Abstract Title Page

Not included in page count.

Title: Causal Inference and the Comparative Interrupted Time Series Design: Findings from Within-Study Comparisons

Authors and Affiliations:

Travis St. Clair, University of Maryland
Kelly Hallberg, American Institutes for Research
Thomas D. Cook, Northwestern University and Mathematica Policy Research

Abstract Body

Background / Context:
Researchers are increasingly using comparative interrupted time series (CITS) designs to estimate the effects of programs and policies when randomized controlled trials are not feasible. In a simple interrupted time series design, researchers compare the pre-treatment values of a treatment group time series to post-treatment values in order to assess the impact of a treatment, without any comparison group to account for confounding factors. The CITS design is a version of the ITS design in which both a treatment and a comparison group are evaluated both before and after the onset of a treatment.

A growing body of literature is employing a within study comparison (WSC) methodology to examine the validity of the CITS model. WSC studies empirically estimate the extent to which a given observational study reproduces the results of an RCT when both share the same treatment group, and represent a rigorous method of evaluating non-experimental designs using real data. A number of recent within-study comparisons have demonstrated that CITS can produce estimates that are comparable to those from a randomized controlled trial (RCT) in practice, including St.Clair, Cook, and Hallberg (2014), Schneeweiss, Maclure, Carleton, Glynn, and Avorn (2004), Fretheim, Soumerai, Zhang, Oxman, and Ross-Degnan (2013), and Somers, Zhu, Jacob, and Bloom (2013). However, these findings have been shown to be dependent on correctly modeling the functional form. In the St. Clair et al. (2014) application, the authors found that correspondence with the RCT was possible when the CITS model accounted for baseline trends, but that additional time points could actually increase bias when the pre-treatment trend was not modeled correctly. Examination of the pretreatment trends in this data set showed clearly that in at least one of the outcomes the treatment and comparison groups had different slopes in the pretreatment period, and as a result the "parallel trends" assumption often invoked in the difference-in-difference literature was clearly violated.

Purpose / Objective / Research Question / Focus of Study:
The initial purpose of this paper was to replicate the finding from the previous work drawing on two new datasets. Our hypothesis was that if we examined the treatment and comparison group pretreatment time series and appropriately modeled the trend we would get estimates that closely corresponded to the RCT benchmark. In the first dataset we examine, our hypothesis was confirmed. Examination of the pretreatment trends revealed parallel slopes in the pretreatment data; accordingly, we fit a baseline mean model (discussed further below), and results closely corresponded with the RCT benchmark. The second dataset presented us with more of a conundrum. The pretreatment data were somewhat sporadic and not easily modeled, and neither the baseline mean model nor the baseline slope model resulted in close approximation to RCT. The mixed results from the replication led us to reconsider approaches to dealing with unclear pretest functional form. We considered two approaches. The first is using a more flexible modeling approach, which employs year fixed-effects rather than trying to parametrically model the pretest trend. The second approach attempts to match treatment and comparison cases to reduce reliance on modeling the pretreatment trend. We also considered a third option, non-parametric modeling, but the limited number of pretreatment data points in most CITS studies in social science research, including those in our three examples, preclude employing this approach.

This paper employs a within study comparison (WSC) methodology to examine the performance of these two approaches and compares them to the performance of the baseline mean and baseline slope models across three datasets. We draw on three datasets that include both an RCT and data for non-experimental comparison cases. The first involves retrospectively applying new approaches to analyzing the data presented in St. Clair et al., while the other two are new applications. Fortuitously for our purposes, the three data sets differ in pretreatment trends: the first data set is characterized by parallel pretreatment slopes, the second is characterized by differential linear pretreatment slopes, and the final dataset is characterized by an unclear pretest functional form. We test the performance of the three modeling approaches (baseline mean, baseline slopes, and year fixed-effects) as well as matching. Our purpose is twofold. First, we aim to examine what approach, if any, works in the unclear functional form case. Second, we aim to examine the relative superiority of the different approaches across the three datasets in terms of both bias reduction and precision.

Significance / Novelty of study:

While CITS designs are frequently used in applied education research, little empirical work has been done to examine the efficacy of different modelling approaches and matching in practice.

Statistical, Measurement, or Econometric Model:

The means for achieving the study goals was to conduct two within study comparisons (WSC). Within-study comparisons estimate the extent of bias remaining in non-experimental causal studies after attempts either to select non-experimental comparison groups as similar as possible to the treatment group or after various matching or regression techniques have been applied to adjust for observed group differences. Using these datasets, we estimate the ability to reproduce RCT results and calculate the degree of bias remaining after implementing three modeling approaches: (1) baseline mean; (2) the baseline slopes; (3) year-fixed effects. In addition, we examine the performance of these three methods when supplemented with identifying a comparison group by matching on pretreatment measures of the outcome.

Usefulness / Applicability of Method:

This study provides evidence of the performance of each CITS modeling approach alone and in combination with matching for estimating the effect of school-level interventions drawing on data from three empirical WSCs: two using the Indiana Formative Assessment System RCT (waves 1 and 2) as the causal benchmark and a third using the P-SELL efficacy study for this purpose.

Data Collection and Analysis:

Indiana Benchmark Assessment Study. The first two datasets we examined come from two cluster RCTs (Konstantopoulos, Miller, & Van der Ploeg, 2014) designed to study the effect of Indiana's benchmark assessment system on student achievement in mathematics and English Language Arts (ELA), using the annual Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) as the data source. In the first year of the study (2009-10), fifty-seven K-8 schools volunteered to implement the system. Of these, 35 were randomly assigned to the state's benchmark assessment system while 22 served as controls. In year 2 (2010-11) an additional 63 school were randomly assigned to the two conditions (32 schools to treatment and 31 to control). While the cluster randomized trial gathered data on students in kindergarten through 8th grade, we focus our analysis on students in 4th through 6th grade. The non-experimental comparison group was constructed from all schools that served 4th through 6th graders in the state.

P-SELL Efficacy Study. The second dataset we examined is a cluster RCT designed to study the efficacy of P-SELL. Sixty-four elementary schools in suburban and urban school districts in Florida agreed to participate in the study. All of the study schools serve a large number of students designated as limited English proficient (LEP).Thirty-two of the elementary schools were assigned to implement the P-SELL curriculum in their fifth grade classrooms while the remaining 32 schools agreed to continue with their standard science curriculum. The non-experimental comparison group was constructed from all schools in that state that served 5th grade students.

Findings / Results:

Figure 1 shows the math and ELA (English Language Arts) test scores for the treatment schools from year 1 of the randomized experiment in Indiana and the all-state comparison group. The data consist of test scores for grades 3-8, standardized by grade and by year. The experimental treatment group performed worse than the state average in both math and ELA, and this difference was fairly stable over time in both math and ELA. Because of the stability in difference between the pre-treatment means, Figure 1 leads us to prefer the baseline mean model a priori; there are no obvious differences in slopes or year-specific shocks in the unconditional pretest information.

[INSERT FIGURE 1 HERE]

Table 1 presents the effect estimates for three modelling approaches implemented with and without matching. The effect estimates from the unmatched year fixed effects model are closest to the benchmark RCT estimates; the ELA estimate is within 0.001 standard deviations, and the math estimate is within 0.022 standard deviations, both extremely strong correspondences. These estimates are superior to the estimates from the unmatched baseline mean model, where the bias for ELA and math is -0.061 and -0.050 respectively. With both models, the unmatched estimates show less bias than the matched estimates. While the model incorporating year fixed effects produces less biased estimates than the baseline mean model, the standard errors are considerably larger, an unsurprising result given the eight additional parameters in the model (four year dummies and corresponding interaction terms). So while we prefer the fixed effects model to the baseline mean model in this case from the perspective of bias reduction, there is a clear trade-off in terms of precision. The baseline slope model produces the most biased estimates. This illustrates the danger of overfitting, that is, applying functional form assumptions unwarranted by the data.

[INSERT TABLE 1 HERE]

Figure 2 present the descriptive results from year 2 of the Indiana study, our second within-study comparison. One difference between Figure 2 and Figure 1 is immediately apparent. The ELA results indicate that the pre-test performance of the schools in the treatment group are declining relative to the rest of the state. There is slight evidence of a similar downward trend in the math scores, though the pattern is not nearly as clear.

[INSERT FIGURE 1 HERE]

Table 2 presents the effect estimates, including those that appeared in St.Clair, Cook, & Hallberg (2014) from the baseline slope model. For ELA, the baseline slope model clearly performs better than either the baseline mean or the fixed effects models in terms of bias reduction, confirming our prediction that the use of slope terms was warranted. The single best estimate comes from the matched baseline slope model, 0.042 standard deviations away from the RCT benchmark. For math, the baseline slope model is also preferred to the other two models, though the differences are less significant; the most biased estimate is still only 0.114 standard deviations from the benchmark. With the baseline mean model producing estimates that are more precise than either the baseline slope or the fixed effects models, this approach may be preferable.

[INSERT TABLE 2 HERE]

Figure 3 shows the results for Florida, once again comparing the performance of the experimental treatment schools to all other schools in the state. Unlike the two studies presented above, there is no obvious functional form pattern visible in the data. The treatment schools exhibit a higher degree of volatility, particularly visible in the science scores.

[INSERT FIGURE 3 HERE]

Table 3 shows the results for all three analytic models, both unmatched and matched. The baseline slope model produces an estimate that met our a priori standard only for science; effect estimates for both math and reading were all at least 0.27 standard deviations away from the benchmark, even with the use of matching. This result led us to investigate the fixed effects model. The fixed effects models appear to improve these results unmatched estimate for math is 0.074 standard deviations away from the benchmark. And while the unmatched reading estimate is still not concordant with the RCT results (with a bias of 0.221 standard deviations), the matched estimate is 0.130 standard deviations from the benchmark. In fact, all three matched fixed effects estimates are within 0.20 standard deviations of the benchmark, making it the only analytic model of the three to reduce bias to acceptable standards for each outcome.

[INSERT TABLE 3]

Conclusions:

Table 4 summarizes our findings with respect to bias reduction. No single analytic approach works best for every situation. With year 1 of the Indiana study, where the pre-treatment slopes were roughly parallel, the year fixed effects model produces the least biased estimates. For year 2 of the Indiana study, the baseline trend model produces the best estimates, consistent with the differential trends that were visible in Figure 2 for ELA, and to a less extent for math. For the Florida study, the fixed effects model again performs best. What effect does matching have? Matching reduces bias in 11 out of the 21 estimates. It is most helpful for the Indiana year 2 study, where 4 out of the 6 estimates improve with matching. It is least helpful for year 1 of the Indiana study, where only one of the six estimates improves with matching.

[INSERT TABLE 4]

Appendix A. References

Bloom, H. (2003). Using “short” interrupted time-series analysis to measure the impacts of whole schools reforms: With applications to a study of accelerated schools. *Evaluation Review*, *27,* 3–49.

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

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

Diaz, J.J., & Handa, S. (2006) An assessment of propensity score matching as a nonexperimental impact estimator: Evidence from Mexico’s PROGRESA Program. *Journal of Human Resources*, 41(2), 319-345.

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

Fretheim, A., Soumerai, S.B., Zhang, F., Oxman, A.D., Ross-Degnan, D. (2013) Interrupted time-series analysis yielded an effect estimate concordant with the cluster randomized controlled-trial result. *Journal of Clinical Epidemiology,* 66(8), 883-887.

Friedlander, Daniel, and Robins, P. (1995) Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods. *American Economic Review*, 85(4), September 1995, pp. 923-937.

Glazerman, S., Levy, D.M., & Meyers, D. (2003) Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science*, 589(1), 63-93.

Heckman, J., Ichimura, H., Smith, J.C., & Todd, P. (1998) Characterizing selection bias. *Econometrica*, 66, 1107-1098.

Heckman, J., Ichimura, H., & Todd, P.E. (1997) Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies,* 64, 605-654.

Jacob, R.T., Goddard, R.D., & Kim, E.S. (2013). Assessing the use of aggregate data in the evaluation of school-based interventions: Implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis*. 36(1), 44-66.

Lalonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *American Economic Review*, 76, 604-620.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics*, 86, 156–179.

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

Smith, J.C., & Todd, P. (2005) Does matching overcome LaLonde’s critique of nonexperimental estimators. *Journal of Econometrics,* 125, 305-353.

Somers, M., Zhu, P., Jacob, R., & Bloom, H., (2013) The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation. MDRC working paper in research methodology. New York, NY.

Steiner, P., Cook, T.D., & Shadish, W. (2011) On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics* 36(2), 213-236.

Wing, Coady, & Cook, Thomas D. (2013) Strengthening the Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison. *Journal of Policy Analysis & Management*, 32(4), 853-877.

Appendix B. Tables and Figures

Figure 1a: Treatment Schools vs All Schools in the State: ELA Scores

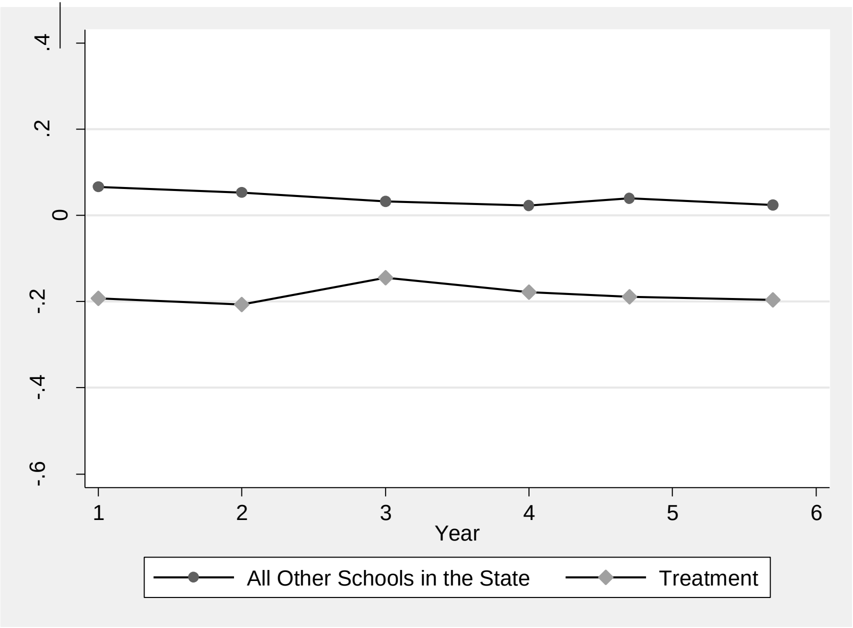

Figure 1b: Treatment Schools vs All Schools in the State: Math Scores

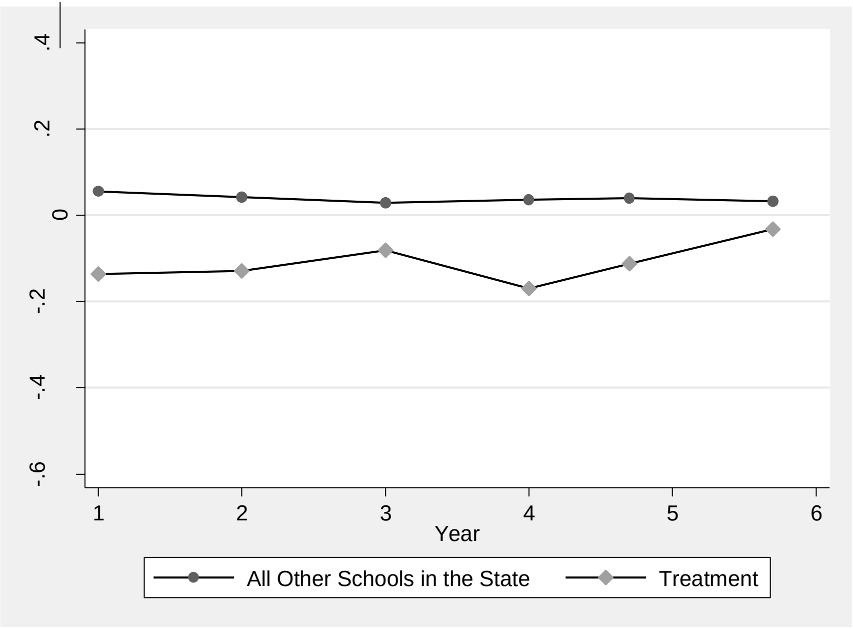

Table 1: WSC#1 - The Difference between Experimental and Quasi-Experimental Results

		RCT	**Without Matching**		**With Matching**	
		Benchmark Estimate	*Estimate*	*Bias*	*Estimate*	*Bias*
Baseline Mean	**ELA Treatment Effect**	0.088 (0.101)	0.027 (0.050)	-0.061	0.013 (0.053)	-0.075
	Math Treatment Effect	0.217 (0.139)	0.167** (0.057)	-0.050	0.126 (0.063)	-0.091
Baseline Slope	**ELA Treatment Effect**	0.088 (0.101)	-0.043 (0.068)	-0.131	-0.008 (0.073)	-0.096
	Math Treatment Effect	0.217 (0.139)	0.079 (0.073)	0.138	0.075 (0.078)	-0.142
Fixed Effects	**ELA Treatment Effect**	0.088 (0.101)	0.087 (0.093)	-0.001	0.023 (0.090)	-0.065
	Math Treatment Effect	0.217 (0.139)	0.239* (0.104)	0.022	0.145 (0.106)	-0.072
	Number of Schools	57	1084		175	

Figure 2a: WSC #2 - Treatment Group vs All School in the State, ELA Scores

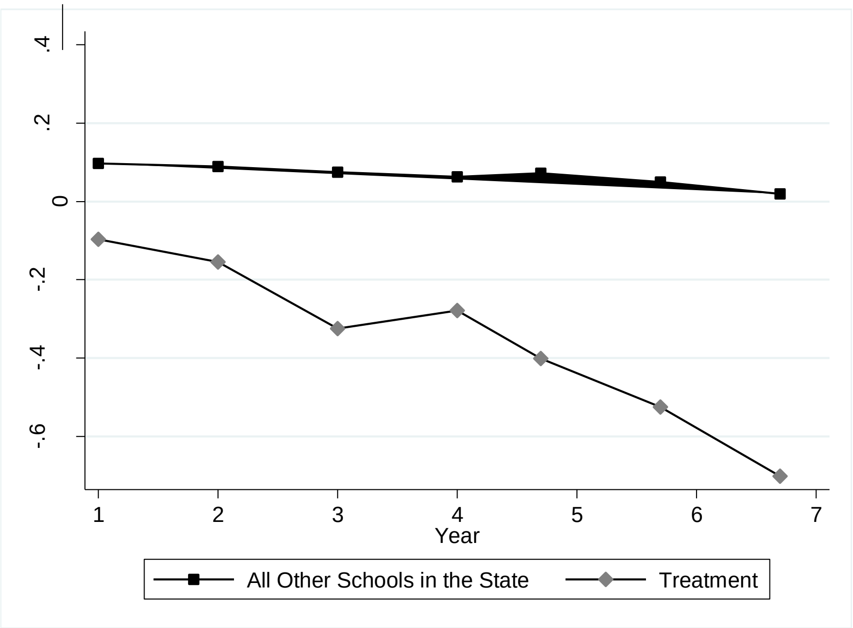

Figure 2b: WSC #2 -Treatment Group vs All Schools in the State, Math Scores

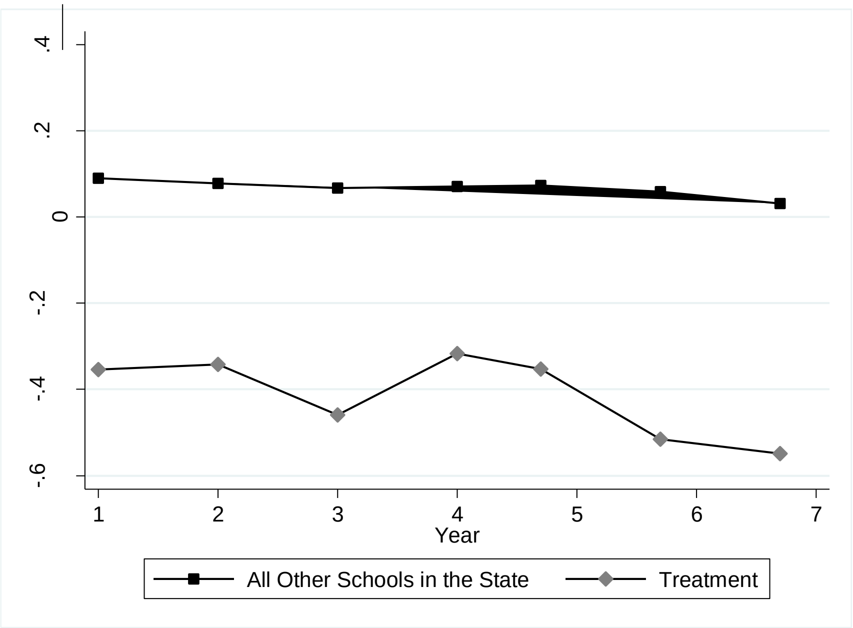

Table 2: WSC #2 - The Difference between Experimental and Quasi-Experimental Results

		RCT	**Without Matching**		**With Matching**	
		Benchmark Estimate	*Estimate*	*Bias*	*Estimate*	*Bias*
Baseline Mean	**ELA Treatment Effect**	-0.099 (0.095)	-0.288** (0.096)	-0.189	-0.179** (0.067)	-0.080
	Math Treatment Effect	-0.014 (0.113)	-0.070 (0.075)	-0.056	0.007 (0.074)	0.021
Baseline Slope	**ELA Treatment Effect**	-0.099 (0.095)	-0.034 (0.074)	0.065	-0.141 (0.087)	-0.042
	Math Treatment Effect	-0.014 (0.113)	-0.020 (0.095)	-0.006	-0.090 (0.108)	-0.076
Fixed Effects	**ELA Treatment Effect**	-0.099 (0.095)	-0.432** (0.096)	-0.333	-0.214* (0.089)	-0.115
	Math Treatment Effect	-0.014 (0.113)	-0.067 (0.164)	-0.053	0.100 (0.153)	0.114
	Number of Schools	63	1008		160	

Figure 3a: WSC #3 - Treatment Schools vs All Schools in the State - Science Scores

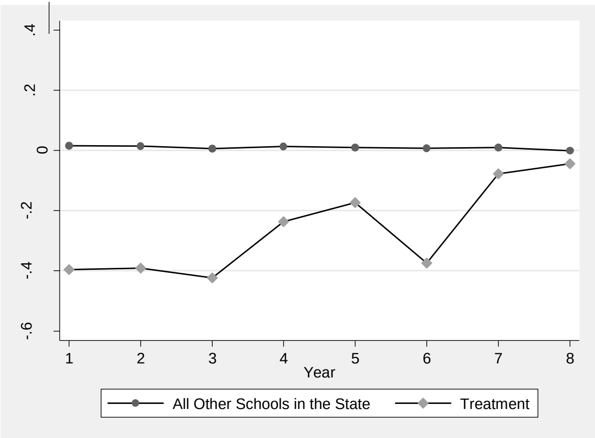

Figure 3b: WSC #3 - Treatment Schools vs. All Schools in the State - Math Scores

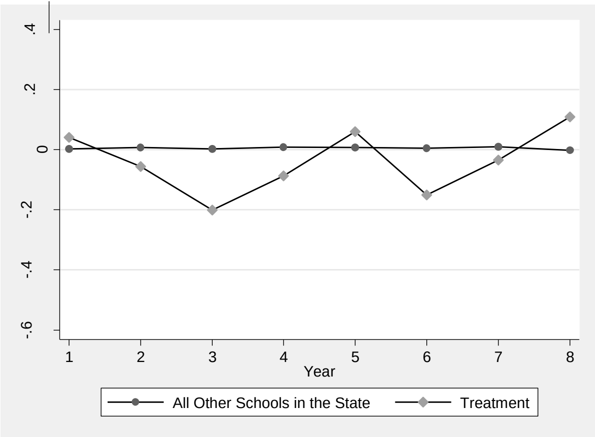

Figure 3c: WSC #3 - Treatment Schools vs. All Other Schools in the State - Reading Scores

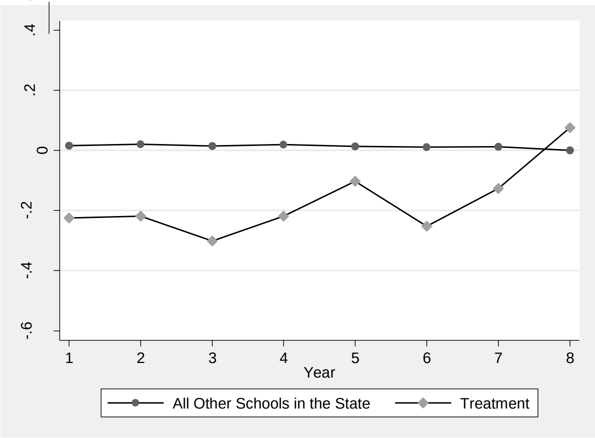

Table 3: WSC #3 - The Difference between Experimental and Quasi-Experimental Results

		RCT	**Without Matching**		**With Matching**	
		Benchmark Estimate	*Estimate*	*Bias*	*Estimate*	*Bias*
Baseline Mean	**Science Treatment Effect**	0.174 (0.135)	0.265** (0.075)	0.091	0.236** (0.084)	0.062
	Math Treatment Effect	-0.107 (0.126)	0.173* (0.083)	0.280	0.118 (0.091)	0.225
	Reading Treatment Effect	0.006 (0.104)	0.335** (0.070)	0.329	0.278** (0.070)	0.272
Baseline Slope	**Science Treatment Effect**	0.174 (0.135)	0.136 (0.090)	-0.038	0.123 (0.100)	-0.051
	Math Treatment Effect	-0.107 (0.126)	0.246* (0.101)	0.353	0.190 (0.107)	0.297
	Reading Treatment Effect	0.006 (0.104)	0.297** (0.084)	0.291	0.282** (0.085)	0.276
Fixed Effects	**Science Treatment Effect**	0.174 (0.135)	0.268** (0.101)	0.094	0.356** (0.108)	0.182
	Math Treatment Effect	-0.107 (0.126)	-0.033 (0.114)	0.074	0.023 (0.119)	0.130
	Reading Treatment Effect	0.006 (0.104)	0.227* (0.092)	0.221	0.136 (0.093)	0.130
	Number of Schools	64	1793		160	

Table 4: Bias Results by Different Modeling Approaches

	Modeling Alone			With Matching		
	Baseline Mean	Baseline Slope	Year Fixed Effects	Baseline Mean	Baseline Slope	Year Fixed Effects
IN 1 (Parallel Trends)						
ELA	-0.061	-0.131	-0.001	-0.075	-0.096	-0.065
Math	-0.050	0.138	0.022	-0.091	-0.142	-0.072
IN 2 (Differential Slopes)						
ELA	-0.189	0.065	-0.333	-0.080	-0.042	-0.115
Math	-0.056	-0.006	-0.053	0.021	-0.076	0.114
FL (Unclear Pretest Functional Form)						
Science	0.091	-0.038	0.094	0.062	-0.051	0.182
Math	0.280	0.353	0.074	0.225	0.297	0.130
Reading	0.329	0.291	0.221	0.272	0.276	0.130