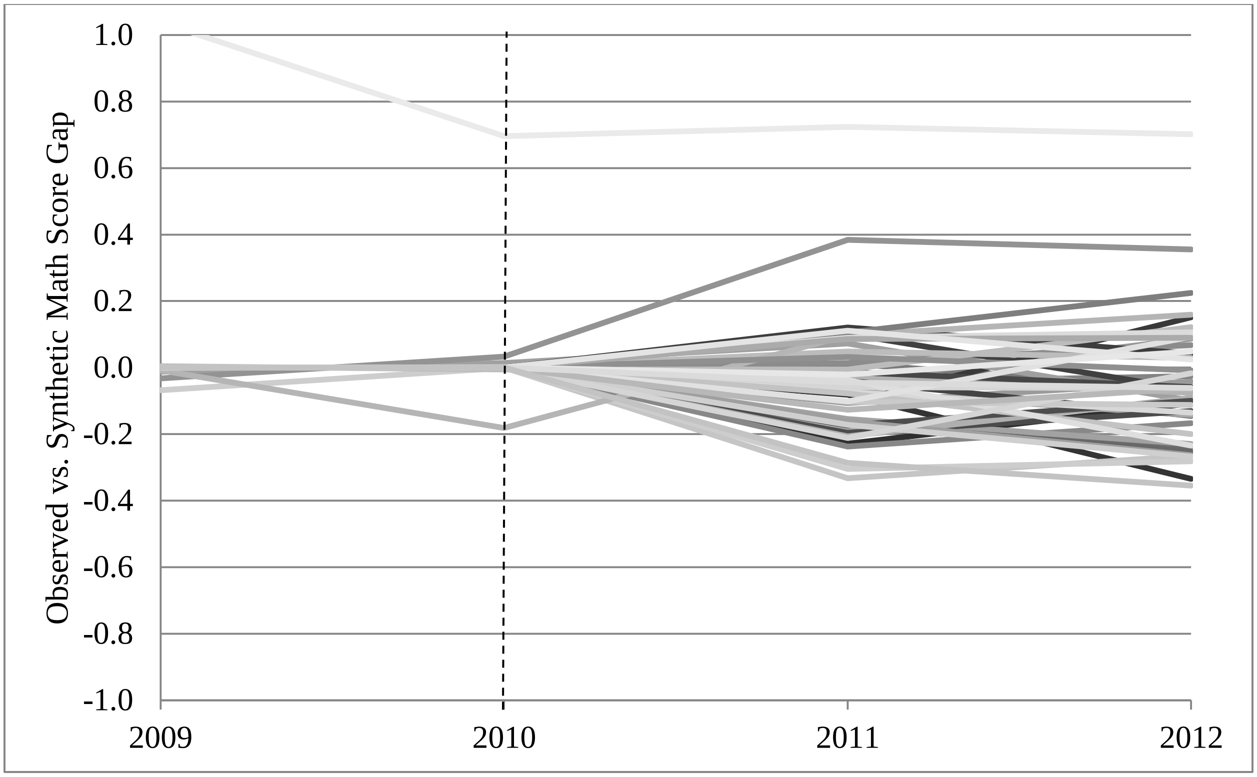


Figure 6. District #30 placebo test gaps for Specification 9 (above) and Specification 13 (below).



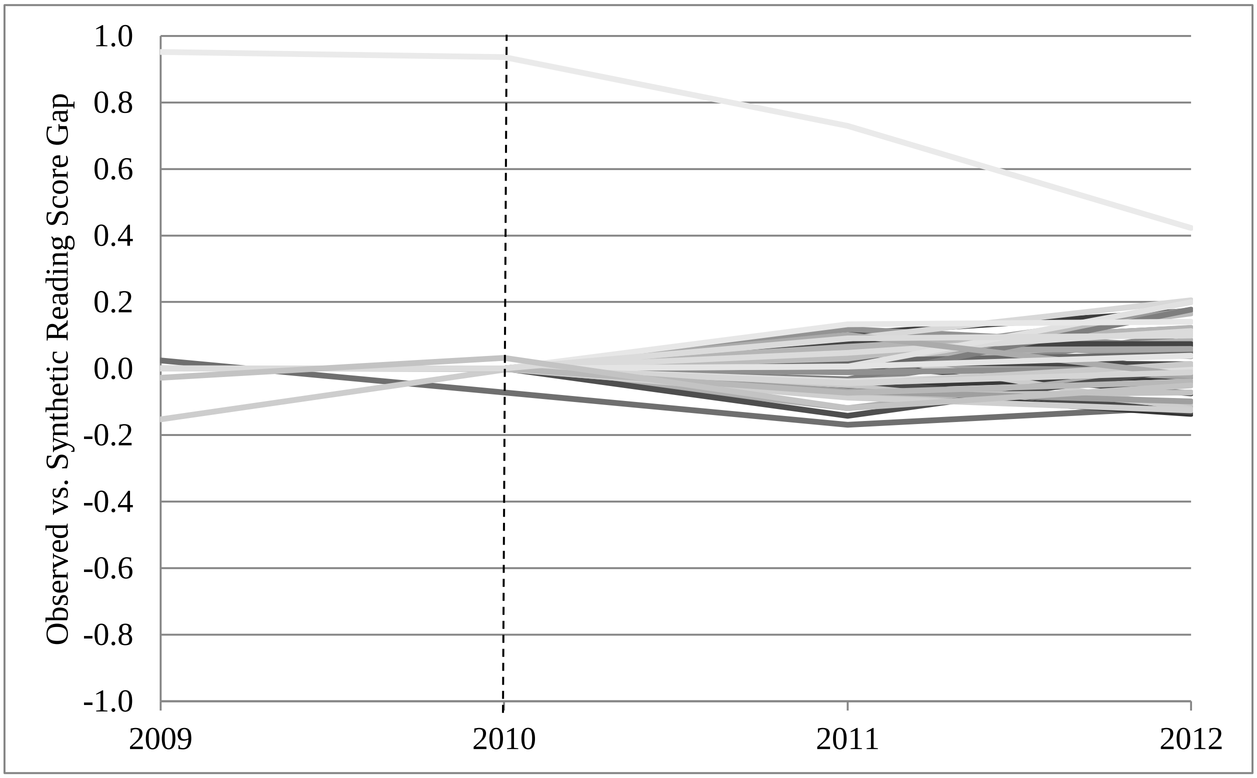
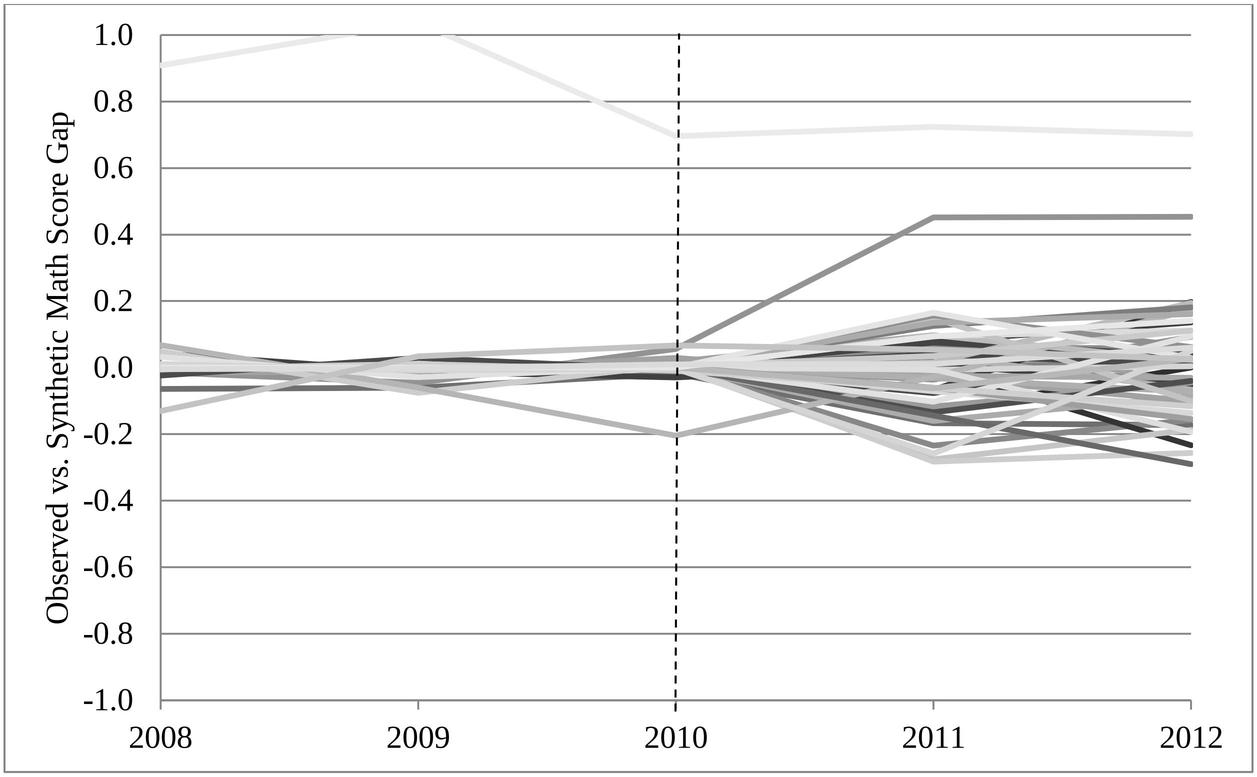


Figure 7. District #61 placebo test gaps for Specification 1 (above) and Specification 5 (below).



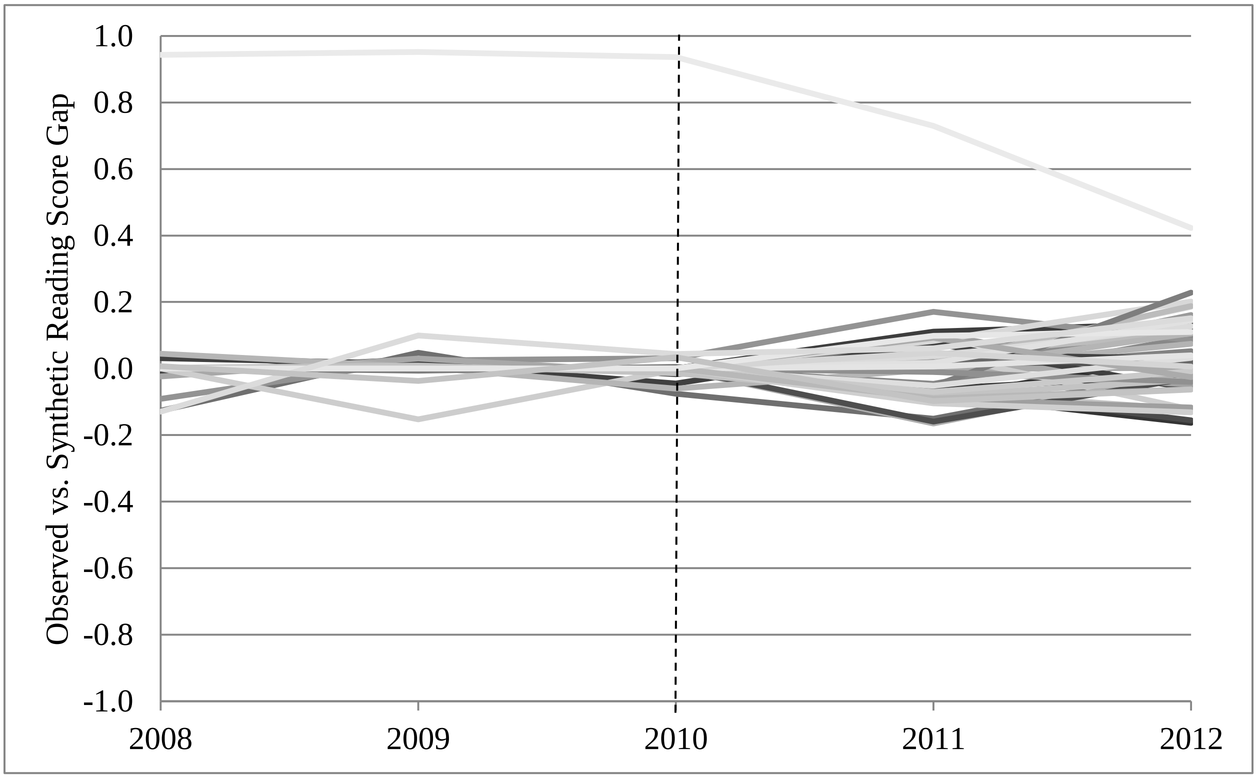
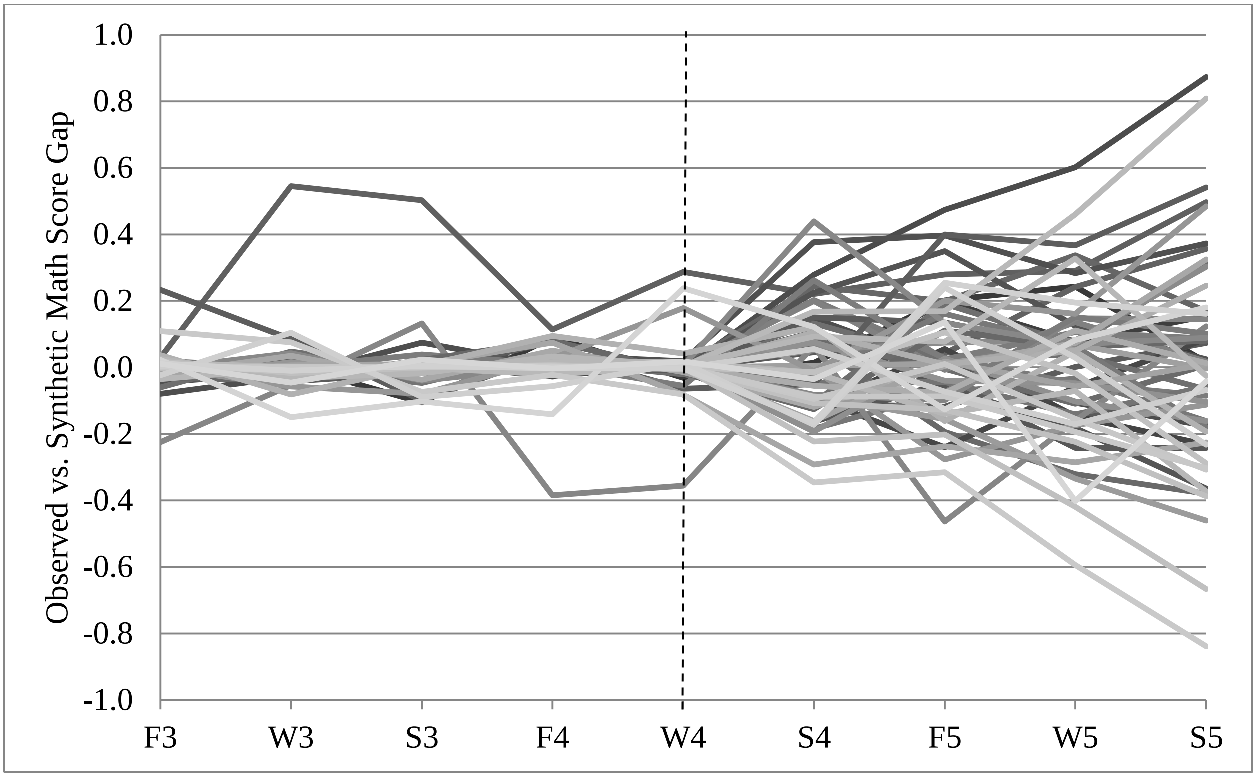


Figure 8. District #61 placebo test gaps for Specification 9 (above) and Specification 13 (below).



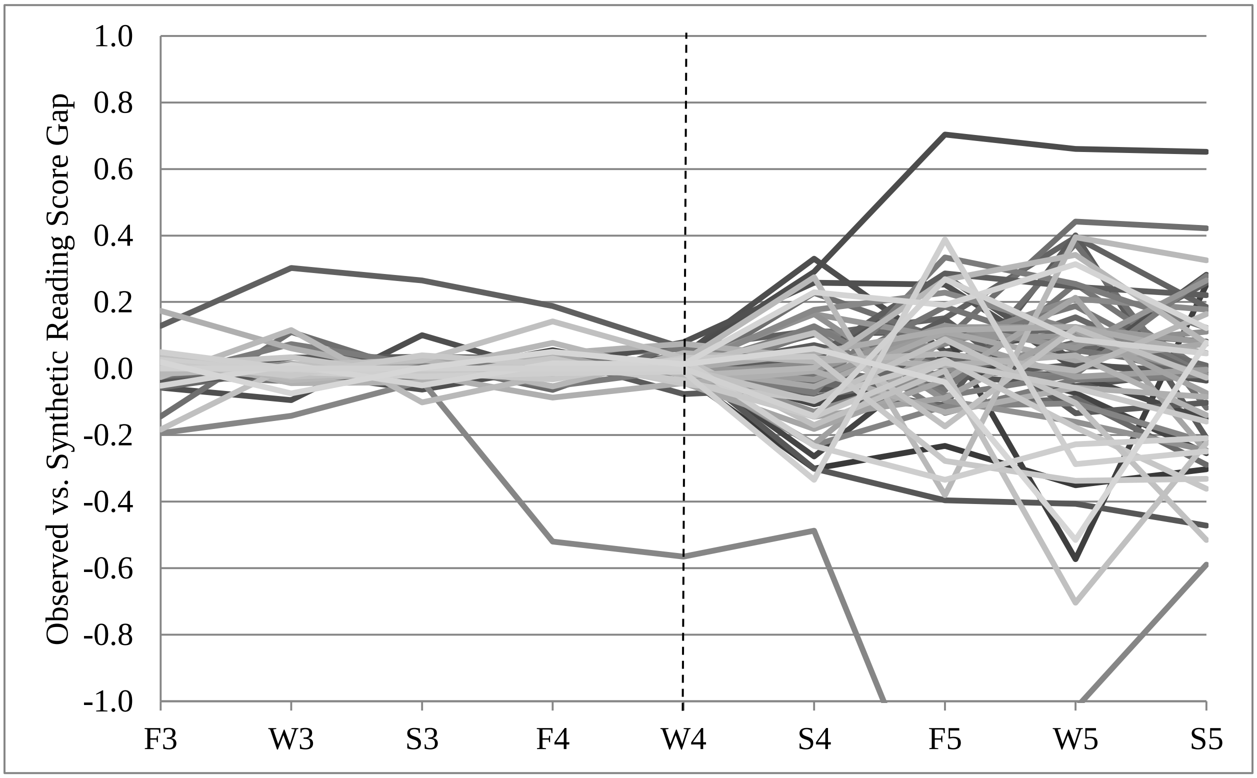
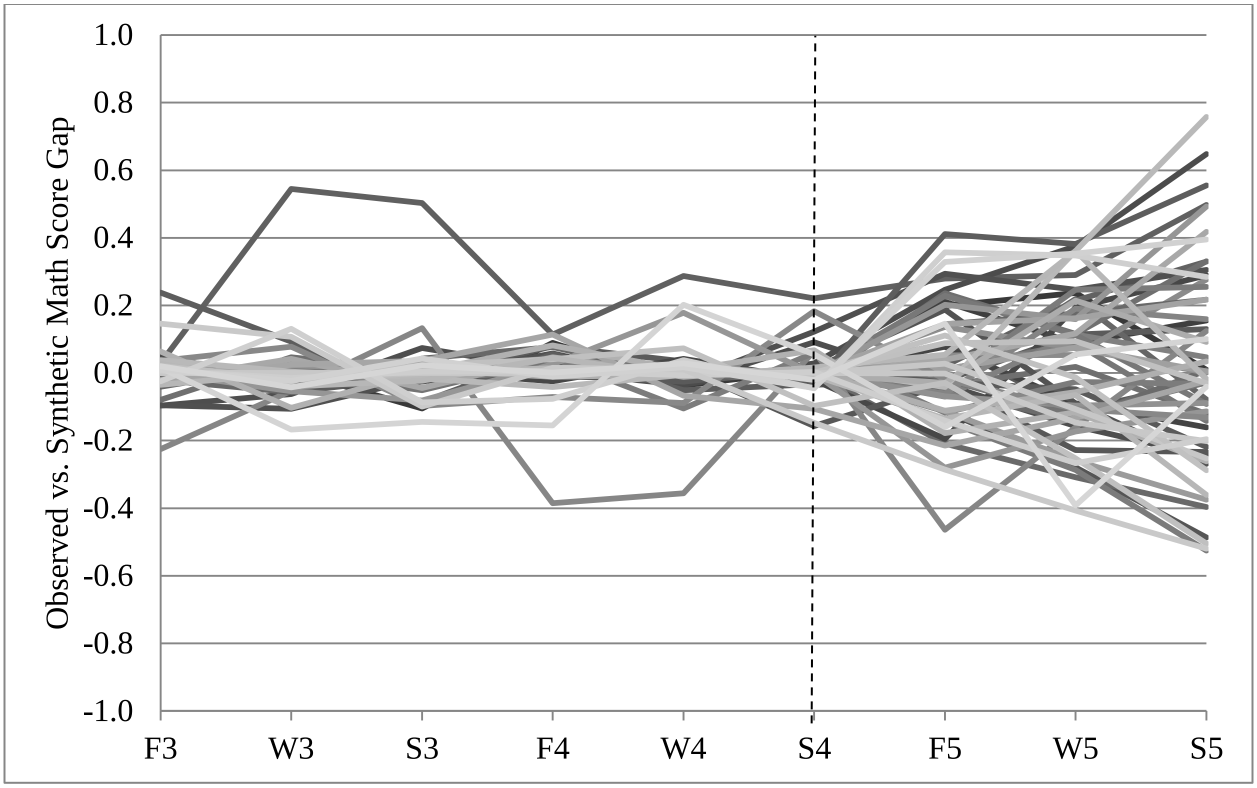


Figure 9. Locale 41 placebo test gaps for Specification 1 (above) and Specification 5 (below).



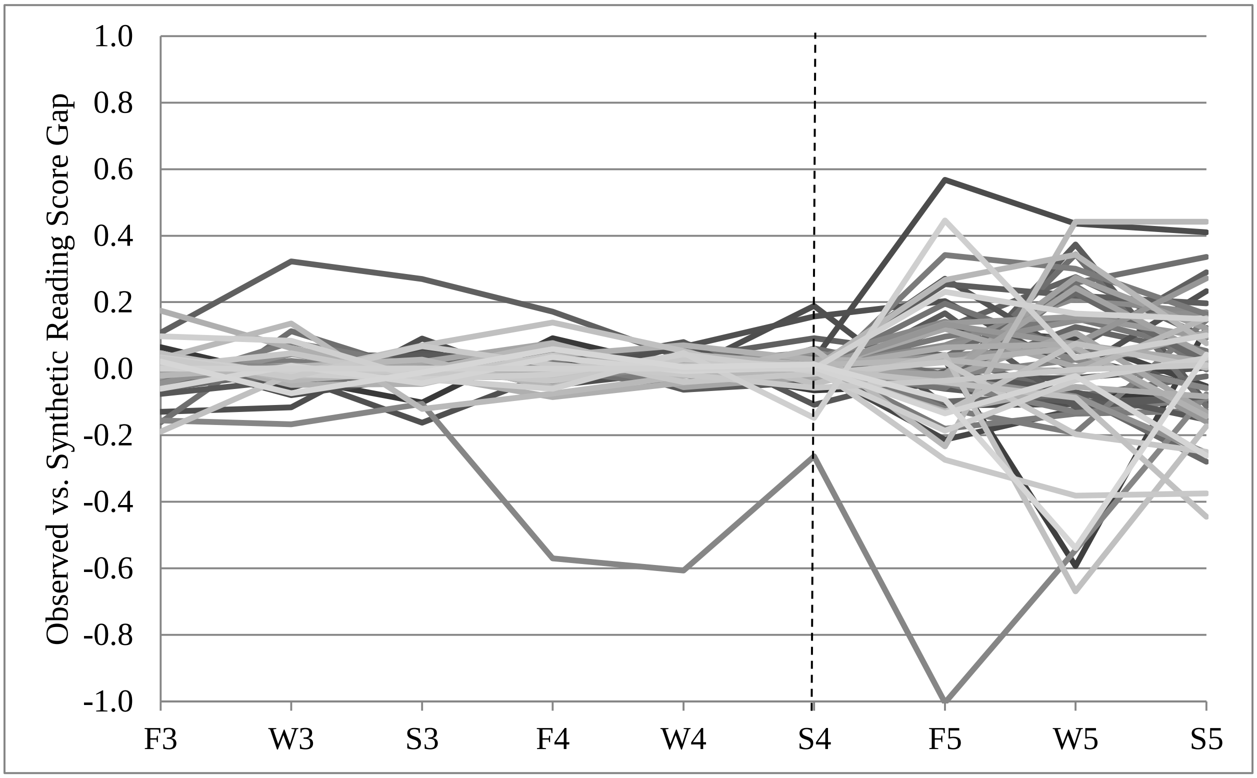
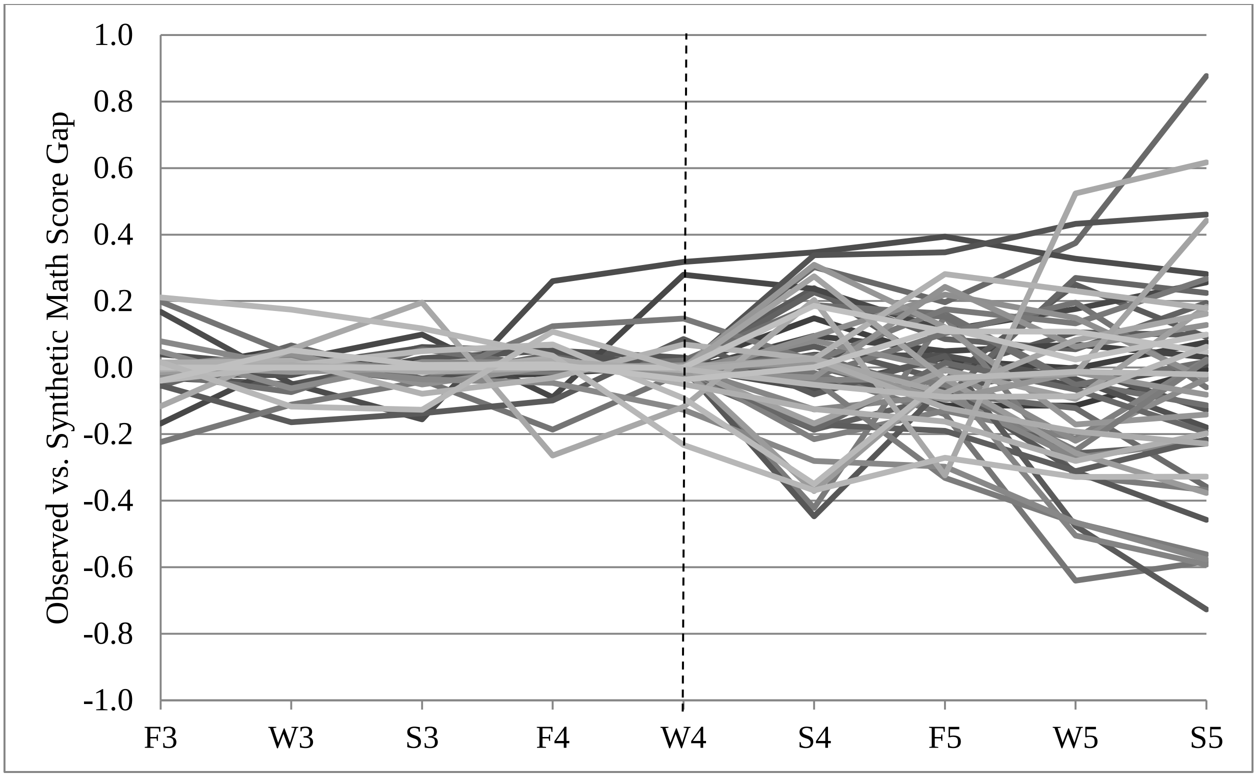


Figure 10. Locale 41 placebo test gaps for Specification 9 (above) and Specification 13 (below).



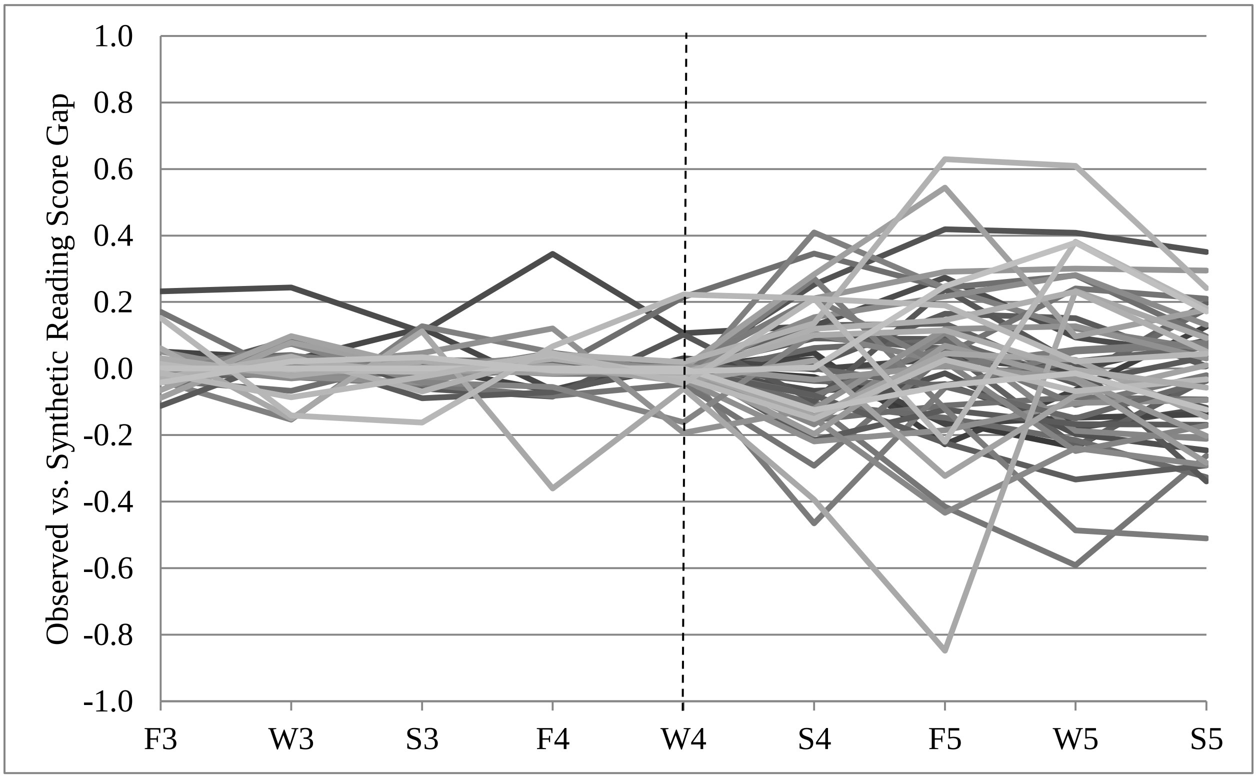
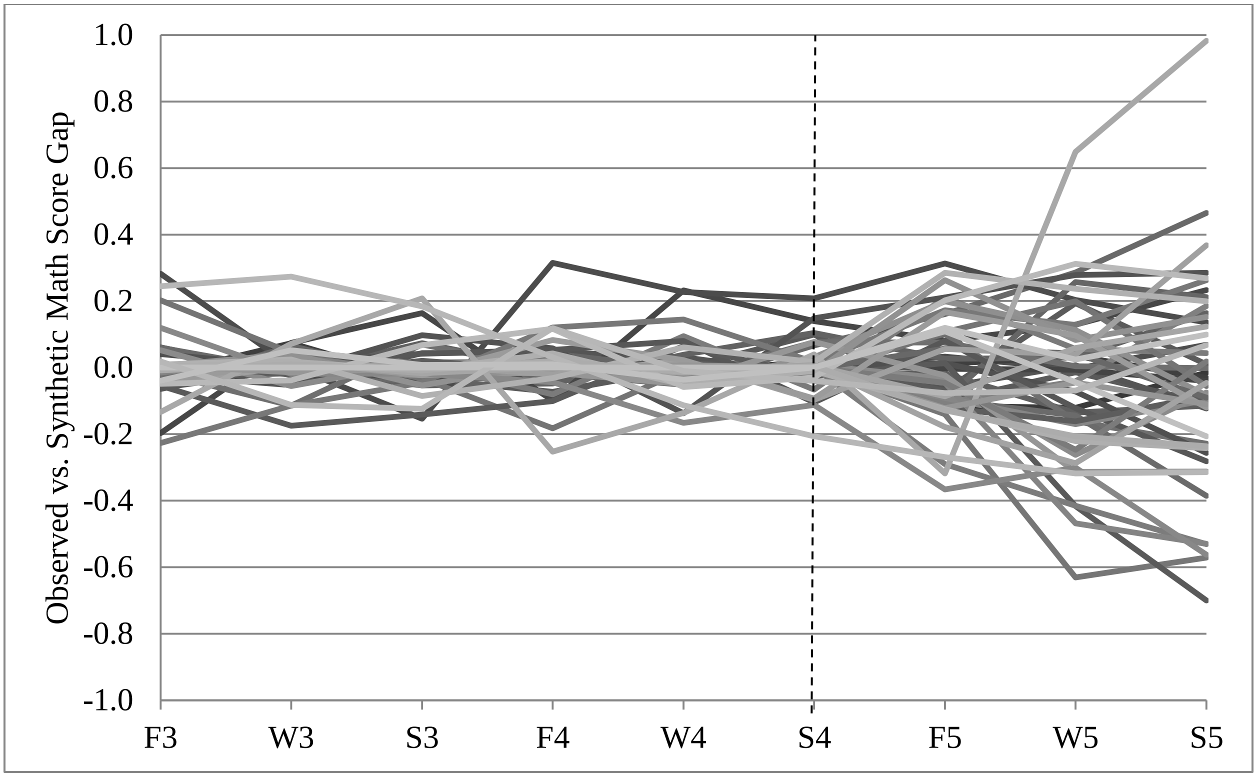


Figure 11. Locale 42 placebo test gaps for Specification 1 (above) and Specification 5 (below).



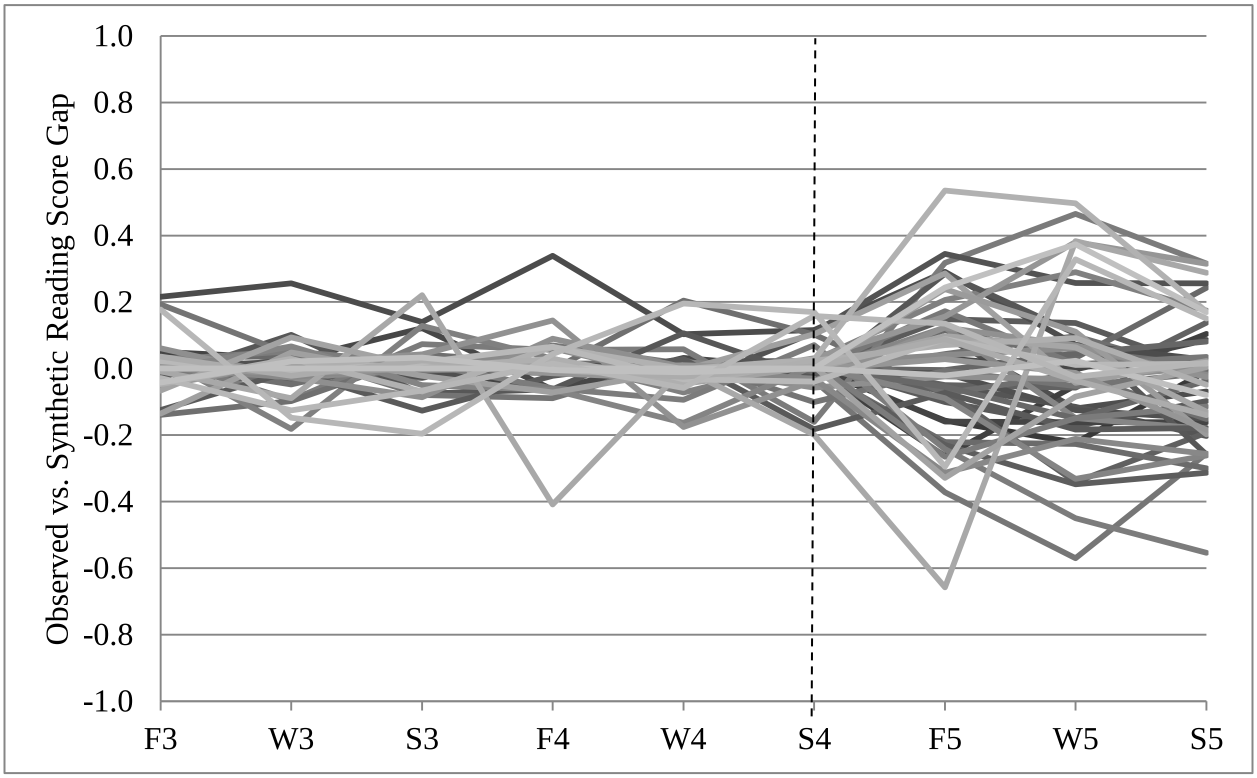


Figure 12. Locale 42 placebo test gaps for Specification 9 (above) and Specification 13 (below).

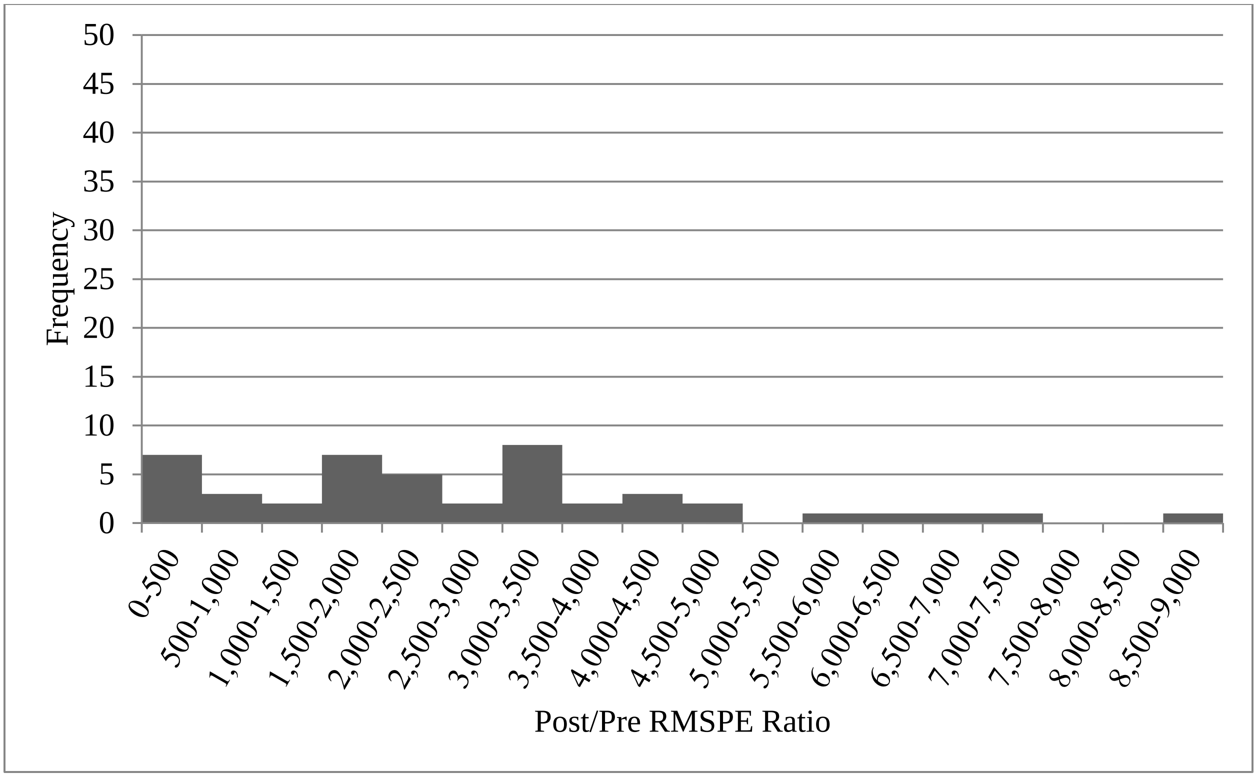
RMSPE Post/Pre Ratios

To form each discrete distribution of placebo test ratios, the proportion of each $\text{RMSPE}_{\text{post}}$ to $\text{RMSPE}_{\text{pre}}$ was calculated to four significant figures as described above. These are the falsification distributions designed to allow for statistical inference in comparison with a treated unit and its synthetic control.

When implemented with the standardized, aggregate achievement measures, many units were able to be perfectly matched across the pretreatment period. With measures significant to four figures, any value of $\text{RMSPE}_{\text{pre}} < 0.0001$ is—within rounding error—no different from zero. These close matches appeared in as many as 42 of the 48 comparison units in District #61 using Specification 5. At the least, six of the 50 units in Locale 42 matched perfectly for Specification 9. Since these pretreatment error values form the denominator in the distribution of ratios, an adjustment was necessary to avoid undefined values resulting from division by zero.

For all values of $\text{RMSPE}_{\text{pre}}$ reported as < 0.0001, I imputed a value of 0.00005 to avoid undefined ratios and prevent the inflation of some large ratios over others due to nonsignificant decimal values and rounding error. This choice of imputed value is justified in that it gives the close matched units the benefit of being the largest value of pretreatment RMSPE that is also smaller than any RMSPE calculated to four significant figures. Since the method of statistical inference using ratio distributions depends only on rank ordering, this choice preserves this property without sacrificing the best-matched data in the process.

Each set of ratios calculated after imputation for denominators of zero is shown as a histogram in Figure 13 through Figure 20. Each horizontal interval represents a range of 500 ratio units. The vertical scale is raw frequency count of placebo units. Placebo units with close matches across pretreatment are those on the right with the largest ratios. Small ratios on the left are the result of either poor pretreatment matches, small posttreatment gaps, or both.



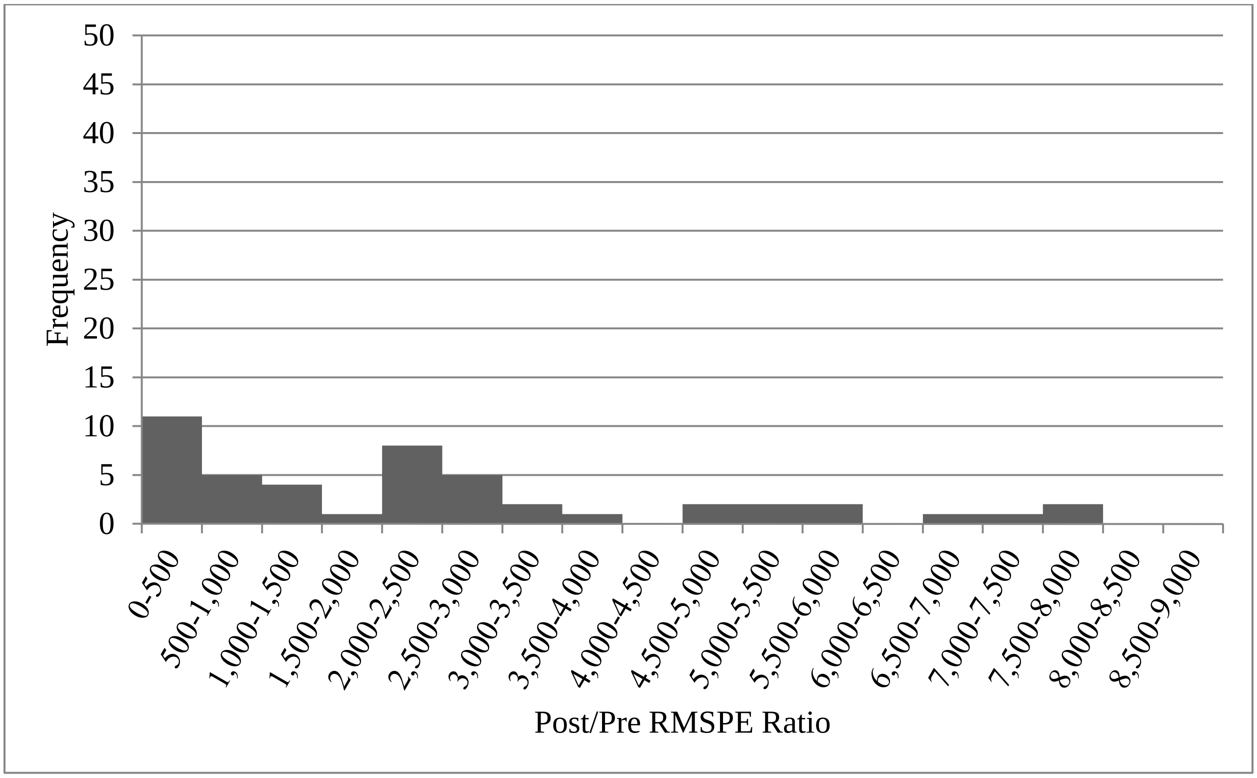
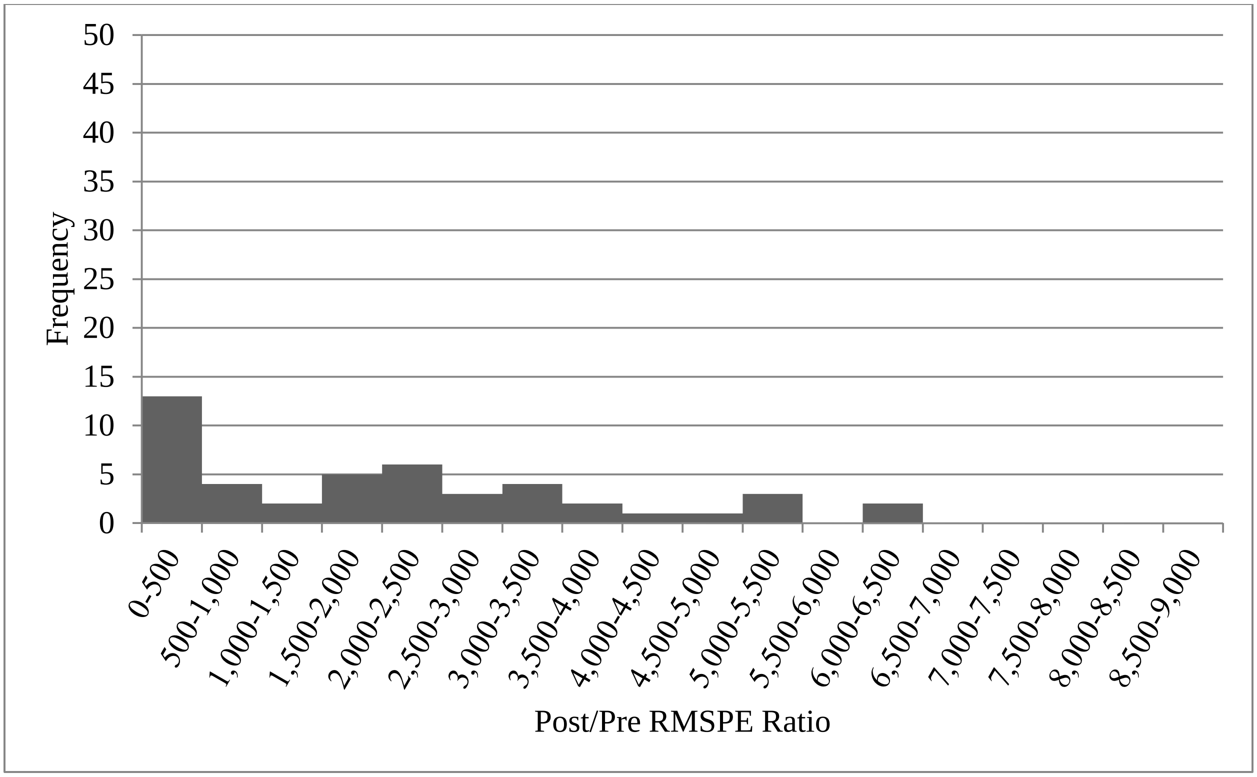


Figure 13. District #30 RMSPE ratios for Specification 1 (above) and Specification 5 (below).



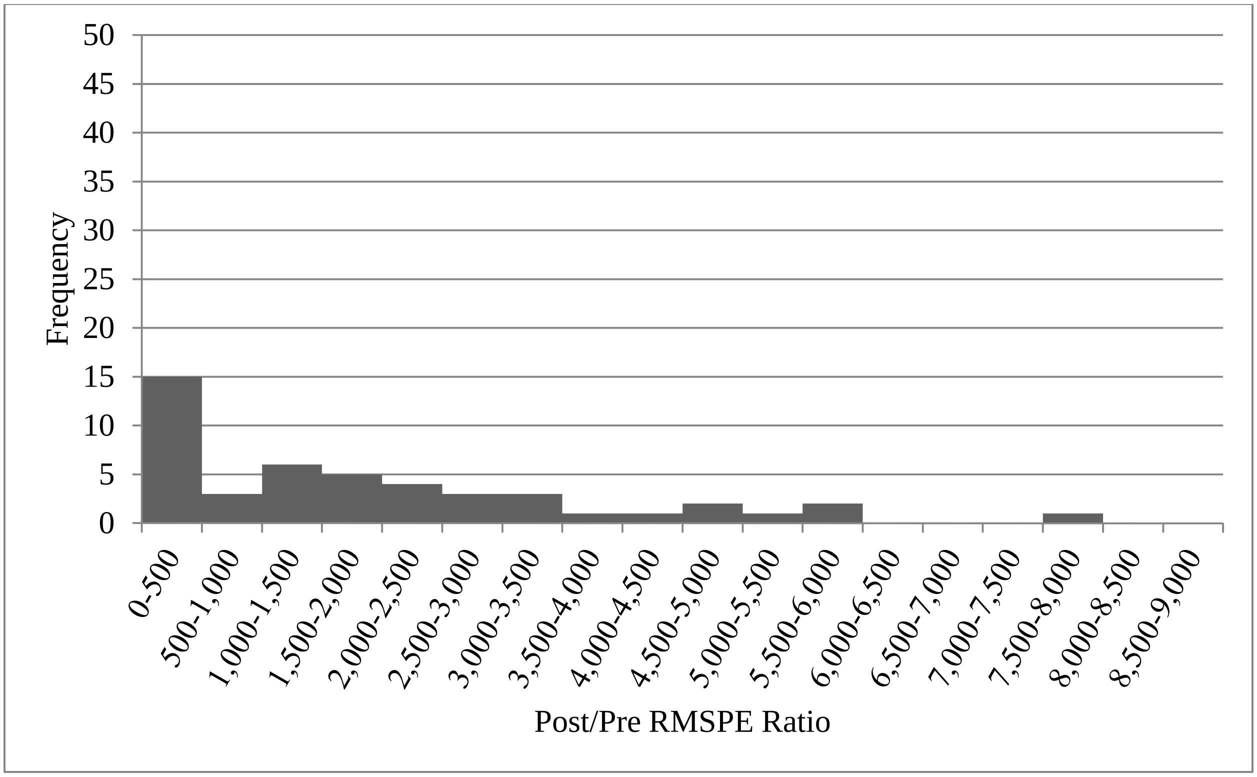
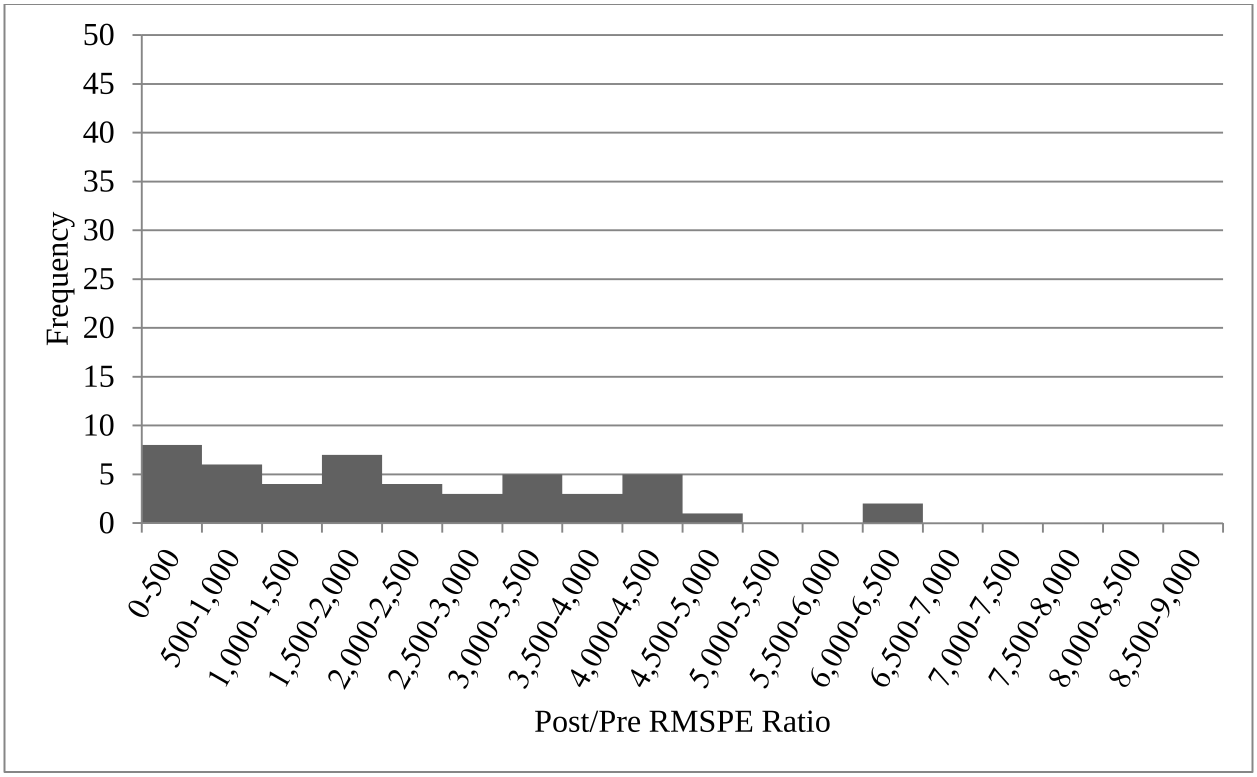


Figure 14. District #30 RMSPE ratios for Specification 9 (above) and Specification 13 (below).



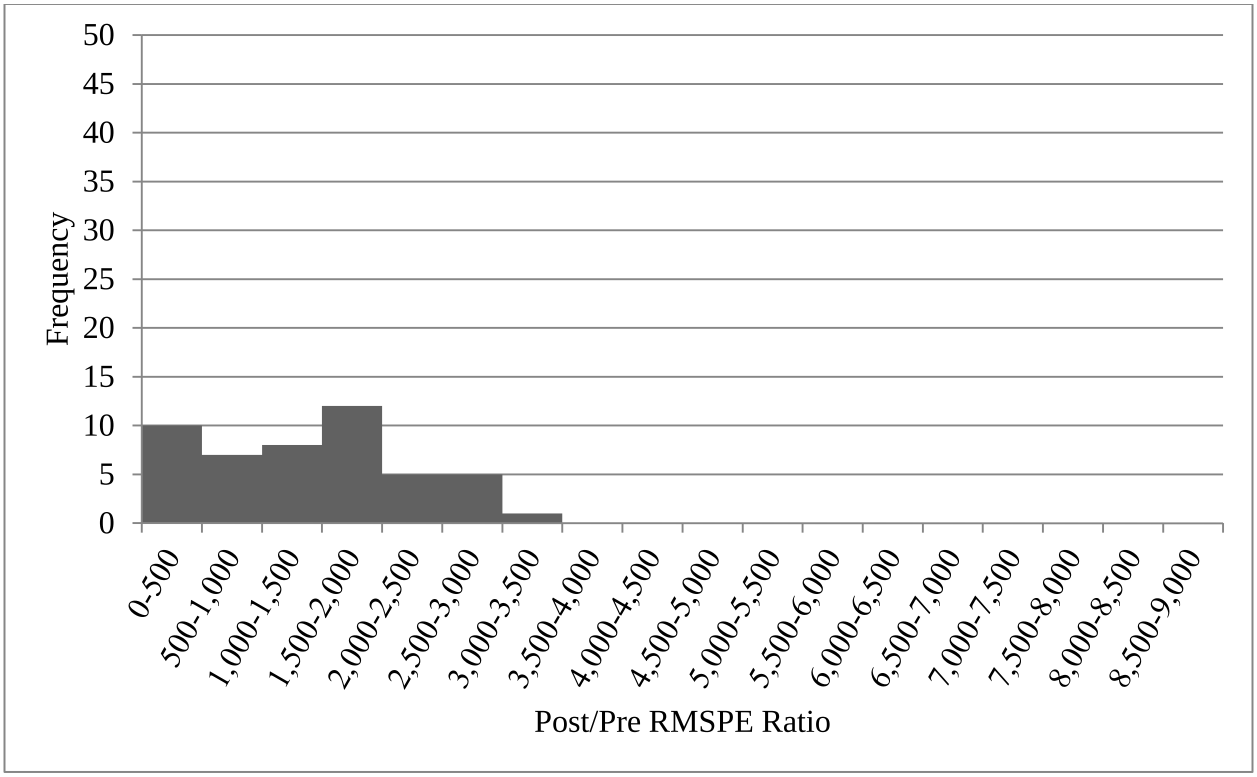
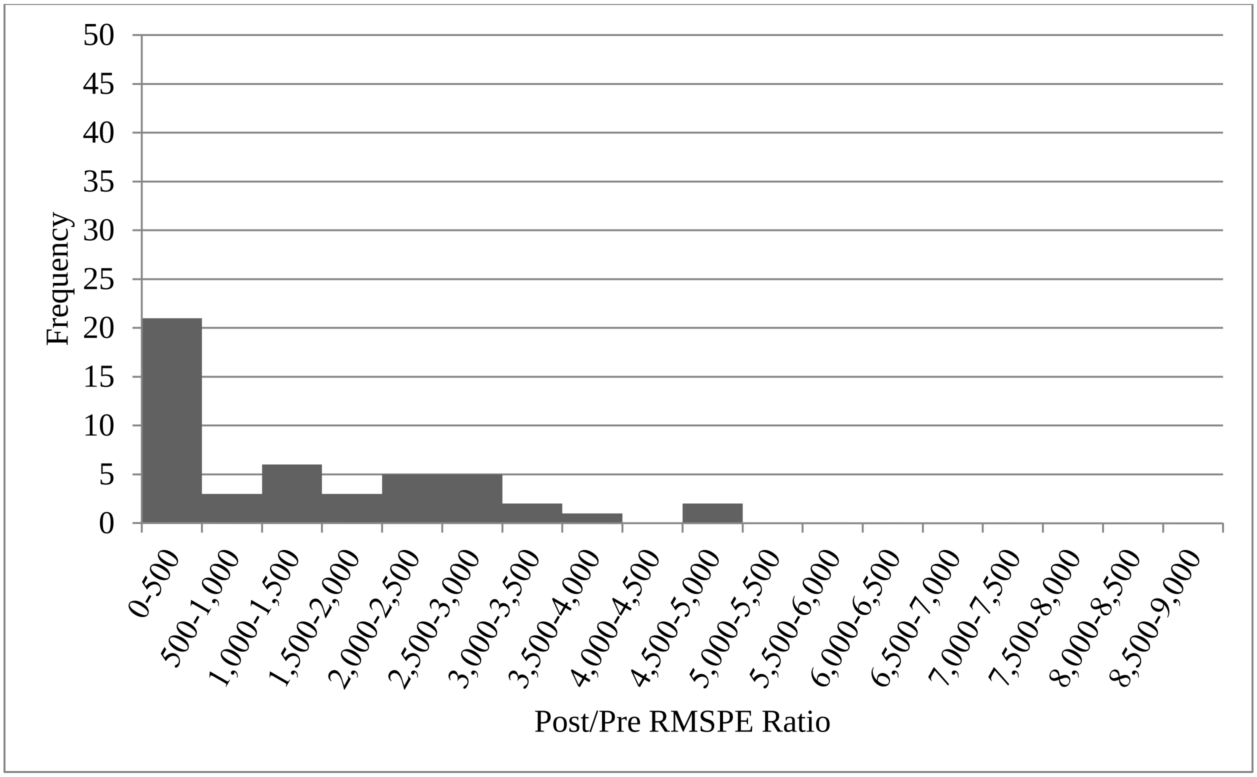


Figure 15. District #61 RMSPE ratios for Specification 1 (above) and Specification 5 (below).



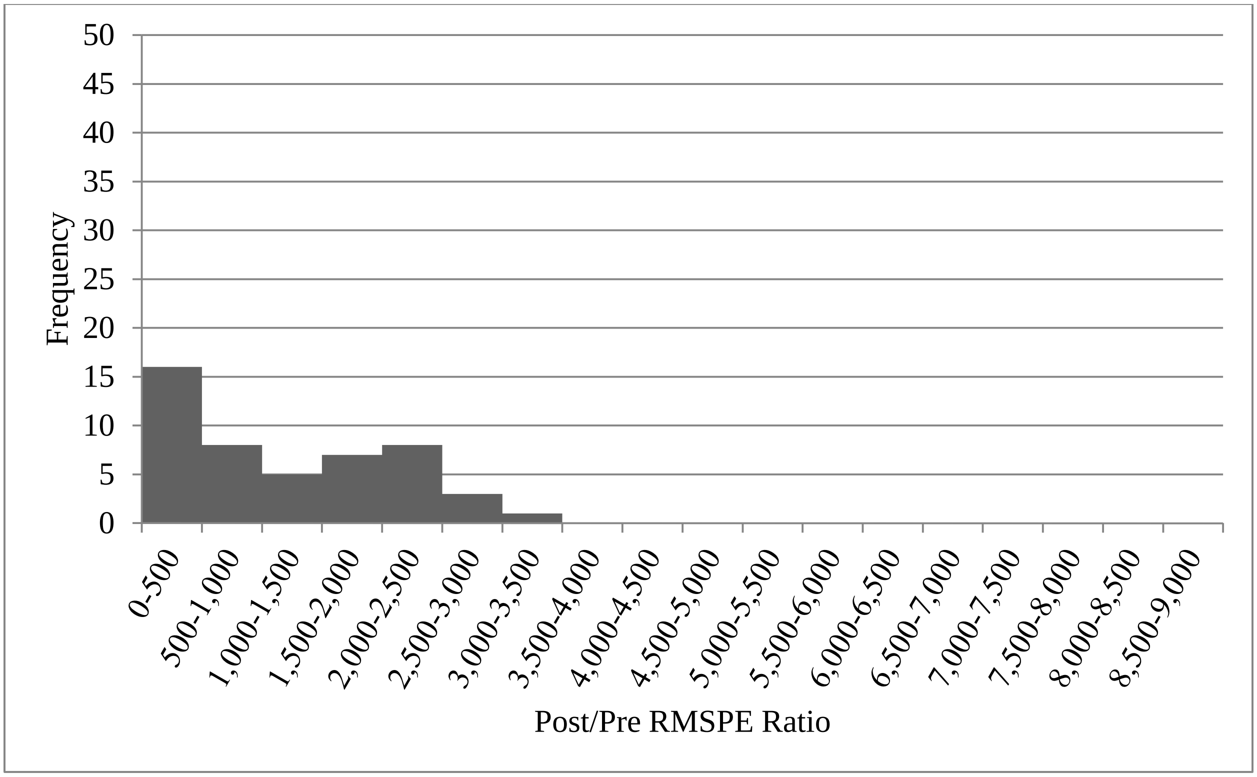
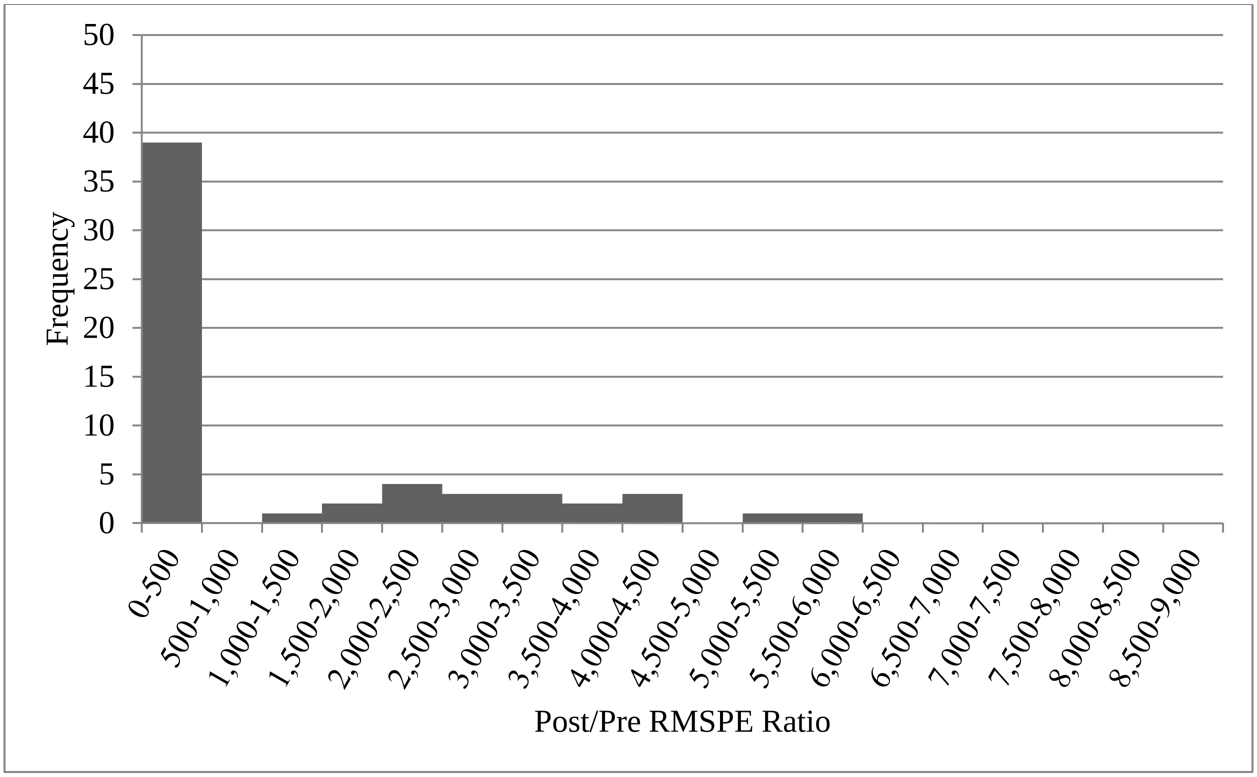


Figure 16. District #61 RMSPE ratios for Specification 9 (above) and Specification 13 (below).



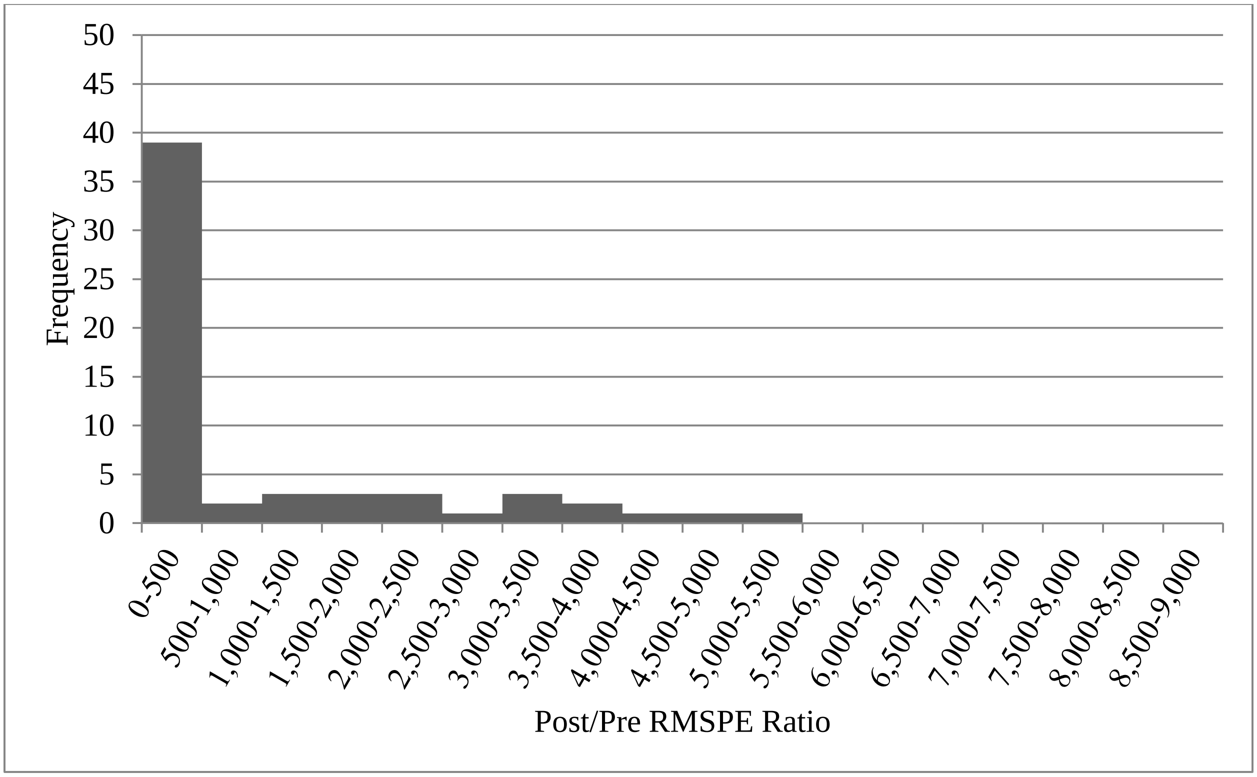
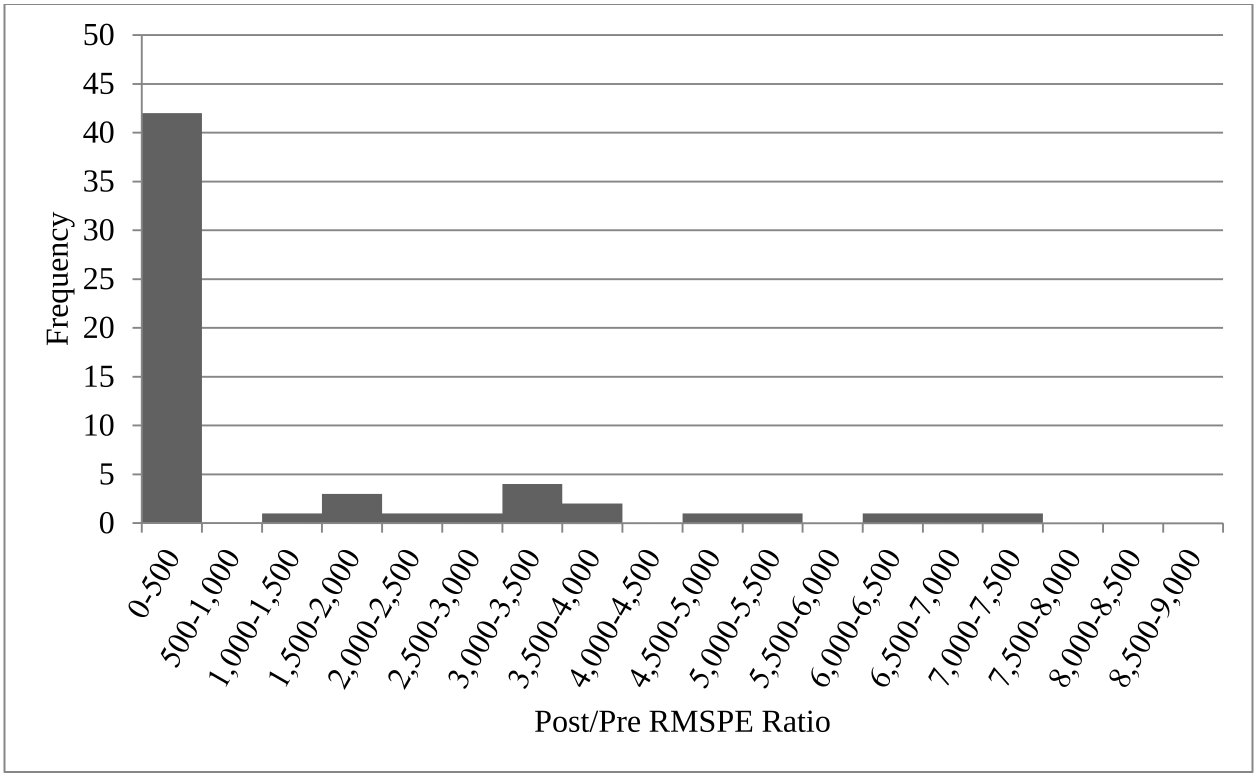


Figure 17. Locale 41 RMSPE ratios for Specification 1 (above) and Specification 5 (below).



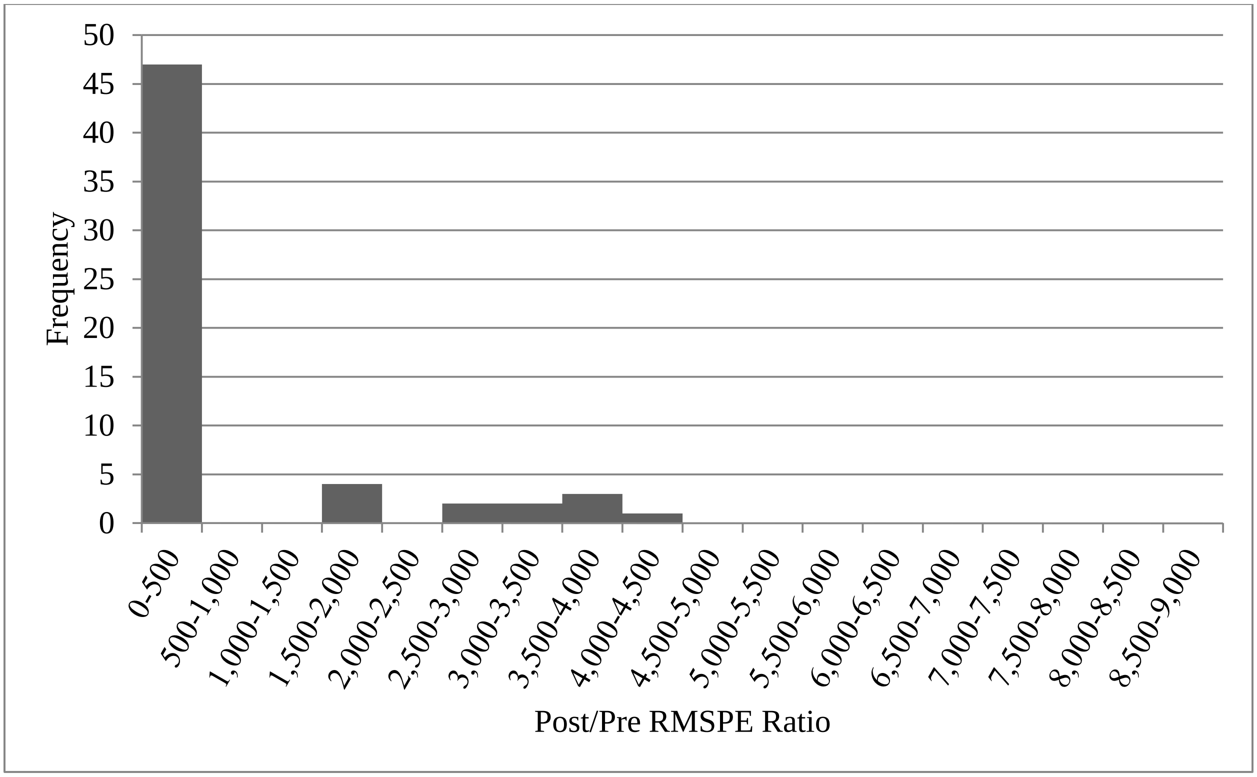
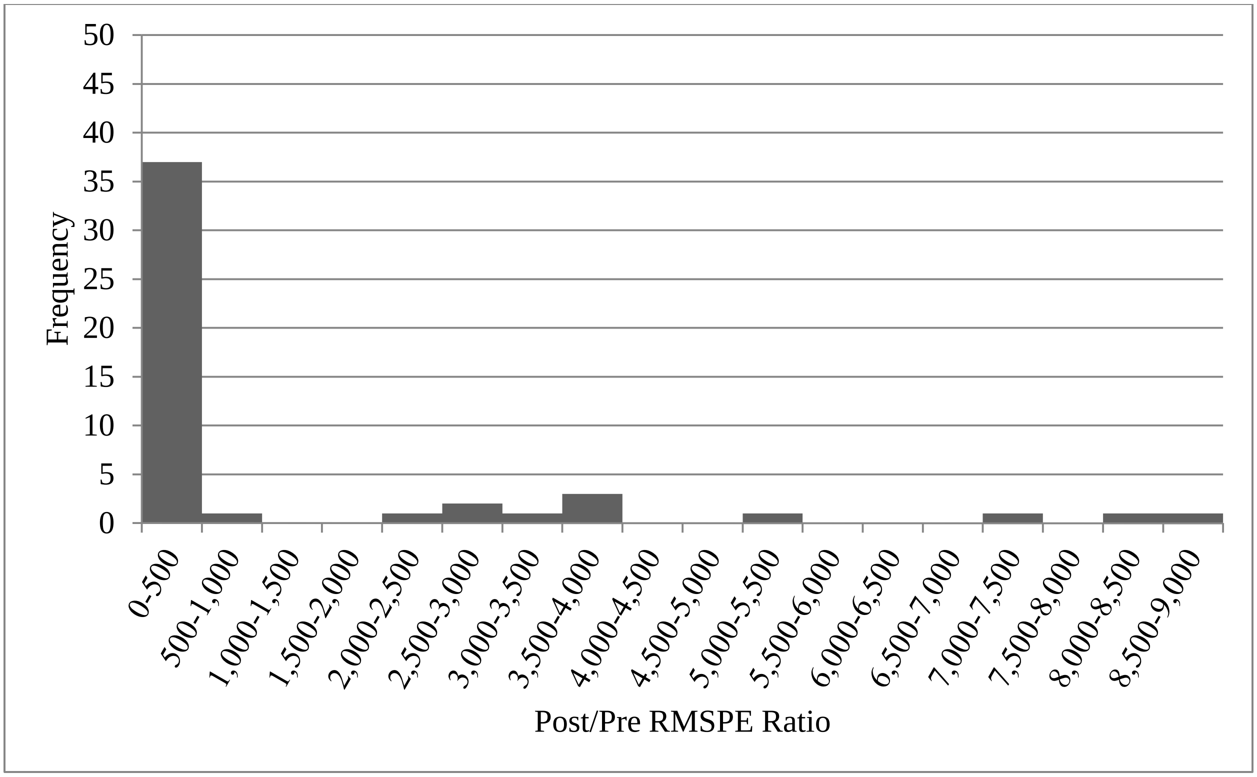


Figure 18. Locale 41 RMSPE ratios for Specification 9 (above) and Specification 13 (below).



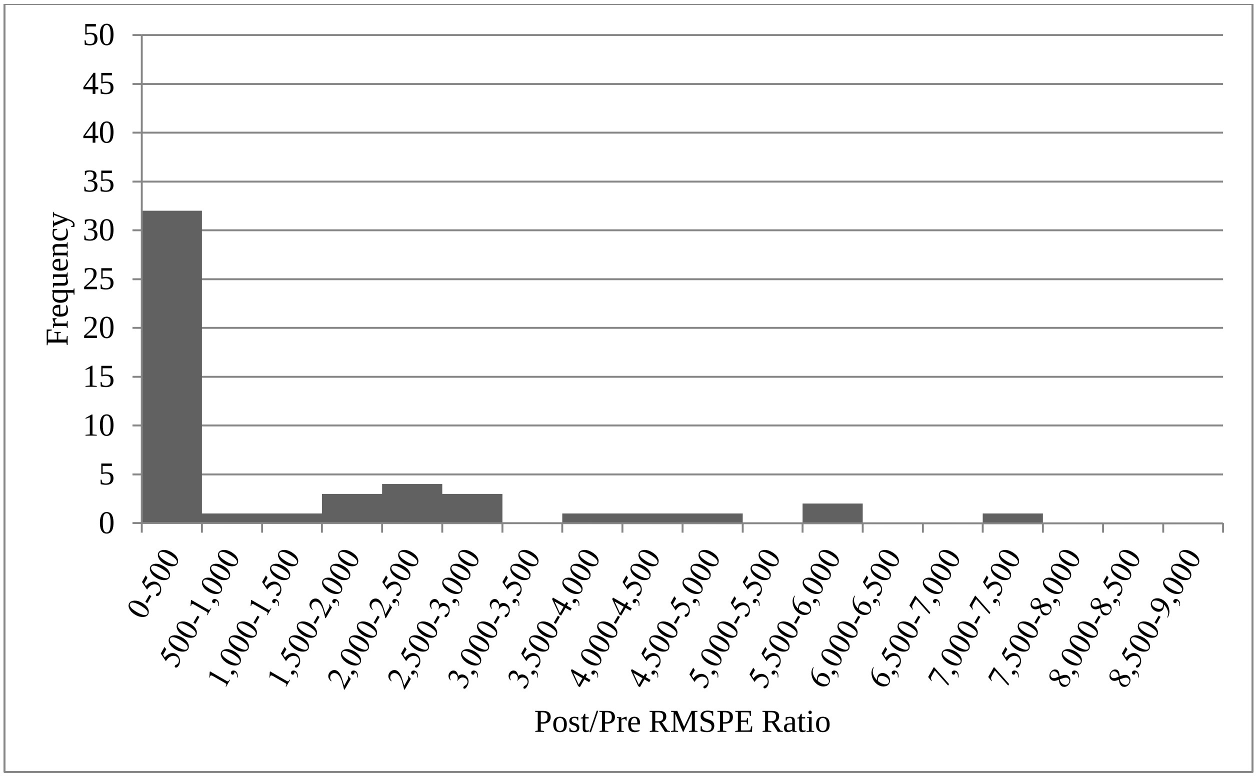
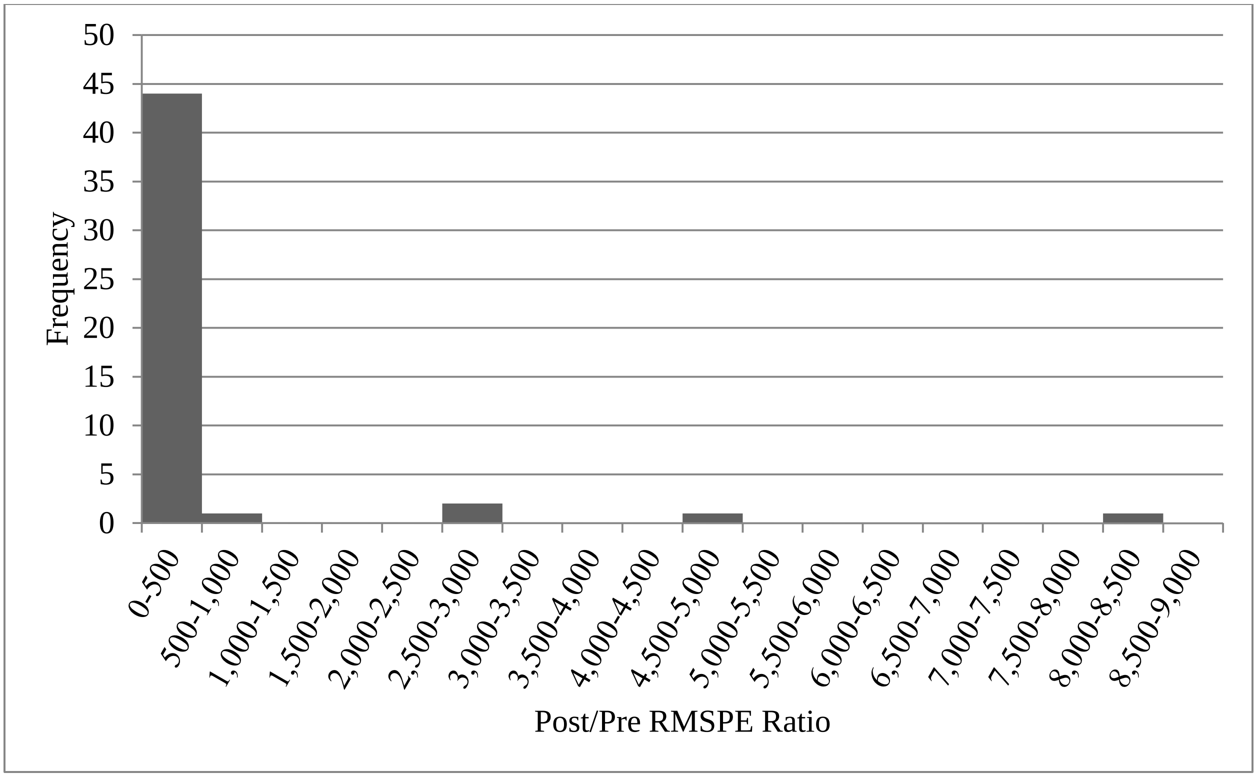


Figure 19. Locale 42 RMSPE ratios for Specification 1 (above) and Specification 5 (below).



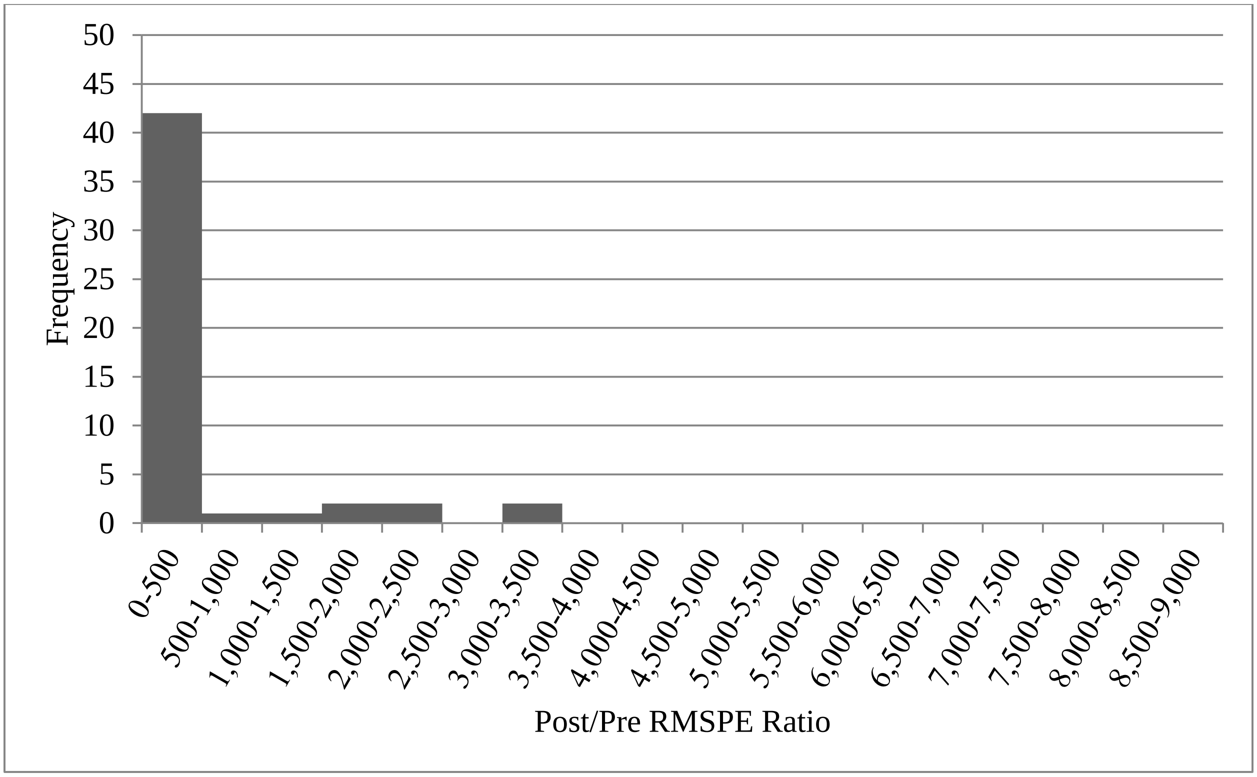


Figure 20. Locale 42 RMSPE ratios for Specification 9 (above) and Specification 13 (below).

Minimum Effect Sizes for Statistical Significance

Based on each distributional set of placebo test ratios, minimum values of posttreatment effects were found for identification of statistical significance at standard levels of α = .05 and α = .10. Given the number of comparison units included in each of the four sets of data, the following *N*th-ranked unit ratios were selected for the chosen values of α, as presented in Equation 5:

- For District #30, the 46th for $p = .043$ and the 44th for $p = .085$,
- For District #61, the 47th for $p = .042$ and the 45th for $p = .083$,
- For Locale 41, the 58th for $p = .034$ and the 55th for $p = .085$, and
- For Locale 42, the 49th for $p = .040$ and the 46th for $p = .100$.

This means that for any of the units included in each set of data to be detected as statistically significant, any treatment that unit might receive must bump its rank within its distribution to the identified position. So for example, if the 25th-ranked school unit in District #30 were to receive a hypothetical treatment, that treatment must cause the school to move to the 46th-ranked position to achieve statistical significance at 5%, or to the 44th position for 10% significance. These *N*th-place ratios are presented in Table 5 and Table 6 that follow.

For each specification with unique results within the cross-sectional data model, Table 5 gives descriptive statistics for each set of $\text{RMSPE}_{\text{pre}}$ followed by the number of close matches within each set, defined as those with $\text{RMSPE}_{\text{pre}} < 0.0001$. Next, the values necessary for statistical significance are shown. Bold values represent those required for 5% Type I error, and values in parentheses are for 10% significance. These are followed by each associated value of $\text{RMSPE}_{\text{post}}$ (the numerator of each ratio) required for one of the close-matched units to achieve the necessary ratio for statistical significance. In order to allow for comparison of patterns between sets of data with differing amounts of variability across units' original test score trends, I further transformed each required $\text{RMSPE}_{\text{post}}$ by rescaling in terms of the pooled standard deviation of posttreatment outcome measures unique to each specification. This measure of effect size is equivalent to $(Y_{observed} – Y_{synthetic}) / SD_{observed}$, and is a representation of the minimum size of difference necessary per posttreatment time point—in terms of standard deviation across comparison units within the dataset—for detection of statistical significance.

For example in regard to Specification 1, for one of the 40 close-matched schools within District #30 to receive a hypothetical treatment and be detected as statistically significant at a 5% level, its resulting RMSPE ratio must be at least as large as 8,907. As a close match with its imputed value, $\text{RMSPE}_{\text{pre}} = 0.00005$, its $\text{RMSPE}_{\text{post}}$ is required to be at least 0.4454 standardized points—an averaged difference between its actual and predicted math scores for both the spring 2011 and 2012 test events. The pooled standard deviation of school math score means in District #30 across 2011 and 2012 is 0.4927 standardized points. The quotient of the required $\text{RMSPE}_{\text{post}}$ and the observed standard deviation results in an effect size measure of 0.9040—a necessary annual gain in math scores equivalent to almost a full standard deviation of its observed annual math score differences. This represents quite a large magnitude of gain to be sustained across two years of math scores. By comparison, these annual gains must be at least 0.7075 standard deviation units for $\alpha = .10$.

All values presented apply only to units that achieved a close match of $\text{RMSPE}_{\text{pre}} < 0.0001$. These values for both cross-sectional datasets are presented in Table 5. Similar values are given for both sets of cohort model data in Table 6.

Table 5

Cross-sectional Model Results of Goodness of Fit for Matching Placebos and Minimum Differences Required for Statistical Significance

	Spec 1	Spec 5	Spec 9	Spec 13
		District #30 (*n* = 47)		
Pretreatment RMSPEs				
M	0.0191	0.0176	0.0279	0.0262
Mdn	< 0.0001	< 0.0001	< 0.0001	< 0.0001
SD	0.0691	0.0654	0.0707	0.0645
Close matches	40 (85%)	38 (81%)	34 (72%)	32 (68%)
Minimum values for statistical significance of close matches, **α = .05** (α = .10)				
Cutoff ratio	**8,907** (6,971)	**7,733** (6,699)	**6,438** (5,481)	**5,897** (5,256)
Post RMSPE	**0.4454** (0.3486)	**0.3866** (0.3350)	**0.3219** (0.2740)	**0.2948** (0.2628)
Post SD_{pooled}	0.4927	0.4888	0.4927	0.4888
ES	**0.9040** (0.7075)	**0.7911** (0.6853)	**0.6534** (0.5562)	**0.6032** (0.5376)
		District #61 (*n* = 48)		
Pretreatment RMSPEs				
M	0.0232	0.0237	0.0294	0.0301
Mdn	< 0.0001	< 0.0001	< 0.0001	< 0.0001
SD	0.1284	0.1368	0.1295	0.1369
Close matches	41 (85%)	42 (88%)	29 (60%)	35 (73%)
Minimum values for statistical significance of close matches, **α = .05** (α = .10)				
Cutoff ratio	**6,018** (4,499)	**2,863** (2,738)	**4,583** (3,299)	**2,668** (2,552)
Post RMSPE	**0.3009** (0.2249)	**0.1432** (0.1369)	**0.2291** (0.1650)	**0.1334** (0.1276)
Post SD_{pooled}	0.3550	0.3369	0.3550	0.3369
ES	**0.8475** (0.6336)	**0.4249** (0.4063)	**0.6453** (0.4646)	**0.3960** (0.3787)

Note. Values in bold reach statistical significance at α = .05; values in parentheses at α = .10. Close matches are those with $\text{RMSPE}_{\text{pre}} < 0.0001$. Specifications not shown had identical results to those given (e.g., Specs 2, 3, and 4 were no different from Spec 1). RMSPE = root mean square prediction error.

Table 6

Cohort Model Results of Goodness of Fit for Matching Placebos and Minimum Differences Required for Statistical Significance

	Spec 1	Spec 5	Spec 9	Spec 13
Locale 41 (*n* = 59)				
Pretreatment RMSPEs				
M	0.0320	0.0319	0.0383	0.0388
Mdn	0.0013	0.0213	0.0179	0.0275
SD	0.0618	0.0557	0.0593	0.0573
Close matches	20 (34%)	20 (34%)	17 (29%)	12 (20%)
Minimum values for statistical significance of close matches, **α = .05** (α = .10)				
Cutoff ratio	**5,386** (4,256)	**4,731** (3,551)	**6,605** (4,552)	**3,716** (3,223)
Post RMSPE	**0.2693** (0.2128)	**0.2365** (0.1776)	**0.3303** (0.2276)	**0.1858** (0.1611)
Post SD_{pooled}	0.4262	0.3501	0.4292	0.3520
ES	**0.6320** (0.4993)	**0.6757** (0.5072)	**0.7694** (0.5303)	**0.5278** (0.4577)
Locale 42 (*n* = 50)				
Pretreatment RMSPEs				
M	0.0376	0.0358	0.0487	0.0454
Mdn	0.0117	0.0197	0.0344	0.0271
SD	0.0534	0.0508	0.0523	0.0516
Close matches	13 (26%)	18 (36%)	6 (12%)	8 (16%)
Minimum values for statistical significance of close matches, **α = .05** (α = .10)				
Cutoff ratio	**8,771** (5,212)	**5,932** (4,317)	**8,470** (2,745)	**3,141** (1,940)
Post RMSPE	**0.4386** (0.2606)	**0.2966** (0.2159)	**0.4235** (0.1372)	**0.1571** (0.0970)
Post SD_{pooled}	0.4756	0.3449	0.4631	0.3376
ES	**0.9221** (0.5479)	**0.8601** (0.6260)	**0.9145** (0.2963)	**0.4652** (0.2873)

Note. Values in bold reach statistical significance at α = .05; values in parentheses at α = .10. Close matches are those with $\text{RMSPE}_{\text{pre}}$ < 0.0001. Specifications not shown had identical results to those given (e.g., Specs 2, 3, and 4 were no different from Spec 1). RMSPE = root mean square prediction error.

Chapter V: Discussion

Although fewer combinations of specifications per dataset contributed unique results than were proposed initially, general patterns of necessary minimum effect sizes emerged in comparison across specifications and datasets. The resulting calculations presented in Table 5 and Table 6 are those discussed here, with regard to patterns of pretreatment match and to the necessary values for statistical significance of a close-matched unit receiving a hypothetical treatment.

Patterns of Results

Influence of covariate measures

The most obvious pattern that emerged from the placebo test analyses was mentioned briefly in the Results section: that the addition of demographic and achievement covariate predictors had no impact on the computation of synthetic control weights. Specifications as originally proposed that included the secondary achievement measure and a matrix of available demographic indicators gave results no different from those whose only predictor was pretreatment values of the outcome achievement measure. This means that the initial set of 64 cases—16 specifications across four datasets—collapsed to 16 unique cases, or four specifications for each dataset. Therefore all findings from Specifications 2, 3, and 4 were redundant and are shown in the Results section as *Specification 1*. In turn, Specifications 5, 9, and 13 represent all unique sets of findings due to the same absence of covariate influence.

While this lack of effect of including covariates is quite different from the seminal examples from political science applications, the phenomenon should come as no surprise in light of the high intercorrelations of the achievement measures as shown in Table 3 and Table 4. With event-to-event, same-subject aggregate test score correlations rarely below *r* = .85 and sometimes as large as *r* = .96, corresponding demographic measures have little chance of offering additional predictive power for matching. By comparison, only the demographic measures of percentage of students of minority racial status and the percentage of students participating in free or reduced-price lunch programs maintain statistically significant, negative correlations with outcome measures of achievement across all datasets. Even with values as large in magnitude as *r* = -.85 to *r* = -.88 within the cross-sectional model for District #30, pretreatment measures of achievement consistently reach even higher levels of association.

The relatively extreme correlations among repeated measures of student achievement, particularly when aggregated over related clusters of students, overrule any concerns for bias amplification on the part of the researcher due to the inclusion of inappropriate covariate measures. By implication, any noise observed in these placebo test analyses beyond observed patterns of effect size must be due to achievement trends of true school- or cohort-level units and not to the reduction or inflation of unobserved bias due to matching. Whereas in the economic examples from Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) where the concern is with including sufficient subsets among many available covariates for the reduction of influential bias, my results across specifications were only sensitive to the inclusion of additional pretreatment test events, since near-perfect matches were common by prediction on pretreatment measures of the outcome only.

Necessity of close synthetic matches

A crucial pattern that emerged from the minimum effect sizes necessary for a statistically significant finding is the importance of a treated unit achieving a close match. Across all specifications and pools of comparison units, every minimum ratio identified as a cutoff for statistical significance resulted from a control unit with a close match across the pretreatment period. Each had a value of $\text{RMSPE}_{\text{pre}} < 0.0001$, or essentially zero given four significant digits. These were the set of comparison units with the value of 0.00005 imputed for pretreatment match in order to avoid division by zero, although my choice of this imputed value is not to blame for the necessity of near-perfect matches.

The distribution of $\text{RMSPE}_{\text{pre}}$ calculated in Specification 1 for the Locale 41 (cohort model) dataset exemplify this finding. As given in Table 6, 20 of the 59 comparison cohorts achieved close matches with $\text{RMSPE}_{\text{pre}} < 0.0001$. The next smallest value of this pretreatment fit was an $\text{RMSPE}_{\text{pre}} = 0.0002$. This is four times the arbitrarily small value of 0.00005 chosen for imputation in place of zeros. This means that the associated effect size necessary for 5% statistical significance, ES = 0.9040, must increase by a factor of four for this next-best matched unit to reach statistical significance. This results in a minimum effect size, ES = 3.616. In comparison with observed magnitudes of unit-level standard deviations, this size of necessary effect is clearly outside of reasonable expectation for increases in annual student achievement.

Notice that, even had I imputed a larger value for close-matched units—say RMSPE = 0.0001, or twice the value selected—the resulting effect size would still remain unreasonably large. The associated minimum cutoff ratio would have been 4,386, and would require that the next-smallest match demonstrate an unreasonably large effect size of 1.844. Since every other set of placebo tests had a next-smallest $\text{RMSPE}_{\text{pre}}$ value at least twice this large (e.g., RMSPE = 0.0004 for Locale 41, Specification 9), the implication is that only close-matched comparison units ($\text{RMSPE}_{\text{pre}}$ values no different from zero) had a reasonable chance of achieving statistical significance due to treatment. In general this would suggest that a pool of control units with a proportion of close matches larger than the desired level of statistical significance would require a close match for the treated unit as well. That is, if more than 10% of comparison units result in close matches, a treatment unit must also achieve a close match with its synthetic control in order to reach a 10% level of significance.

Number of pretreatment time points

A third evident pattern that emerges from the necessary minimum effect sizes lies in the influence of the number of time points matched across the pretreatment period. In general, as more time elements are included in pretreatment, $\mathrm{RMSPE}_{\mathrm{pre}}$ increases, and the total number of closest matching placebo tests decreases. Regardless of whether this results in improved matches on a qualitative level, in quantitative terms the creation of perfect synthetic matches becomes more difficult.

As shown in Table 5, the cross-sectional model placebos matched on three years of pretreatment scores consistently result in fewer close matches and larger values of $\mathrm{RMSPE}_{\mathrm{pre}}$ as compared with two pretreatment time points. This trend was shown to continue in general for cohort model matches, where increasing the matching time period from five events to six further reduced the frequency of close matches, seen in Table 6. Unfortunately this pattern's appearance in the cohort data format served to overrule the intended comparison of an arbitrary shift in the time point of treatment, where I hypothesized a lack of sensitivity to time of treatment. Further analyses might shift the time of treatment within the available data while maintaining the total number of measured time periods both pre- and posttreatment to better investigate whether results are robust to this specific factor.

While close matches were more difficult to find where more pretreatment points were included, necessary effect sizes for identification of statistical significance were reduced overall. So generally, it might be concluded that close matches across more time points are actually better matches than those across narrower pretreatment intervals. Smaller treatment effects would be necessary for detection when compared with close matches over longer pretreatment periods. A clear departure from this pattern is demonstrated by Specification 9 for the Locale 41 data within the cohort model however. As shown in Table 6, when compared with Specification 1 where math is the outcome matched over five pretreatment measures, six pretreatment events actually led to larger necessary RMSPE ratios and final effect sizes. While pretreatment matches were still less close as a result, posttreatment gaps needed to be larger. Individual well-matched cohort units tended to remain well matched when a sixth pretreatment point was added. However, many of their average posttreatment gaps increased as posttreatment events decreased from four to three periods. This occurred often enough that the overall ratio distribution led to larger necessary effect sizes for six pretreatment times compared with five. By comparison, the opposite trend occurred for identical specifications across the Locale 42 set of data. The implication is that inherent noise within the pool of comparison units is capable of leading to unpredictable patterns of results for synthetic controls matches.

Although the comparison of effect size across specifications reflects standardization across dataset and achievement measure, erratic patterns of results still give evidence of other sources of noise. For example, math scores tended to demonstrate more initial variability than reading scores, which was taken into account by larger pooled standard deviations in effect size calculations. But overall trends still suggest that math gaps require larger effect sizes than reading for statistical significance with synthetic controls. The evident source of noise that remains is in the trend of repeated measures. Inspection of initial achievement trends in Figure 1through Figure 4 suggest less overall variability in reading scores among units but also smoother trends over time as compared to math scores. Season-to-season math scores show more erratic trends overall. In turn, sets of smoother trends led to more stable sets of placebo gaps and finally to smaller required effect sizes for statistical significance. These trends seem to contribute the additional sources of noise leading to unpredictable matches after standardization of results.

Results for Locale 42 data within the cohort model also appear to deviate from the overall pattern of smaller necessary effect sizes' association with more pretreatment points. Also shown in Table 6, the largest calculated effect sizes appear in connection to these specifications for a 5% level of statistical significance, even reaching larger values than many of those within the cross-sectional model. Effect size results associated with statistical significance at $\alpha = .10$ are more in line with larger patterns however. This effect is due to a very few cohort units within Locale 42 that have both an unusually close level of pretreatment match at the same time as an unusually large posttreatment effect. These units appear more as outliers within their own distribution, but are common enough to require a treated unit to outrank in comparison. For 10% significance as these few outlying ratios are disregarded, patterns of effect size fall in line with expectation.

In total, larger sets of highly similar comparison units offer greater potential for successful synthetic control matching, as expected. Greater variability among comparison units and stability across repeated measurements led to closer matches that allow for detection of smaller effects. Furthermore, where outcome measures made it possible to achieve near-perfect pretreatment match, the necessity of a similar goodness of fit for a treatment unit of interest is critical. Next I formally restate these general conclusions in terms of my original hypotheses.

Conclusions Regarding Hypotheses

Pretreatment measures

I hypothesized that the addition of pretreatment measures of student- and school-level demographics as covariate predictors would improve the fit of synthetic control matches. This was based on recommendations in literature for the inclusion of all available data for reduction of bias for causal inference, such as by Rubin (2006). Instead I found that matches formed using only pretreatment measures of achievement were so successful that no other predictors contributed to improved fit. This was due in large part to extremely high intercorrelations among repeated measures of aggregate MAP achievement scores.

Pretreatment time factors

Second—based on the same suggestions from the literature—I hypothesized that access to additional pretreatment events would improve goodness of fit between synthetic controls and observed data. The addition of a third year of pretreatment achievement showed the very opposite in terms of RMSPE for fit; $\text{RMSPE}_{\text{pre}}$ values tended to increase overall. However, required effect sizes for statistical significance decreased in turn, suggesting a benefit of qualitatively improved synthetic fit in spite of quantitative increases in the measure of lack of fit.

As for my hypothesis of cohort results being robust to arbitrary shifts in treatment time point, findings were largely overruled by the additional difficulty of matching across a sixth pretreatment point. As discussed above, a future model respecification might better untangle the sensitivity of the model to treatment shift from the effects of added pretreatment points.

Synthetic match level

Overall patterns of effect size calculations suggested that neither the cross-sectional nor the cohort model applied synthetic controls more effectively. However, the effects of matching across a larger set of pretreatment time points had most significant influence on results, as discussed above. Comparison units formatted to demonstrate more stable trends over time allowed for the closest matches and required smaller effect sizes in return.

Synthetic control donor pool

I further hypothesized that synthetic control matches would be robust to choice of comparison units. I examined this theory by forming pairs of comparable pools of data. Overall patterns of final effect size supported this hypothesis in general, with statistical noise remaining in the results due to influential outlier units and erratic achievement trends. As just repeated, sets of units demonstrating consistency in achievement trends led to smaller necessary size of effects as compared with groups showing trends that were less smooth over time.

Research Limitations

As only 16 of the originally proposed 64 specifications resulted in unique analyses, the final set of patterns available for observation was reduced by 75% as well. Because of this, the patterns emerging from effect size calculations could more likely be due to random statistical fluctuations than would be noticeable within a larger set of comparable results.

While every effort was made to create duplicate sets of comparable comparison units for both the cross-sectional and the cohort data models, education and testing policies in place in South Carolina over the period of data collection may have had influence on the consistency of the available measures of achievement. South Carolina data were chosen for analysis due to convenient data availability, as MAP exams are administered to all but a few of this state’s public school students. However policies and stakes should not be assumed constant from district-to-district or year-to-year. Some exams may have been administered in a no-stakes environment while other locations or testing seasons may have been tied to incentives for students or teachers. While it could be argued that these differences should largely be balanced away in aggregate, the possibility of the influence of these realities of achievement data must be considered.

Further, limits on the generalizability of findings should be addressed. As all results were associated with only a single, specific measure of student achievement—MAP—implications must be interpreted accordingly. MAP was selected for these analyses under the assumption that the sophistication of the scale offers a best-case scenario among measures for forming matched comparisons of student achievement. Other scales, such as state-administered annual proficiency exams or standards-based interim assessments—where comparison across locations or grade levels might require rescaling for comparison—likely add additional sources of measurement error and noise to sets of comparison units. It is likely that where matches were worse or more difficult to find using MAP as outcome, other measures of student achievement might fare no better. It is expected that outcome measures with high intercorrelations such as those seen with MAP would perform similarly in synthetic controls analyses, while measures with lower levels of association over time would perform less successfully. Further, other alternatives to cross-sectional or cohort comparisons of aggregate student performance might demonstrate larger or smaller magnitudes of intercorrelations, leading to patterns of results different from my own. I conclude with the assumption that concerns with the effectiveness of synthetic matches using MAP scores are likely no less problematic for other student achievement measures.

Finally, while these methods allowed me to examine patterns of behavior for matching over many datasets and model specifications, the validity of accuracy in causal inferences cannot be directly addressed. In order for direct causal claims to be assessed, access to comparisons with true randomized experimental designs would be necessary. I further discuss this possibility in the section containing suggestions for future research that follows.

Further Research

Beyond the research limitations just presented, this study only begins to give insight into the behavior of synthetic controls in combination with student achievement data. Of greatest interest is the validity of the quasi-experimental method in replicating causal effects as detected in true randomized control experimental research. As appearances of high quality experimental evaluations increase within educational research, within-study comparisons of quantitative results compared among alternative analysis methods become more feasible. As was given for example in a large scale within-study analysis by Mathematica Policy Research (Fortson et al., 2012), the accuracy of causal conclusions from synthetic controls matches could be investigated most directly by comparison among other alternative observational methodologies alongside the results of a true random assignment study. The lack of availability of detailed data from existing experimental studies using MAP achievement measures made within-study comparison impossible for the current paper. Future opportunities to tie synthetic controls to accurate causal implications would provide the best insight into issues of necessary data availability and sufficient reduction of bias in matching as well. As presented by Cook, Shadish, and Wong (2008) the within-study comparison literature provides best evidence of the ability of observational studies to generate unbiased alternatives to randomized experiments.

Further analyses for establishing the generalizability of the current findings are justified as well. Future studies might replicate these data models and variable specifications using measures of student achievement other than MAP for comparison. Besides analyzing additional academic measures, adjustment of intercorrelations among available covariates could be performed through either manipulation of software defaults or by simulation of data. In this way baseline levels of necessary correlations among predictors for improved matches would further inform not only educational evaluations but also those using measures with smaller magnitudes of association. Simulated data that would allow comparison of units with erratic time trends against others with trends exhibiting more stability would assist further in generalizing the current findings.

Analysis of models alterative to the cross-sectional and cohort designs presented here would further inform generalizability of synthetic controls as well. Future research might compare school-level analyses alongside similar units at the larger district level or the smaller teacher level. Teacher-level analyses might offer further insight into models for teacher evaluation such as those using value added approaches. District-level analyses could inform report card performance evaluations that are currently popular. Furthermore, although the method's intended purpose is to match in aggregate, analyses of student performance at the individual level might provide a useful alternative methodology for some applications. As most of the demographic measures included here served more as constant indicators than as varying trend measures, these aggregated data performed no differently than student-level covariates might. Although the total number of individual students for comparison would be restrictive, there is no practical reason that a student-level analysis could not be performed. Results could inform the performance of the aggregate measures for matching by comparison.

Implications for Educational Evaluations

While these analyses and results may raise more follow-up questions than they definitively answer, the findings offer some practical implications for researchers using synthetic controls for comparing student achievement outcomes. To consider the comparability of an evaluation with the data specifications presented here, intercorrelations among the available measures, the size of standardized gains after treatment of the aggregated unit of interest, and the aggregate-level standard deviation of the outcome measure should be examined.

In comparison to descriptive statistics presented in Table 3 and Table 4, analyses of achievement measures with high aggregated correlations from event-to-event—and comparably low correlations with available demographic covariates—should be expected to behave similarly in creating synthetic controls. Evaluation of an intervention that includes similar data as these should expect demographic measures to be inconsequential in matching as compared with more highly correlated measures of prior achievement. Further, as these correlations are expected to be high for aggregate measures, the researcher should be aware of the possibility of finding near-perfect matches across pretreatment times, i.e., $\text{RMSPE}_{\text{pre}}$ near or no different from zero. This means that it may be necessary for some real value, however small, to be substituted for RMSPE values of near-perfect matches to avoid undefined values for RMSPE ratios and make the inferential process impossible.

Although the results given here suggest that creating near-perfect matches across more pretreatment time points is more difficult than with fewer events, the researcher should include as many pretreatment measures of the outcome measure as possible. My results suggest that this reduces the effect sizes necessary for detection of statistical significance. However the researcher's primary goal is to reduce the value of $\text{RMSPE}_{\text{pre}}$ between the treated unit and its synthetic control. This means that, while close matches across more pretreatment time points is ideal, some test events may have to be excluded in the interest of reaching goodness-of-fit values near zero in light of my findings that less than perfect matches may make statistical significance impossible to achieve.

Next, average z-score differences for a treated unit should be compared to the values of estimated post RMSPE across specifications given in Table 5 and Table 6. These offer standardized estimates of the size of necessary gaps after treatment has been administered for statistical detection. Since these were calculated assuming that the treated unit achieved a near-perfect match, these values only apply when the creation of a close-matched synthetic control is possible. But for close-matched cases, these values of $\text{RMSPE}_{\text{post}}$ are comparable to differences necessary among student-level test scores.

It should be noted that the final effect size calculations reported in Table 5 and Table 6 were calculated at an aggregated level in order to compare across the multiple specifications and cannot be interpreted to represent student-level differences. For the researcher, the variability among the available pool of aggregated comparison units should be examined with respect to these estimates of effect size. The inclusion of comparison units that are too dissimilar to the treated unit may be responsible for further diminishing close matches across the pretreatment period. Also outcome measures that tend to behave more erratically over time can negatively impact matching as well. As presented above, any efforts toward lowering $\text{RMSPE}_{\text{pre}}$ for the treated unit—including adding or omitting units from the comparison pool—are immune from criticism for data fishing so long as the researcher avoids being influenced by values of posttreatment measures.

Finally, the level of sensitivity available in achievement measures in the form of significant figures must be considered to avoid artificially inflating ratios by compounding rounding error since values of $\text{RMSPE}_{\text{pre}}$ near zero are possible. In addition, a researcher should look for a possible loss of sensitivity in results due to unexpected rounding steps within the software package syntax—at least in the case of analyses using Synth version 9.2 for Stata. An evaluation of academic achievement measures using synthetic controls that is able to show close matching across a pretreatment period along with a comparably extreme value of $\text{RMSPE}_{\text{post}}/\text{RMSPE}_{\text{pre}}$ for statistical significance should be in position to present a well-supported, intuitive case for meaningful group-level effects of an educational intervention.

Contributions to Education Research

Statistical methods for the social science disciplines are a young field of study. Effective, robust, accessible methodologies for addressing the types of questions that arise in educational research are crucial to efforts toward establishing the body of scientifically based research that is the stated goal of the National Center for Education Statistics and the U.S. Institute of Education Sciences. As new methodological techniques arise across the social sciences, and as cross-pollination increases between disciplines that have been traditionally segregated, the need for investigation and validation of new methods increases as well. In response, the goal of this paper is to offer insight into the utility of a recently developed, innovative statistical method as it applies to the specific analysis needs and data availability of educational evaluations.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2007). *Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program* (Working Paper No. 12831). Washington, DC: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w12831

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. doi:10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A., & Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, *42*(13). Retrieved from http://www.jstatsoft.org/v42/i13/paper

Abadie, A., Diamond, A., & Hainmueller, J. (2012). *Comparative politics and the synthetic control method* (Research Paper No. 2011-25). Cambridge, MA: MIT Political Science Department. Retrieved from http://ssrn.com/abstract=1950298

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *The American Economic Review*, *93*(1), 113–132. doi:10.1257/000282803321455188

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, *76*(6), 1537–1557. doi:10.3982/ECTA6474

Almer, C., & Winkler, R. (2011). *The effect of Kyoto emission targets on domestic CO2 emissions: A synthetic control approach* (SSRN Scholarly Paper No. 1752282). Rochester, NY: Social Science Research Network. Retrieved from http://ssrn.com/abstract=1752282

Bassok, D., Fitzpatrick, M., & Loeb, S. (2012). *Does state preschool crowd-out private provision? The impact of universal preschool on the childcare sector in Oklahoma and Georgia* (Working Paper No. 18605). Washington, DC: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w18605

Belot, M., & Vandenberghe, V. (2009). *Grade retention and educational attainment: Exploiting the 2001 reform by the French-speaking community of Belgium and synthetic control methods* (Discussion Paper No. 2009022). Louvain-la-Neuve, Belgium: Université catholique de Louvain, Institut de Recherches Economiques et Sociales. Retrieved from http://ideas.repec.org/p/ctl/louvir/2009022.html

Betts, J., Levin, J., Miranda, A. P., Christenson, B., Eaton, M., & Bos, H. (2010). An evaluation of alternative matching techniques for use in comparative interrupted time series analyses: An application to elementary education. In *Festschrift for Professor Charles Beach*. Queen's University, Kingston, Ontario, Canada. Retrieved from http://jdi.econ.queensu.ca/sites/default/files/Matching%20Study%20Report%2003-23-10.pdf

Brady, H. E. (2008). Causation and explanation in social science. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 217–270). New York, NY: Oxford University Press.

Coffman, M., & Noy, I. (2012). Hurricane Iniki: Measuring the long-term economic impact of a natural disaster using synthetic control. *Environment and Development Economics*, *17*(2), 187–205. doi:10.1017/S1355770X11000350

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199. doi:10.3102/01623737024003175

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724–750. doi:10.1002/pam.20375

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, *95*(3), 932–945. doi:10.1162/REST_a_00318

Economist 5ae7. (2012, March). "Synthetic control" modeling questions [Online forum comment]. Retrieved from http://www.econjobrumors.com/topic/synthetic-control-modeling-questions

Eren, O., & Ozbeklik, I. S. (2011). *Right-to-work laws and state-level economic outcomes: Evidence from the case studies of Idaho and Oklahoma using synthetic control method* (Working paper). Retrieved from University of Nevada, Las Vegas website: http://faculty.unlv.edu/oeren/eren_ozbeklik_paper2.pdf

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver and Boyd.

Fitzpatrick, M. D. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. *The B.E. Journal of Economic Analysis & Policy*, *8*(1), 1–38. doi:10.2202/1935-1682.1897

Fortson, K., Verbitsky-Savitz, N., Kopa, E., & Gleason, P. (2012). *Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates* (NCEE Report No. 2012-4019). Retrieved from http://ies.ed.gov/ncee/pubs/20124019/

Gelman, A. (2009, July). Resolving disputes between J. Pearl and D. Rubin on causal inference. *Statistical modeling, causal inference, and social science* [Web log post]. Retrieved from http://andrewgelman.com/2009/07/disputes_about/

Griliches, Z. (1985). Data and econometricians--The uneasy alliance. *The American Economic Review, 75*(2), 196–200. doi:10.2307/1805595

Hainmueller, J. (2011). Synth (Version 9.2) [Stata software package]. Retrieved from Massachusetts Institute of Technology website: http://www.mit.edu/~jhainm/synthpage.html

Hainmueller, J., Abadie, A., & Diamond, A. (2010). *Synth: Synthetic control methods for comparative case studies*. Cambridge, MA: Massachusetts Institute of Technology. Retrieved from http://www.mit.edu/~jhainm/synthpage.html

Hainmueller, J., & Diamond, A. J. (2012). Package: Synth. *inside-R: A community site for R*. Retrieved from http://www.inside-r.org/packages/cran/Synth/docs/synth

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*(2), 219–236. doi:10.1214/08-STS254

Hinrichs, P. (2012). The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities. *Review of Economics and Statistics*, *94*(3), 712–722. doi:10.1162/REST_a_00170

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. doi:10.2307/2289064

Hoxby, C. (2012, January). *Applications of the synthetic control method*. Research panel presented at the American Economic Association, Chicago, IL. Retrieved from http://www.aeaweb.org/aea/2012conference/program/preliminary.php

Hudson, S. (2010). *The effects of performance-based teacher pay on student achievement* (Undergraduate thesis, Stanford University). Retrieved from http://ideas.repec.org/p/sip/dpaper/09-023.html

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(2), 481–502. doi:10.1111/j.1467-985X.2007.00527.x

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86. doi:10.1257/jel.47.1.5

Institute of Education Sciences. (2011). *What Works Clearinghouse procedures and standards handbook: Version 2.1*. Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Jaciw, A., & Newman, D. (2011, March). *External validity in the context of RCTs: Lessons from the causal explanatory tradition*. Paper presented at the conference of the Society for Research on Educational Effectiveness, Washington, DC. Retrieved from https://www.sree.org/conferences/2011/program/downloads/abstracts/191.pdf

Klasik, D. (2013). The ACT of enrollment: The college enrollment effects of state-required college entrance exam testing. *Educational Researcher*, *42*(3), 151–160. doi:10.3102/0013189X12474065

Kroopf, S., Murphey, M. S., Soisson, J. (Producers), & Herek, S. (Director). (1989). *Bill & Ted's excellent adventure* [Motion picture]. United States: Orion Pictures.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Millimet, D. L. (2011). The elephant in the corner: A cautionary tale about measurement error in treatment effects models. *Missing Data Methods: Cross-Sectional Methods and Applications*, *27*(1), 1–39. doi:10.1108/S0731-9053(2011)000027A004

Mueser, P. R., Troske, K. R., & Gorislavsky, A. (2007). Using state administrative data to measure program performance. *Review of Economics and Statistics*, *89*(4), 761–783. doi:10.1162/rest.89.4.761

National Center for Education Statistics. (n.d.). Common Core of Data: Identification of rural locales. Retrieved from Institute of Education Sciences web site: http://nces.ed.gov/ccd/rural_locales.asp

Northwest Evaluation Association. (2011a). *RIT scale norms: For use with Measures of Academic Progress (MAP) and MAP for Primary Grades*. Portland, OR: Author. Retrieved from http://www.nwea.org/support/article/rit-scale-norms-study

Northwest Evaluation Association. (2011b). *Technical manual for Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG)*. Portland, OR: Author.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, United Kingdom: Cambridge University Press.

Pearl, J. (2009). *Myth, confusion, and science in causal analysis* (Technical Report No. R-348). Retrieved from University of California, Los Angeles website: http://www.cs.ucla.edu/~kaoru/r348.pdf

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, *174*(11), 1223–1227. doi:10.1093/aje/kwr352

Quasi-experimental design. (n.d.). In *What Works Clearinghouse glossary of terms*. Retrieved from http://ies.ed.gov/ncee/wwc/Glossary.aspx

Rosenbaum, P. R. (2002). *Observational studies*. New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi:10.2307/2335942

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi:10.1037/h0037350

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58. doi:10.1214/aos/1176344064

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge, United Kingdom: Cambridge University Press.

Sekhon, J. S. (2008). The Neyman-Rubin model of causal inference and estimation via matching methods. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 271–299). New York, NY: Oxford University Press.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, *12*(1), 487–508. doi:10.1146/annurev.polisci.11.060606.135444

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston, MA: Houghton Mifflin.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, *36*(4), 187–198. doi:10.3102/0013189X07303396

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. doi:10.1214/09-STS313

Stuart, E. A., & Rubin, D. B. (2008a). Matching methods for causal inference: Designing observational studies. In J. W. Osborne (Ed.), *Best practices in quantitative methods*. Thousand Oaks, CA: Sage Publications.

Stuart, E. A., & Rubin, D. B. (2008b). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics, 33*(3), 279–306. doi:10.3102/1076998607306078

Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. New York, NY: Springer.

Wooldridge, J. M. (2007, July). *What's new in econometrics?* Lecture presented at the Summer Institute of the National Bureau of Economic Research, Cambridge, MA. Retrieved from http://www.nber.org/minicourse3.html

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics, 86*(1), 91–107. doi:10.1162/003465304323023705

Appendix

University of Arkansas research compliance protocol approval letter

UNIVERSITY OF ARKANSAS

Office of Research Compliance
Institutional Review Board

May 3, 2013

MEMORANDUM

TO:	Clay Johnson Ronna Turner
FROM:	Ro Windwalker IRB Coordinator
RE:	New Protocol Approval
IRB Protocol #:	13-04-689
Protocol Title:	*Compared to What? The Effectiveness of Synthetic Control Methods for Causal Inference in Educational Assessment*
Review Type:	☒ EXEMPT □ EXPEDITED □ FULL IRB
Approved Project Period:	Start Date: 05/03/2013 Expiration Date: 05/02/2014

Your protocol has been approved by the IRB. Protocols are approved for a maximum period of one year. If you wish to continue the project past the approved project period (see above), you must submit a request, using the form *Continuing Review for IRB Approved Projects*, prior to the expiration date. This form is available from the IRB Coordinator or on the Research Compliance website (http://vpred.uark.edu/210.php). As a courtesy, you will be sent a reminder two months in advance of that date. However, failure to receive a reminder does not negate your obligation to make the request in sufficient time for review and approval. Federal regulations prohibit retroactive approval of continuation. Failure to receive approval to continue the project prior to the expiration date will result in Termination of the protocol approval. The IRB Coordinator can give you guidance on submission times.

If you wish to make *any* modifications in the approved protocol, you must seek approval *prior to* implementing those changes. All modifications should be requested in writing (email is acceptable) and must provide sufficient detail to assess the impact of the change.

If you have questions or need any assistance from the IRB, please contact me at 210 Administration Building, 5-2208, or irb@uark.edu.