Abstract Title Page
Not included in page count.

Title:

Comparison Groups in Short Interrupted Time-Series: An Illustration evaluating No Child Left Behind

Author(s):

Manyee Wong, PhD, Institute of Policy Research at Northwestern University

Thomas D. Cook, Professor, Institute of Policy Research at Northwestern University

Peter M. Steiner, PhD, Institute for Advanced Studies in Vienna and Institute of Policy Research at Northwestern University

Abstract Body
Limit 5 pages single spaced.

Background/context:
Description of prior research and/or its intellectual context and/or its policy context.

Interrupted time-series (ITS) are often used to assess the causal effect of a planned or even unplanned shock introduced into an on-going process. The pre-intervention slope is supposed to index the causal counterfactual, and deviations from it in mean, slope or variance are used to indicate an effect. However, a secure causal inference is only warranted if: (1) The intervention is exogenous and not a product of prior time series values; (2) the intervention is not correlated with some other force that abruptly affects the outcome at the same time as the intervention; (3) onset of the intervention is abrupt or its dissemination is otherwise well described; (4) the response occurs abruptly or with a theoretically known delay; and (5) correlated error is controlled so that the standard error of any effect is unbiased. It also helps if (6) the effect is large relative to the size of the inter-temporal variation prior to the intervention. Although this is a long series of contingencies, there are nonetheless many examples of interrupted time-series that meet these conditions (Cook & Campbell,1979; Shadish et al., 2002). Some of the examples presented to date require no statistical analysis; their inter-occular impact is striking because the effect is so specific to the intervention time point and so large relative to the prior inter-temporal variation.

Unfortunately, there has been little educational research using interrupted time-series designs. This dearth may be due, in part, to the difficulty of collecting educational time-series that meet the requirement of the standard Box & Jenkins (1970) framework. Their use of autoregressive integrated moving average models (ARIMA) requires many time points to estimate the error structure, with 50 to 100 considered the minimum. Except for studies of daily attendance, it is rare in education to have so many observations on the same or similar children, especially since the observations have to be separated into the pre-intervention ones necessary for estimating the causal counterfactual and the posttest ones that estimate the form of any effect. In most educational research, longitudinal data are collected at many fewer time points, perhaps only three or four before the intervention and even fewer after it. This renders the Box tradition inapplicable to most educational researchers, however important it might be in other spheres of application. Alternative approaches are needed for capturing the separate advantages of multiple pre- and post-intervention time points for when random assignment is not possible but multiple pre-intervention time waves are. Educational researchers should then explore the use of abbreviated interrupted time series (AITS) even though the fewer pre-intervention time points reduces our confidence that the true pre-intervention functional form has been correctly estimated. Confidence is further reduced when one realizes that many educational interventions are implemented slowly rather than abruptly, that many effects are delayed rather than immediate, and that minimally detectable effect sizes of .20 are now deemed desirable whereas Shadish et al. illustrate single-group ITS examples with effects of more than five standard deviations, ITS with a single series does not seem to be practical for educational research. The requirements for clear use seem too stringent, however well they work in engineering and medicine.

Purpose/objective/research question/focus of study:
Description of what the research focused on and why.

One purpose of the proposed paper is to briefly illustrate the case made above. More important is (1) to argue that AITS can help with causal identification even with as few as three pre-intervention time points provided that some form of a non-equivalent comparison series is available; (2) to briefly illustrate the range and quality of non-equivalent time series comparisons; and (3) to illustrate how one kind of comparison time series helps identify the short-term effect of No Child Left Behind (NCLB) on academic achievement. The present proposal tilts towards the last purpose since elaborating the NCLB example allows the other purposes to be explored at the same time.

The 2001 No Child Left Behind Act requires that all students meet proficiency level by 2014. Past studies of NCLB's short-term effects have used simple interrupted time-series analysis based on national data, examining changes in student test scores from before to after the law's implementation (See Figure 1 and Figure 2). The results suggest possible positive effects, particularly in the lower grades. However, the results are far from definitive (Fuller, 2007). Many have argued that the observed rise in student achievement post intervention is just a continuation of prior trends (Hoff and Manzo, 2007), and some states' measures are likely to differ from before to after NCLB, casting doubt on whether the observed change is due to NCLB or changes in test content. The moral is that the pre-intervention time points are not by themselves sufficient for causal identification and estimation.

To reduce some of this uncertainty, researchers need to better understand what would have happened had NCLB not been implemented. In fields other than education it is customary to complement the useful but inadequate pre-intervention series with a comparison series. Sometimes, the comparison series is a non-equivalent independent group, as when West, Hepworth, McCall & Reich (1989) used San Diego, California, and El Paso, Texas, to compare with Phoenix, Arizona, where a drunk-driving ordinance had been introduced. At other times, the comparison is a non-equivalent dependent variable series, as when in his study of the British Breathalyzer Ross (1973) used the hours when pubs were closed as a comparison for the intervention hours when pubs were open. At other times, a switching replication is used, as when television was introduced into some communities at one time and into the original comparison communities six years later (Hennigan et al., 1982).

At first glance, it does not seem possible to add an independent comparison time-series in order to evaluate NCLB since the law applies to all public schools. Where are the comparison schools to come from? Nor does a non-equivalent dependent variable series seem plausible. This requires justifying the identification of some outcomes that the law would not affect but that would be affected by all other historical changes that occurred at about the time of NCLB. But what could these be? The situation looks grim.

However, it is possible to create two groups that vary in their level of treatment, thus in their dosage level of NCLB. NCLB requires that all students be proficient in all basic subjects by 2014, but each state has considerable freedom about how it charts its path to this goal. If it wants to blunt the full weight of the law (i.e., treatment), it can use a relatively easy state test, it can set a low proficiency cutoff score, it can use test content that is particular sensitive to gains by low performing students, it can choose to reach the 2014 goal in steps that begin immediately or are largely postponed until after 2009, or it can choose to do any combination of the above. Starting in 2002, states had a lot of freedom to use NCLB for the immediate improvement of education or to blunt the reform process until at least 2009. We first demonstrate this state variation in treatment dosage between 2002 and 2008 and then use it to create a non-equivalent comparison group that allows for a more rigorous test of NCLB than the pretest time series alone provides. Of course, the test is inevitably conservative since there are no states whose schools are so universally proficient that they are all making adequate yearly progress (AYP). Indeed, states with similar NAEP results report widely different levels of student proficiency according to their own assessments (See our Table 1 and also Fuller et. al., 2006, 2007; Lee, 2006; Skinner, 2005).

Combining simple AITS with a low dosage comparison series allows a summative evaluation of NCLB's short term impact that has greater empirical rigor than prior studies, though the causal question is shifted. Instead of asking, "What is the effect of NCLB compared to the absence of NCLB?" we ask, "How does varying levels of dosage from NCLB affect student test scores?" If there is a direct policy effect, then higher dosage states should see a greater increase in their average percentage of students deemed proficient from before to after NCLB than should states with lower levels of treatment dosage. More specifically, there should be a relative shift in mean and slope at the intervention time point in 2002.

The tests specified above require that the mean and slope of the pre-intervention achievement time series be reliably observed. It does not require that the two series be identical in mean or slope. However, we also match the high and low dosage states on pre-intervention achievement, thus partially reassuring all those researchers who (mistakenly) believe that group comparability is a necessary condition for secure causal inference. But the matching is limited to observables, though we use in this study state-level achievement data where the annual correlations are very high indeed. Even so, unobserved covariates can be a problem to the extent that, after 2002 the high dosage states change their tests and cutoffs differently from the low dosage states, to the extent that the high dosage states introduce more or different educational changes after 2002 that are not part of NCLB, and to the extent that acceptably small difference in causal estimates result when high and low dosage states are compared relative to, say, high and no dosage states or even median and no-dosage states.

Setting:
Specific description of where the research took place.

This research focuses on student achievement in the United States.

Population/Participants/Subjects:
Description of participants in the study: who (or what) how many, key features (or characteristics).

The study's population includes all 50 states and results are based on a representative sample of each state's public school students.

Intervention/Program/Practice:
Specific description of the intervention, including what it was, how it was administered, and its duration.

The policy intervention examined in this study is the 2001 reauthorization of the Elementary and Secondary School Act (EASA). The 2001 No Child Left Behind Act passed into law in January of 2002. NCLB aims to strengthen the assessment and accountability provisions of Title I and to more aggressively hold schools accountable for the academic achievement of disadvantaged students. The law specifies a broad range of requirements. All teachers are to be highly qualified by 2006-2007. This means that teachers must have a bachelors' degree, state certification, and demonstrate expertise in their subject area. For paraprofessionals, they must have completed two years of college or passed a test that demonstrates their ability to support teachers in reading, writing, and math instruction. Schools are required to use scientifically based research teaching strategies in classroom and all students are subjected to a series of tests or assessments aligned with states' curriculum standards. Specifically, it requires that reading and math tests be given to 95% of all students in 4th, 8th, and 12th grade every two years after the law's enactment and annually by 2005-2006 for 3rd thru 8th grade, including at least one high school year. After the 2007-2008 academic year, testing in science will also be required once during grades 3 thru 5, 6 thru 9, and 10 thru 11.

In addition to state assessments, states are required to participate in NAEP, which is to be administered every two years after the law's enactment and annually after 2007. Prior to NCLB, state participation in NAEP was voluntary but after the new law it is compulsory for the receipt of federal funds (Department of Education, 2002). In addition, all test results from each school are to be reported annually to the public and must include all students as a whole and broken down for various subgroups of students (i.e., children with disabilities, limited English proficiency, racial minorities, and children from low-income families). The aim is to provide parents and the community with information on whether a school has been successful in teaching to all children, particularly those most in need.

Perhaps the most important change in NCLB is the requirement that each school makes adequate yearly progress (AYP) so that all students meet "proficiency" in all basic subjects by 2014. Thus, the law not only expanded the 1994 Improving America School Act (IASA) requirements but tied them to concrete expectation of results. A school is said to meet AYP if the percentage of students deemed proficient in a subject area meet or exceed the percentage set by the state. Essentially, NCLB requires states to establish a rising series of competency levels over time where the initial percentage is usually based off of the lowest-achieving student group or school's performance and increases thereafter (Department of Education, 2002). The goal is for schools to make yearly progress toward the preset rising levels so that by 2014, all students are proficient in all subject areas. While NCLB require states to make AYP, it does not specify the amount of progress states must make each year toward full proficiency, nor does it define what is considered proficient. The law only requires that schools make annual "incremental progress". States must decide on their own what they deemed to be proficient and how much progress schools must make each year in order to meet full proficiency by 2014.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

Hierarchical piecewise linear modeling is used to conduct the proposed short interrupted time-series analysis and empirically test the effect NCLB has on states' average student achievement. HLM is most appropriate because it accounts for the nested structure of the data (i.e., time-series data are clustered within states and states are clustered within region), allowing the mean outcome of states and regions to vary from one another. An interrupted time-series design is one of the strongest quasi-experimental approaches to evaluate causal effect of a policy intervention *if* a reasonable comparison group can be specified.

The hierarchical piecewise linear regression model is specified to have two time-series segments with the policy effective date serving as the cutoff. The concept is similar to a piecewise linear growth model that compare growth rates during two separate periods (Raudenbush, 2003). The model also specifies two types of parameters for each segment, the intercept and slope. Also formalized in the model is a dummy for high and low treatment dosage groups and interaction variables that assess dosage group differences in their average test score and growth rate before and after policy intervention. Alternatively, a continuous variable is specified for the supplemental analyses. Any significant change in intercept or slope after the intervention date may be attributed to the policy. However, for an unambiguous and clear policy effects, a *difference* in intercept or slope change between the treatment and comparison group must be observed (See Figure 3). Average state level school characteristics prior to the policy intervention are also included as covariates to control for hidden selection bias. We also test the robustness of the results by using both 2002 (the policy enactment date) and 2001 as the cutoff. Using 2001 as a cutoff allows us to assess any possible pre-policy effects but 2002 as the official year of law passage.

Data Collection and Analysis:

Description of plan for collecting and analyzing data, including description of data.

The analyses for this paper use data from main NAEP assessments administered by the NCES. Main NAEP aims to periodically assess how each state's students are doing in several core subjects. Main NAEP began in year 1990 as part of NCES effort to provide states information on how their students are doing. Assessment instruments are formatted to match the nation's most current instructional practices and so can change over time. Using data from main NAEP, this paper assesses NCLB's effect on student achievement by comparing states that differ in intervention dosage level. Main NAEP data are not as appropriate for time-series analyses as the trend NAEP data (the original survey of NAEP) but trend NAEP data are available only for the nation as a whole so do not allow for group comparisons at the state level. We recognize that the main NAEP data are not as ideal as trend NAEP data given its short time frame and possible changes in test content and structure, but it should still provide valid results as any changes in test content affect all states. We also used NCES Common Core Data to obtain data on student population and state education characteristics. State assessment data from NCES are used to assign states to either the high or low treatment dosage group in core analyses.

Findings/Results:
Description of main findings with specific details.

In general, after NCLB there are no significant difference in average test score change and growth rate change when the higher and lower dosage groups of states are compared for 4th grade reading. However, for 4th grade math there are reliable differences in average test score change between higher and lower dosage groups, and there is a difference in rate of change between 4th and 8th grade math (See Figure 4a-c). For 4th grade reading, the maximum difference in average test score change is 0.04 standard deviation units. For 4th grade math, the difference is 0.14. And for 8th grade math, the difference is 0.09. Using the same criteria, the maximum difference in growth rate change is 0.01 per year for 4th grade reading, is 0.02 per year for 4th grade math, and is 0.022 per year for 8th grade math. From 2003 through 2007, the high and low dosage slope differences translate into cumulative effect sizes of 0.05, 0.10, and 0.11.

The above effect sizes are not large; indeed, they are lower than many existing guidelines about minimal desired effect sizes. However, based on a nonlinear normative benchmark of annual test score change (Hill 2007), the observed 4th grade mean reading effect is roughly equivalent to 1.2 months worth of learning; the 4th grade math intercept shift to 3 months worth of learning; and the 8th grade math effect worth 5 months of learning . For cumulative difference in growth rate change from 2003 to 2007, the effect size is equivalent to 1.5 months worth of learning for 4th grade reading, to 2 months worth of learning for 4th grade math and to 6 months for 8th grade math. Even in this metric, the effect sizes are not large but neither are they negligible.

This same pattern of results holds when high and low dosage states are matched by their prior mean achievement levels, and also when states with middle levels of NCLB dosage are added to the analysis so that all US states are included. However, no analysis with no-treatment states is possible. Only modeling exercises can extrapolate into this unobserved space, but they necessarily entail strong distributional assumptions. So the results we present are necessarily underestimates of the true effect of NCLB.

Conclusions:
Description of conclusions and recommendations of author(s) based on findings and over study. (To support the theme of 2009 conference, authors are asked to describe how their conclusions and recommendations might inform one or more of the above noted decisions—curriculum, teaching and teaching quality, school organization, and education policy.)

The results suggest that, wherever possible, a control or comparison group should complement simple time-series. This will improve on the counterfactual, for fewer alternative interpretations are plausible than when the pre-intervention mean, slope and slope variance provide the only source of counterfactual information. Several kinds of non-equivalent comparison series exist in the quasi-experimental literature, though few have been used in education and we would like to see more, particularly as regards use of the stronger non-equivalent control time series. One of the weaker ones involves contrasting a high versus low dosage series, and we illustrate that here in the hope that it will further educational research on, and using, time series methods even when the pre-intervention time series is far shorter than the Box ARIMA tradition requires.

Appendixes
Not included in page count.

Appendix A. References
References are to be in APA format. (See APA style examples at the end of the document.)

Baker, Peter. (2007). "An Extra 'S' on the Report Card." *The Washington Post* A10.

Billing, Shelley H. (1997). Title I of the Improving America's Schools Act: What It Looks Like in Practice. *Journal of Education For Students Placed At Risk* 2 (4): 329-343.

Billing, Shelley H. (1998). Implementation of Title I of the Improving America's Schools Act: A 1997-1998 Update. *Journal of Education For Students Placed At Risk* 3 (3): 209-222.

Borman, G.D., D'Agostino, J.V. (1996). Title I and Student Achievement: A Meta-analysis of Federal Evaluation Results. *Educational Evaluation and Policy Analysis* 18 (4): 309-326.

Campbell, D, and J Stanley. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago, Illinois: Houghton-Mifflin.

Center on Education Policy. (2006). *From the Capitol to the Classroom: Year 4 of the No Child Left Behind*.

Center on Education Policy. (2007). *Answering the Question that Matters Most: Has Student Achievement Increased Since No Child Left Behind*.

Center on Education Policy. (2008). *Many States Have Taken a "Backloaded" Approach to No Child Left Behind Goal of All Students Scoring "Proficient"*.

Chapman, Laura H. (2007). An Update on No Child Left Behind and National Trends in Education. *Arts Education Policy Review* 109 (1): 25-36.

Department of Education. (2002). No Child Left Behind Act of 2001, Public Law print of PL 107-110. Retrieved August 8, 2008 from http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf.

Department of Education. (1994). Improving America's Schools Act of 1994. Archived Information. Retrieved August 6, 2008 from http://www.ed.gov/legislation/ESEA/toc.html.

Dillon, Sam. (2007). New Study Finds Gains Since No Child Left Behind. *New York Times*.

FairTest. (2005). Flatline NAEP Scores Show Failure of Test-driven School Reform: "NO CHILD LEFT BEHIND" has not improved academic performance.

Fuller, Bruce, Kathryn Gesicki, Kang Erin, and Joseph Wright. (2006). *Is the No Child Left Behind Act Working?: The Reliability of How States Track Achievement*. Policy Analysis for California Education.

Fuller, Bruce, Joseph Wright, Kathryn Gesicki, and Erin Kang. (2007). Gauging Growth: How to Judge No Child Left Behind. *Educational Researcher* 36 (5): 268-278.

Goertz, Margaret E. (2005). Implementing the No Child Left Behind Act: Challenges for the States. *Peabody Journal of Education* 80 (2): 73-89.

Goertz, Margaret E., and Mark C. Duffy. (2007). *Assessment and Accountability Across the 50 Statess. CPRE Policy Briefs*. Consortium For Policy Research in Education.

Goertz, Margaret, and Mark Duffy. (2003). Mapping the Landscape of High-Stakes Testing and Accountability Programs. *Theory Into Practice* 42 (1): 4-11.

Hennigan, K.M., Del Rosario, M.L., Heath, L., Cook, T.D., Wharton, J.D., & Calder, B. J. (1982). Impact of the introduction of television on crime in the United States. Journal of Personality and Social Psychology, 55, 239-247.

Hernandez, Raymond. (2004). Bush Carries His Attack Against Kerry to Pennsylvania. *New York Times* 23.

Hoff, David J., and Kathleen Kennedy Manzo. (2007). Bush Claims About NCLB Questioned. *Education Week* 26 (27): 1-27.

Lee, Jaekyung. (2002). Racial and Ethnic Achievement Gap Trends: Reversing the Progress Toward Equity? *Educational Researcher* 31 (1): 3-12.

Lee, Jaekyung. (2006). *Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-Depth Look into National and State Reading and Math Outcome Trends*. Cambridge, MA: Harvard Civil Rights Project.

McDermott, Kathryn A., and Laura Jensen. (2005). Dubious Sovereignty: Federal Conditions of Aid and the No Child Left Behind Act. *Peabody Journal of Education* 80 (2): 39-56.

McDonnell, Lorraine M. (2005). No Child Left Behind and the Federal Role in Education: Evolution or Revolution? *Peabody Journal of Education* 80 (2): 19-38.

Milbank, Dana. (2002). With Fanfare, Bush Signs Education Bill. *The Washington Post*.

National Council of Teachers Mathematics. (2008). Rise in NAEP Math Scores Coincides with NCTM Standards. *NCTM News Bulletin* 2008 (January/February): 1-2.

Raudenbush, Stephen W., and Anthony S. Bryk. (2002). *Hierarchial Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, California: Sage Publications, Inc.

Ross, H.L. (1987), Law, science and accidents: The British Road Safety Act of 1967. Journal of Legal Studies, 2, 1-75.

Rudalevige, Andrew. (2003). No child Left Behind: Forging a Congressional Compromise. In P.E. Peterson and M.R. West (Eds), *No Child Left Behind?* (pp. 23 -54).

Schwartz, Robert B., and Marian A. Robinson. (2000). Goals 2000 and the Standards Movement. *Brookings Paper on Education Policy* 2000 (1): 173-206.

Skinner, Ronald A. (2005). State of the States. *Education Week* 77-80.

Smith, Emma. (2005). Raising Standards in American Schools: The Case of No Child Left Behind. *Journal of Educational Policy* 20 (4): 507-524.

Stullich, Stephanie, Elizabeth Eisner, and Joseph McCrary. (2007). *National Assessment of Title I: Final Report*. U.S. Department of Education.

Superfine, Benjamin Michael. (2005). The Politics of Accountability: The Rise and Fall of Goals 2000. *American Journal of Education* 112 (1): 10-43.

The White House. (2008). The State of the Union Address. Retrieved August 6, 2008 from http://www.whitehouse.gov/stateoftheunion/2008/

U.S. Department of Education. (2005). *The Achiever* 4 (12): 2.

Wanker, William Paul, and Kathy Christie. (2005). State Implementation of the No Child Left Behind Act. *Peabody Journal of Education* 80 (2): 57-72.

West, S.g., Hepworth, J.T., McCall, M.A., Reich, J.W. (1989). An Evaluation of Arizona's July 1982 drunk driving law: Effects on the city of Phoenix. Journal of Applied social Psychology, 19, 1212-1237.

Appendix B. Tables and Figures

Not included in page count.

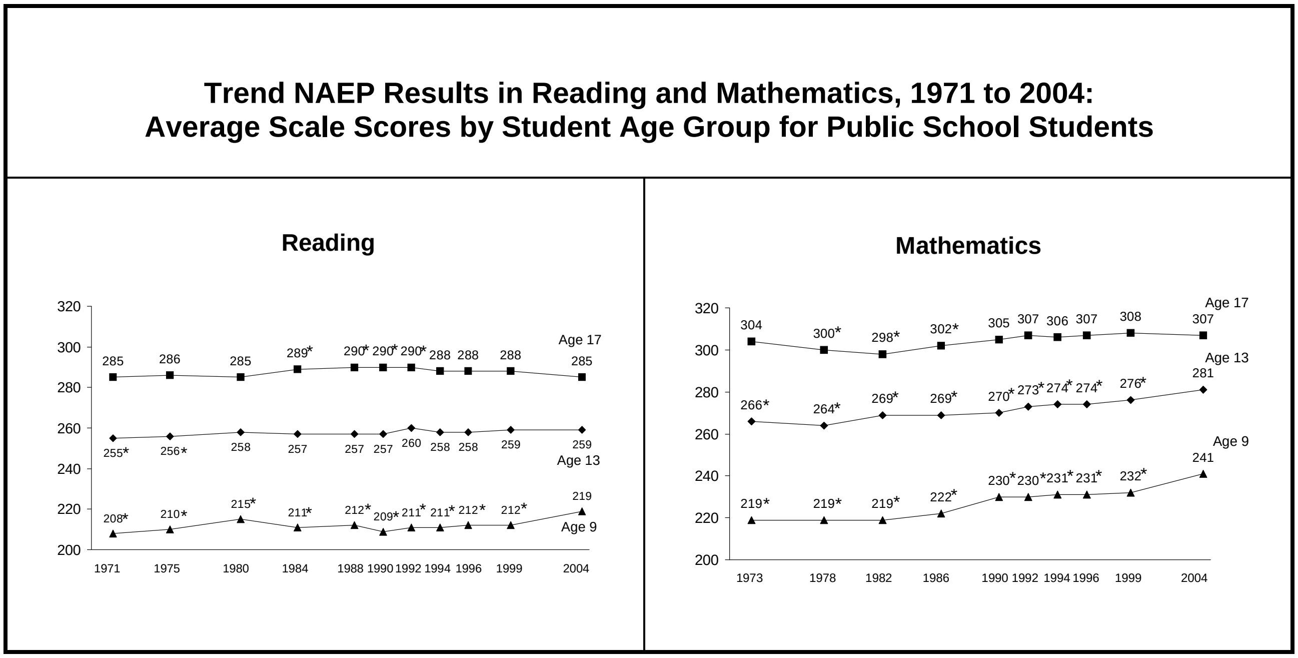
Table 1. Average Percentage of Students At or Above Proficiency Level By Test Standards and Region: A Comparison of State and NAEP Results Averaged Across Grades, Subjects, and Years

	High Test Standards States			Low Test Standards States		
Region	State	Avg. State Score	Avg.NAEP Score	State	Avg. State Score	Avg. NAEP Score
Northeast	Maine	37	35	New Hampshire*	76	40
	Rhode Island	49	28	Connecticut	75	38
West	California	38	23	Idaho	77	32
	Nevada*	49	22	Colorado	84	35
	Hawaii*	33	21	Utah	75	32
	Wyoming	41	35			
South	South Carolina	29	28	North Corolina	87	33
	Arkansas	49	27	Georgia	77	26
				Tennessee*	83	25
				Texas	80	30
				Viriginia*	78	26
Midwest	Missouri*	30	31	Nebraska	84	34
				South Dakota	76	36
				Wisconsin	76	36

Note: High standards group determination based on having less than 50% of students below proficiency level.
Low standards group determination based on having at least 75% of students above proficiency level.

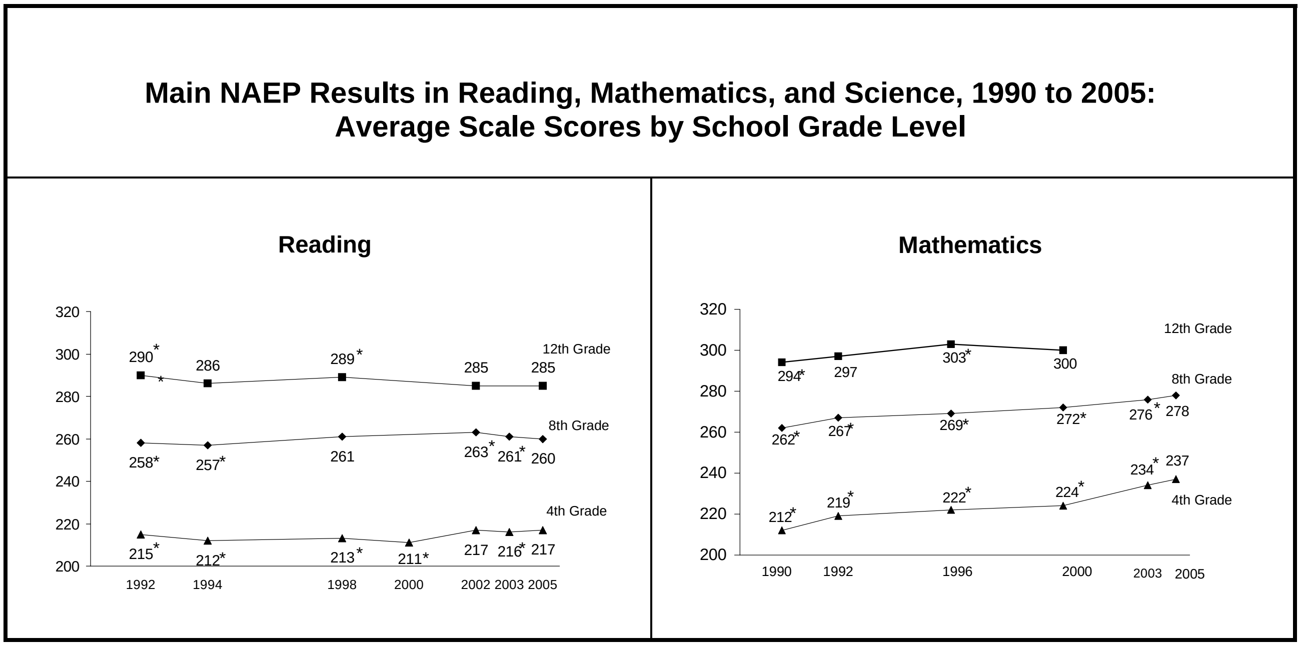
* States where assessment data are not available, the next lower nearby grade data are used. If next lower grade data are not available, the next higher grade data are used.

Figure 1. Trend NAEP Reading and Math Time Series Data


Trend NAEP Results in Reading and Mathematics, 1971 to 2004: Average Scale Scores by Student Age Group for Public School Students

* Indicates that the score is significantly different from the most recent score (p<.05).
Source: National Center for Education Statistics, Trend NAEP.

Figure 2. Main NAEP Reading and Math Time Series Data.


Main NAEP Results in Reading, Mathematics, and Science, 1990 to 2005: Average Scale Scores by School Grade Level

* Indicates that the score is significantly different from the most recent score (p<.05).

Source: National Center for Education Statistics, Trend NAEP.

Figure 3: Hypothetical Policy Intervention Effects on Treatment and Control Group

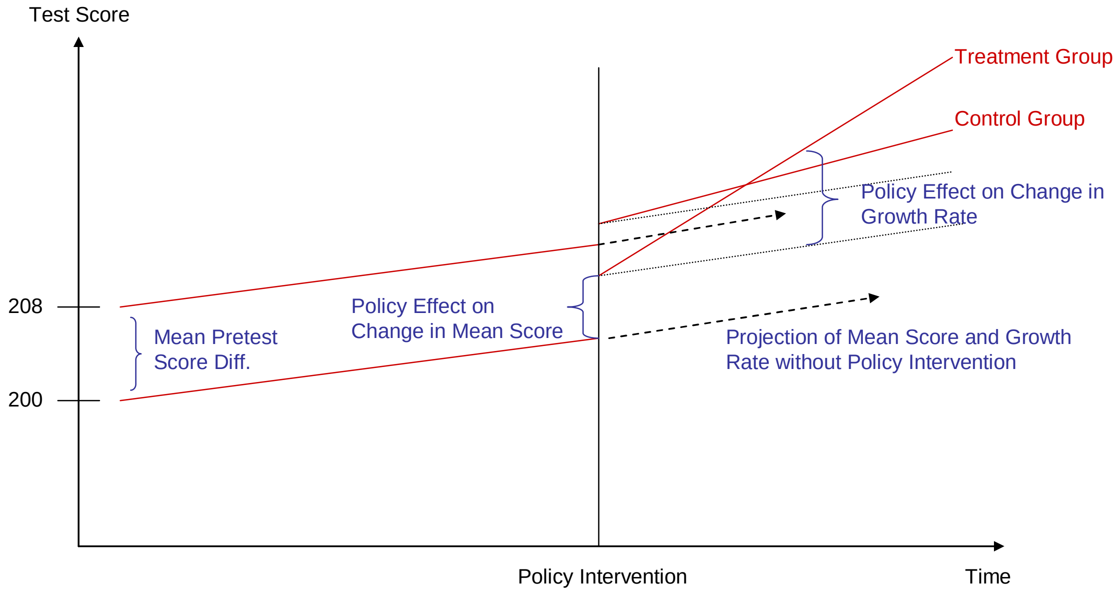


Figure 4a-c. Predicted NAEP Test Scores Over Time Before and After Policy Intervention for 4th Grade Reading, 4th Grade Math, and 8th Grade Math

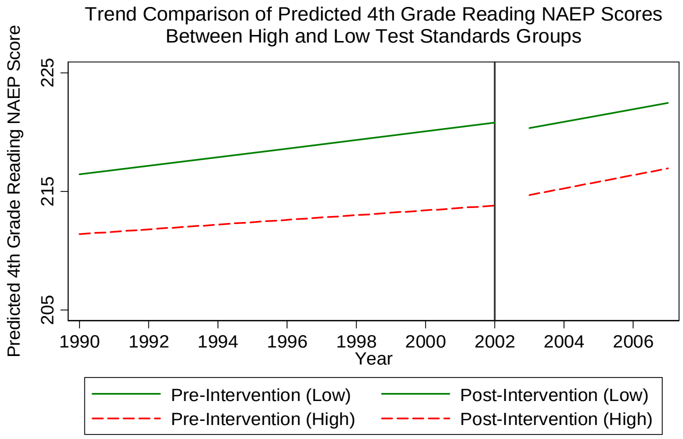


Figure 4a. Predicted NAEP 4th Grade Reading Test Scores Over Time Before and After Policy Intervention for High and Low Test Standards Group

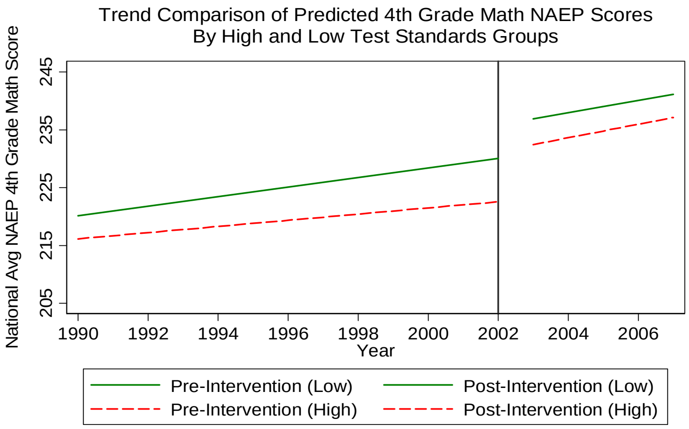


Figure 4b. Predicted NAEP 4th Grade Math Test Scores Over Time Before and After Policy Intervention for High and Low Test Standards Group

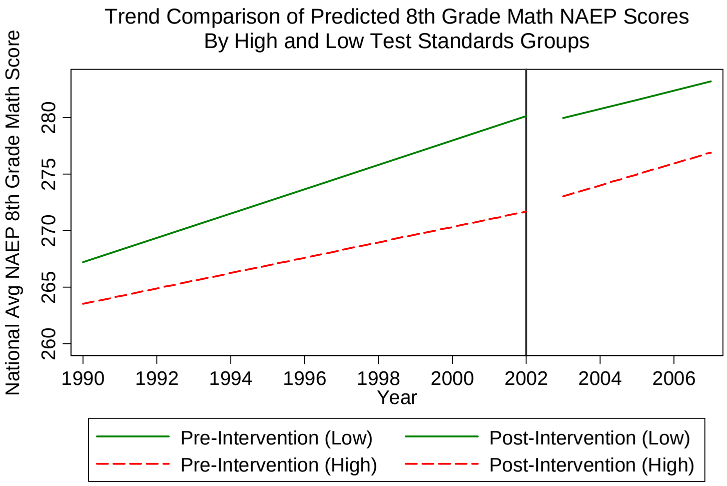


Figure 4c. Predicted NAEP 8th Grade Math Test Scores Over Time Before and After Policy Intervention for High and Low Test Standards Group