GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

DOCUMENT REVIEWED:	**"Answering the Question That Matters Most: Has Student Achievement Increased Since No Child Left Behind?"**
AUTHOR:	**Center on Education Policy**
PUBLISHER/THINK TANK(S):	**Center on Education Policy**
DOCUMENT RELEASE DATE:	**June 5, 2007**
REVIEW DATE:	**June 26. 2007**
REVIEWER:	**John T. Yun**
E-MAIL ADDRESS:	**jyun@education.ucsb.edu**
PHONE NUMBER:	**(805) 893-2342**

Summary of Review

A new report released by the Center on Education Policy, "Answering the Question That Matters Most: Has Student Achievement Increased Since No Child Left Behind?" has received a great deal of attention in the press and is likely to be cited often in the upcoming debate on the reauthorization of the No Child Left Behind Act (NCLB). Using states as their unit of analysis, this report concludes that since the implementation of NCLB in 2002, on average, test scores have increased, the achievement gap has narrowed, and achievement gains post-NCLB have increased faster than before NCLB. Despite its attempt and intent to carefully analyze the complex issue of test score improvement before and after the implementation of NCLB in 2002, however, there are some important weaknesses in the analysis that may have resulted in a much more optimistic picture of the impact of the legislation than the data warrant. The report acknowledges several important methodological weaknesses, but other such weaknesses are never mentioned. Among these additional problems are issues of scope, measurement, and selection—all of which ultimately call into question the robustness of the findings, rendering the report's conclusions far from definitive.

I. INTRODUCTION

The report released by the Center on Education Policy, "Answering the Question That Matters Most: Has Student Achievement Increased Since No Child Left Behind?"[1] has received a great deal of attention in the press and is likely to be cited often in the upcoming debate on the reauthorization of the No Child Left Behind Act (NCLB). The report attempts to carefully analyze the complex issue of test score improvement before and after the implementation of NCLB in 2002. There are some important weaknesses in the analysis, however, that may have resulted in a much more optimistic picture of the impact of the legislation than the data warrant.

In addition to these methodological issues, which bring into question the robustness of some of their results (and will be discussed more completely in the methodology section below), a secondary concern exists about the overstating of results. The title of the report sets the stage for an examination of the direct impact of NCLB on student achievement as measured by state level test scores. The report does not do this, however, and was never designed to do this. As the authors themselves explain concerning one of their main conclusions,

> It is very difficult if not impossible, to determine the extent to which these trends and test results have occurred *because* of NCLB. Since 2002, states, school districts and schools have simultaneously implemented many different but interconnected policies to raise achievement (p. 1).[2]

Given this appropriate disclaimer (and many similar disclaimers throughout the report) it is unclear why the Center chose to release the report under a title that blurs the line between causal and descriptive analyses. While it may be unfair to criticize a report for how others interpret it, the report's title is not the only source of confusion. The wording of the numerous findings and key conclusions in the report imply a much stronger connection between the implementation of NCLB in 2002 than can be substantiated by the data, resulting in a possible over-estimate of the impact of the legislation on the actual changes in test scores.

Notwithstanding these concerns, this new report offers some thoughtful analyses that take into account important issues that have not been often addressed in the past. For instance, the authors use multiple measures of growth (change in percent proficient and effect sizes); they acknowledge the problem of changes in the testing system across years and states; they describe the weakness of using cut scores to determine proficiency levels; and they identify and correct errors in the data-reporting systems state by state. Yet these positive steps are undermined by several issues not adequately addressed in the report that weaken the overall approach and make the findings far less compelling. The methodological problems with this report include issues of scope, measurement, and selection bias,[3] all of which ultimately call into question the robustness of the findings, rendering the report's conclusions far from definitive.

In general, while this report represents progress toward a more comprehensive way to examine outcomes from a complex and wide-ranging federal policy like NCLB, the main emphasis of the report probably should not have been the tenuous connection between the implementation date of NCLB and subsequent achievement changes. In fact, a second, much stronger, point made in the report is largely hidden behind the causal

headline: the current lack of coordination and support for comprehensive, state-by-state data reporting and analysis. Such improved data sets would likely assist states, districts and schools in their instructional decision-making—which is arguably the ultimate goal of all good data analysis.

II. FINDINGS AND CONCLUSIONS OF THE REPORT

The scope of the report is extremely wide-ranging and ambitious. Within this relatively short report (approximately 100 pages) the authors discuss the issue of data collection and analysis across all 50 states and then examine achievement at multiple grade levels using multiple tests across every state over multiple years. In addition, they use two different measures of achievement and growth (change in percent proficient and change in effect sizes[4]) across all available states, years, and grade levels, and they then try to expand their analysis to determine how achievement gaps have changed among various ethnic/racial groups.

Such a complex analysis is difficult under the best of circumstances, as suggested by this report's many conflicting findings. The key summary conclusions (those that are most likely to be used in ongoing policy debates) cannot capture the reasons for these many conflicts, they can only provide a summary picture over multiple, incomparable and occasionally conflicting measures.

In their report, the CEP authors begin by presenting their five major summary conclusions. They then present numerous findings that they use to substantiate their conclusions. Because of its broad scope and the large numbers of findings described in the report, I will simply present the main conclusions in the study, plus one key finding described later in this review. In addition, I will detail in the methodology section below several critical problems with the report's underlying analysis that substantially weaken the support for several of its conclusions.

One of the main problems is evident in the report's five main conclusions quoted below. For each main conclusion, very different numbers of states were included in the underlying analysis. This is due to the fact that the authors chose to include or exclude states from analyses based on the amount and type of available data. While from one perspective this choice improves the comparability of test data from year to year, it also introduces the possibility of selection bias and artificial test score gains. These dangers will be explored later in this review.

The report's five main conclusions are as follows (quoting from page 1):

1. In most states with three or more years of comparable test data, student achievement in reading and math has gone up since 2002, the year NCLB was enacted.
2. There is more evidence of achievement gaps between groups of students narrowing since 2002 than of gaps widening. Sill the magnitude of the gaps is often substantial.
3. In 9 of the 13 states with sufficient data to determine pre- and post-NCLB trends, average yearly gains in test scores were greater after NCLB took effect than before.
4. It is very difficult if not impossible, to determine the extent to which these trends and test results have occurred *because* of NCLB. Since 2002, states, school districts and schools have simultaneously implemented many different but interconnected policies to raise achievement

5. Although NCLB emphasizes public reporting of state test data, the data necessary to reach definitive conclusions about achievement were sometimes hard to find or unavailable, or had holes or discrepancies. More attention should be given to issues of the quality and transparency of state test data.

Together, if supported by the data and analyses, the first three of these conclusions would seem to provide substantial support for the position that NCLB has indeed contributed a great deal to improving average achievement on state examinations. In addition, the final two conclusions suggest that a degree of caution and care should be taken in the pursuit of the answers to the research questions, and they suggest an important focus on the future availability and use of data in state testing systems.

It is also important to note, however, a key finding reported by the CEP authors but not listed as a main conclusion of the report. The authors state that there is almost no correlation between the findings of the study, which are based on various state assessments, and scores reported by the National Assessment of Educational Progress (NAEP). This lack of consistency between multiple measures is troubling since one would expect true achievement gains to not be test-specific. True gains should be reflected on multiple examinations of similar content. The authors suggest that this lack of consistency may be due to a lack of alignment of NAEP with the state standards, a lack of motivation among students on the NAEP versus the state examinations, different inclusion criteria, score inflation, or differences in the grades analyzed (pp. 71-72). These are all legitimate possibilities that should be considered and explored. Some of these explanations are more damaging for the CEP report than others, however. In fact, as discussed below, the possibility of score inflation on state examinations may be the most likely explanation for this inconsistency, which throws into question the robustness of the report's overall conclusions.

III. THE REPORT'S USE OF RESEARCH LITERATURE

The format of the report is not designed around an examination of the research literature; instead it is focused on an analysis of the new datasets the authors created, and the literature they cite is largely in support of the methodological choices or rival explanations for the results they find. For instance, the report invokes previous research[5] to support explanations as to why the NAEP test scores do not match the analyses performed in the report and the possibility that test score inflation is a factor in their state test results. There is no examination, however, of past research on the exact topic of how test scores have changed over time on state examinations.[6]

A 2002 paper by Linn found that on the NAEP in the 1990s (well before NCLB was implemented) average yearly changes in percent proficient were modest. For instance, between 1992 and 1998, only three of the 33 states administering the assessment showed one percent or more yearly gains in scores in 4th grade reading. Similarly, only 17 of 34 states showed such gains in mathematics from 1992 to 2000. In addition, Linn found that very few states showed any decreases (seven of 33 states in reading and one of 34 states in mathematics between the same years, respectively).[7] These earlier findings, had they been included and discussed in the report, could have helped place several of the report's findings in context, since the numbers of states showing increases and declines on their own state ex-

aminations were very similar to these NAEP findings prior to NCLB—particularly the relatively small numbers of states showing declines during this period.

In addition, there exists a very rich literature using the NAEP to examine trends concerning achievement gaps between racial and ethnic groups.[8] Yet no literature of this type was cited or used. In fact, NAEP data on the widening or narrowing of the achievement gap was not used at all even as a comparison for the gap findings in this report. Such context could have helped readers understand how large or small the changes reported in this report are compared to other authors' estimates both before and after NCLB. Thus, readers would be able to better determine whether these changes are part of a consensus overall trend toward narrowing gaps, or a change from a static or increasing situation. Given the care used by the authors in designing their trend studies, this omission is somewhat puzzling.

IV. REVIEW OF THE REPORT'S METHODOLOGIES

The report makes good use of the data that the authors cleaned and collected within the states. The approach used in this report illustrates, in a very clear way, the problems inherent in creating strong cross- and within-state comparisons, particularly given the lack of transparency of many state data systems and the rapid changes occurring in state testing systems. Some critical problems remain, however. Below, I discuss three of the most serious, concerning the report's trend, gap, and pre/post analyses.

Trend Analysis

Trend analysis was used to support conclusion #1: "In most states with three or more years of comparable test data, student achievement in reading and math has gone up since 2002, the year NCLB was enacted." But this analysis suffers from fundamental problems with measurement, selection and robustness brought about by analytic design decisions.

Trends were defined by the average change in test scores across a minimum of two to three years. Moderate-to-large changes were defined as changes larger than one percentage point per year, and slight changes as less than one percentage point per year (p. 2). In effect, this means that the actual average yearly increase relies on the selection of the first year of data and the last year; the years in-between do not play any role in the overall change. If either the first or last year is lower or higher than the 'true' score (a hypothetical perfect measurement), either through random process like measurement error or non-random process like score inflation, the estimates of growth will be biased. As explained below, this is in fact very likely, given how the states were selected into the various samples.

Data were included in the "trend" analysis if they met particular criteria. Among these criteria is the condition that the state data set must have no "breaks" in the data. Breaks are defined in the report as changes in the state testing system that have the effect of making consecutive years of data non-comparable (pp. 78-79). Examples of such changes could be a new test, a new set of content standards, or changes in the cut scores that set proficiency levels. In any states with these breaks, only data that were comparable—that is, data subsequent to the breaks—were analyzed. This inclusion criterion was cited in the report as a way to ensure comparability. But it resulted in the exclusion of some or all data from 39 states.[9]

This approach, however, simply trades one problem for another. While the authors were able to compare similar tests across years, their approach had the likely unintended effect of biasing the sample toward states more likely to have inflated test scores. A 2002 analysis by Koretz provides strong evidence that new tests lead to a depression of initial scores, followed by rapid increases in subsequent years.[10] Accordingly, there is a strong possibility the introduction of a new examination would be followed by relatively large increases in change scores.

For this reason, the sampling approach that includes states in the analysis once tests have been introduced may ensure strong growth measures over a constricted sample of years. For example, consider two hypothetical states, A and B. State A has a consistent state assessment in place from 1998 to 2005. In this case the year 2002 will be used as the start point for the post-NCLB analysis, and 2005 will be used as the end date to calculate growth per year. State B changed its test in 2003 and also has data up through 2005; thus the start date for State B will be 2003 and the end date will be 2005. Since State B's test is new, there is a strong possibility of test score depression in the first year and rapid growth in the next few years will artificially bias the growth estimates upwards.[11]. Since State A's system is long established, the test score changes in this system are not as likely to be subjected to this type of bias. Since there are quite a few states that introduced new testing systems post 2002, this is likely to have a real impact on the Report's estimates of growth per year.

Had the NAEP data shown independent confirmation of the changes observed in the state tests, there would be more evidence to suggest that the changes were real and not a function of test-score inflation. But given the weak to nonexistent correlation between the NAEP and state test results, these questions about the validity of the reported improvements on the state examinations should have been more fully addressed.

In addition, state graphs of percent proficient provided on the CEP website show a great deal of variability,[12] which suggests that if different endpoints were selected, the results of the analysis could radically change—particularly given the small-to-average changes (1 percent per year) that define slight or moderate increases. This is important for two reasons. First, for states with breaks in the data, the initial score point is defined as the first year of the new testing regime, not necessarily 2002, the NCLB implementation year. Second, for those states without breaks, the first year (2002), may not be the year the state actually responded to the implementation of NCLB. According to *Education Week*, by the 2004-2005 school year only 23 of the 50 states and the District of Columbia were testing students in math and reading in grades 3 through 8, and even fewer were additionally implementing other aspects of NCLB.[13] Thus, the problem of random variability is compounded by the fact that the state's actual implementation date for NCLB is unknown, resulting in an inability to draw reliable conclusions relating the implementation of NCLB to state test score changes.

Gap Analysis

Gap analysis was used to support conclusion #2: "There is more evidence of the narrowing of achievement gaps between groups of students since 2002 than of gaps widening. Still, the magnitude of the gaps is often substantial."

All of the methodological concerns discussed in the above trends section also apply

to the gap analysis. Further, the gap analysis is complicated by the small sample size among some of the included racial and ethnic groups. Due to such small sample sizes, these ethnic and racial groups are more likely to show greater variability in their estimates of percent proficient. This variability may well render the gap measures even less robust than the trend measures because unreliability is built into each of the percent-proficient estimates.

Pre- and Post-NCLB Analysis

The pre- and post-NCLB analysis is used to support conclusion #3: "In nine of the 13 states with sufficient data to determine pre- and post-NCLB trends, average yearly gains in test scores were greater after NCLB took effect than before."

The value of this analysis is seriously damaged by issues of selection. Only 13 states had sufficient data pre- and post-NCLB to allow for this analysis—a very small sample size. Moreover, these 13 were not randomly selected. The states included are those that had not changed any key elements of their testing policies from 1999 to 2006, the span of this study.

These states all had accountability systems that they had designed well before NCLB. Some of those systems met the NCLB guidelines enough not to be changed. With the others, the states either a) decided that the current system served them well under the NCLB regime, or b) decided that they would not alter the current system even though it may result in considerable NCLB-related sanctions.

Either way, it is exactly the wrong group to be examining in order to determine whether or not NCLB had an impact on student achievement, since any comprehensive investigation of the impacts of a policy intervention should most certainly consider whether the intervention induces changes in the behavior of the state. This report, in contrast, excludes very important information about such state policy changes. It includes those states that instituted a testing program and set of standards well prior to NCLB, so any change before and after the NCLB implementation date cannot reliably be attributed to NCLB's testing or standards requirements.

Robustness

The report should be praised for several cautionary notes, helping readers to understand some of its limitations. The three above-described and serious limitations were not among those discussed in the report, however. Moreover, the report's authors could have addressed some of these concerns by, for instance, using slightly different criteria for the start- and end-dates to see if the results changed due to random variation. In addition, the authors might have attempted to equate tests across breaks, as an alternative approach for maintaining comparability.

Without such explorations and without a rationale for dismissing or minimizing these problems, the data cannot be seen to convincingly support the report's conclusions.

V. REVIEW OF THE VALIDITY OF THE FINDINGS AND CONCLUSIONS

The report is on its most solid ground with regard to its more cautionary conclusions: conclusion #4 (concerning the causal warrant to suggest that NCLB is improving test scores) and conclusion #5 (concerning the difficulty in obtaining data). In particular, the difficulty that the CEP authors describe in obtaining and analyzing their data is well

documented and supported, and it points to an important need for the future.

The report's analytical findings rest on very weak foundations, however. The approaches employed to fix data problems may well have created other problems that were not addressed in the report and that ultimately undercut their substantive conclusions.

VI. USEFULNESS OF THE REPORT FOR GUIDANCE OF POLICY AND PRACTICE

The description of the difficulty the Center on Educational Policy researchers encountered in simply trying to gather what should be publicly available data is instructive and should be noted by policymakers trying to understand what types of supports would help states in their implementation of NCLB.

The substantive data analysis provided on the potential impact of NCLB on the achievement of students is suspect, however. While the conclusions of this study are sure to be cited in the debate around reauthorization, the data and analyses should be viewed with great caution, and should not overshadow the more important and concrete findings regarding data difficulties.

NOTES & REFERENCES

[1] Center on Education Policy (2006, June 5). *Answering the Question That Matters Most: Has Student Achievement Increased Since No Child Left Behind?* Washington. D.C.: Author. (No individual author is listed for the report.)

[2] Center on Education Policy (2006, June 5). *Answering the Question That Matters Most: Has Student Achievement Increased Since No Child Left Behind?* Washington. D.C.: Author.

[3] Selection bias is defined as bias in estimates due to how samples are selected and are unrelated to the actual underlying phenomenon that is being estimated. For instance, a survey of parents conducted only at university daycare centers would be biased toward parents with more formal education and could not justifiably be generalized to the entire population of parents.

[4] NCLB requires states to move all students to full proficiency on their state language arts and mathematics examinations by 2014 with the definition of proficiency to be determined state by state. Effect sizes are simply measures of how large a change in scores is, relative to a given state's distribution of the scores. A full definition of these terms and a description of the authors' approaches is provided in the CEP report on pages 21-26.

[5] For examples of such literature cited see:

Hamilton, L. (2003). Assessment as a policy tool. In Flooden, R. E. (ed), *Review of Research in Education*, 27, 25-68.

Linn, R. L., Graue, M. E., and Sanders, N. M. (1990). Comparing states and district test results to national norms. *Journal of Educational measurement: Issues and Practice*, 9, 5-14.

Koretz, D. (2005). *Alignment, high stakes, and the inflation of test scores.* Los Angeles: National Center for Research and Evaluation, Standards, and Student Testing.

[6] For a description of one such study see Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). "Is the no child left behind act working? The reliability of how states track achievement." Policy Analysis for California Education, Working paper 06-1.

[7] Linn, R. L., Baker, E. L., Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001, *Educational Researcher*, 31(6), 3-16.

[8] See examples:

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

Hedges, L. & Nowell, A. (1999). Changes in the Black-White Gap in Achievement Test Scores, *Sociology of Education*, 72(2), pp. 111-135.

Jenks, C. & Phillips, M. (1999). *The Black-White Test Score Gap*. Washington, D.C.: Brookings Institution Press

[9] See report Table 22, page 80.

[10] Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity, *Journal of Human Resources*, 37(4), pp. 752-777.

[11] Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity, *Journal of Human Resources*, 37(4), pp. 752-777.

[12] See the CEP website for state profiles http://www.cep-dc.org/index.cfm?fuseaction=document.showDocumentByID&nodeID=1&DocumentID=201.

[13] Olsen, L. (2004, Dec. 8). "Taking Root." *Education Week*. Retrieved June 15, 2007, from http://www.edweek.org/ew/articles/2004/12/08/15nclb-1.h24.html?print=1

The Think Tank Review Project is made possible by funding from the Great Lakes Center for Education Research and Practice.