Nonparametric Comparisons of High-Stakes and Low-Stakes Trends: 2003-2007

Brad Thiessen
Tracey Magda
Andrew Ho
University of Iowa, Iowa City

Please address all correspondence to:
Brad Thiessen
340 Lindquist Center
The University of Iowa College of Education
Iowa City, IA 52242
Phone: 563-333-6160
E-mail: ThiessenBradleyA@sau.edu

Paper presented at the
Annual Meeting of the American Educational Research Association
New York City, March 24-28, 2008

Nonparametric Comparisons of High-Stakes and Low-Stakes Trends: 2003-2007

Brad Thiessen
Tracey Magda
Andrew Ho
University of Iowa, Iowa City

The inflation of high-stakes score trends is a major threat to the validity of No Child Left Behind (NCLB). A flourishing approach to the validation of NCLB score gains compares NCLB trends to corresponding "low-stakes" results such as those from the National Assessment of Educational Progress (NAEP). However, statistics chosen for these comparisons frequently confound trend comparison with choices of cut scores or score scales. A nonparametric framework that overcomes these shortcomings is used to compare State and NAEP trends from 2003-07 in reading and mathematics for 4th and 8th grade. The assumptions underlying cross-test comparisons are discussed to offer possible explanations of State-NAEP trend discrepancies.

In an effort to increase the academic achievement of all students and confront the "soft bigotry of low expectations" (Bush, 2000), the *No Child Left Behind* (NCLB) *Act* was passed into law on January 8, 2002. Like the *Improving America's Schools Act* (IASA) signed in 1994, NCLB required all states to develop content and performance standards; implement assessment systems to track student performance against those standards; and create adequate yearly progress (AYP) goals to ensure all students reach a proficient level of achievement (IASA, 1994; NCLB, 2001). Believing the IASA was ineffective in improving student achievement due to its status as "an undertaking without consequences," (Rotherham, 1999) NCLB granted the federal government the authority to impose sanctions upon states failing to meet AYP goals. The sanctions were intended to provide an incentive for educators to improve the quality of education provided to students and, ultimately, to improve student achievement (Laffont & Martimort, 2001; Jacob, 2007).

Decisions to sanction schools, school districts, and states are made on the basis of state test score trend statistics. The validation of these high-stakes sanction decisions is difficult, given that educators may engage in activities that artificially inflate test score gains. Since the passage of NCLB, newspapers have published more than 150 incidents of public school educators artificially increasing test scores (Thiessen, 2008, p.19). These incidents, along with other research into test score inflation, indicate that educators can artificially increase test scores by manipulating the teaching process (practicing with actual items from the test or focusing instruction only on near-proficient students), manipulating the examinee pool (excluding students from testing or bribing students to increase scores), manipulating the test administration (changing student answers or providing students with extra time), or manipulating score reports (removing or changing scores on official reports) (p. 12). With anywhere from 1%-88% of educators engaging in any one of 36 manipulations of test scores (pp. 20-21), the validation of decisions made from test score trends is extremely important.

One way to provide confirmatory evidence for state test score trends is by comparing trends on state tests to trends on audit tests (Koretz & Hamilton, 2006). The logic is that if state test score gains are legitimate, then those gains should generalize across other tests of the same domain. Score gains artificially inflated by manipulations would be specific to a single test, so they would not generalize to other tests. Thus, significant discrepancies in score trends between state tests and an external audit test could provide evidence of the legitimacy of state test score trends.

A state-level audit test exists in the form of the National Assessment of Educational Progress (NAEP), which has reported state-level Reading and Mathematics results for all 50 states in 2003, 2005, and 2007 for grades 4 and 8. The NAEP is designed and administered in a way that has made it more robust against manipulations. First, NAEP results are not used to make high-stakes decisions regarding the performance of individual educators or schools. Because of this, educators should feel no pressure to manipulate NAEP scores. Second, although some items are released after testing, educators do not have access to NAEP items before it is administered. This eliminates the potential for manipulations due to piracy. Third, by forcing make-up testing for classrooms with less than 90% attendance and by comparing sample demographics to state demographics, the NAEP provides some level of protection against manipulations of the examinee pool (NCES, 2007). Finally, because the U.S. Department of Education hires staff to administer the NAEP and classroom teachers can monitor the administration, the test administration cannot easily be manipulated. Thus, the NAEP provides manipulation-free score trends to which state score trends can be compared.

These State-NAEP trend comparisons have formed the basis of a number of high-profile reports and articles. Reports using 2005 NAEP data for State-NAEP comparisons include those by the Thomas B. Fordham Foundation (2005), the Education Trust (Hall & Kennedy, 2006), Education Week (2006), the Civil Rights Project at Harvard University (Lee, 2006), the Center on Education Policy (2007), and a recent feature in Educational Researcher (Fuller, Wright, Gesicki, & Kang, 2007). Earlier comparisons of State and NAEP trends can be found in articles by Linn, Graue, and Sanders (1990); Koretz and Barron (1998); Klein, Hamilton, McCaffrey, and Stecher (2000); and Linn, Baker, and Betebenner (2002). Trend discrepancies are provocative; they seem to suggest that score gains are legitimate and artificial, achievement is both increasing and decreasing, schools are both improving and declining, and policies are both working for and failing our students.

One reason why these reports provide contradictory results is that the reports track score trends using changes in Percents of Proficient Students (PPS). Trend comparisons based on changes in PPS are known to be deeply dependent on the choice of cut-score for proficiency (Holland, 2002; Ho & Haertel, 2005; Koretz & Hamilton, 2006) and are largely unsuitable for State-NAEP comparisons. Figure 1 illustrates test score distributions from two simulated administrations of the same test. The data were simulated so that the Time 2 distribution has a slightly higher mean and a slightly lower standard deviation; representing the ostensible goals of NCLB to increase overall student achievement and decrease achievement gaps.

The figure shows that 63% of students at Time 1 and 84% of students at Time 2 scored above a cut-score of 500. If this cut-score were defined as the proficiency standard, the school producing these results would be lauded for increasing proficiency by 21%. If, instead, a cut-score of 700 defined proficiency, we would conclude the school was not effective at increasing achievement (26% of students scored above this cut-score at both Time 1 and Time 2). Using a cut-score of 800, PPS-based trend statistics would lead us to conclude that score trends were actually negative (proficiency dropped from 5% to 2%). Thus, the choice of cut-score impacts the conclusions drawn from PPS-based trend statistics.

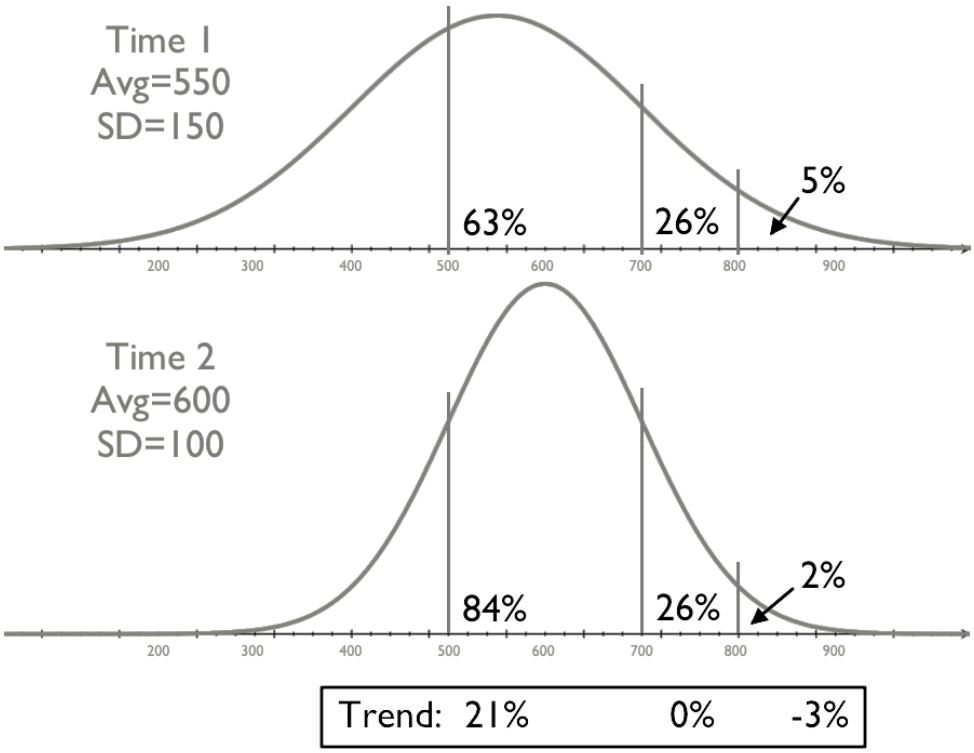


FIGURE 1. *PPS trend discrepancy estimates are dependent upon the choice of cut-score.*

PPS-based trends are not incorrect, but they are shortsighted and certainly misleading as distribution-wide representations of the trends of interest. These problems become compounded in cross-test comparisons, where there are broad discrepancies between cut-scores on NAEP and State tests (McLaughlin & Bandeira de Mello, 2005; Braun & Qian, 2007). Average-based calculations like effect sizes are much more appropriate for these purposes. However, average-based calculations are also susceptible to distortion, including sign-reversal, under monotone transformations of scale (Spencer, 1983; Ho & Haertel, 2005). When the argument for equal-interval scale properties are weak for either or both tests, effect-sizes calculation for trends will rely on pliable scales that preclude straightforward comparison.

Framework: Nonparametric State-NAEP Trend Comparisons

A solution to the vagaries of cut-score and score scale choices across tests is a nonparametric framework based on Probability-Probability (P-P) plots (Haertel, Thrash, & Wiley, 1978; Spencer, 1983; Ho & Haertel, 2005; Livingston, 2006). Again consider the simulated data from a test administered at Time 1 and Time 2. The cumulative distribution functions (CDFs) displayed in Figure 2 display $F_1(x)$ and $F_2(x)$, the percent of students scoring at or below cut-score x at Time 1 and Time 2. The labeled values of 21%, 0%, and -3% show that vertical gaps between two CDFs represent score trends from Time 1 to Time 2.

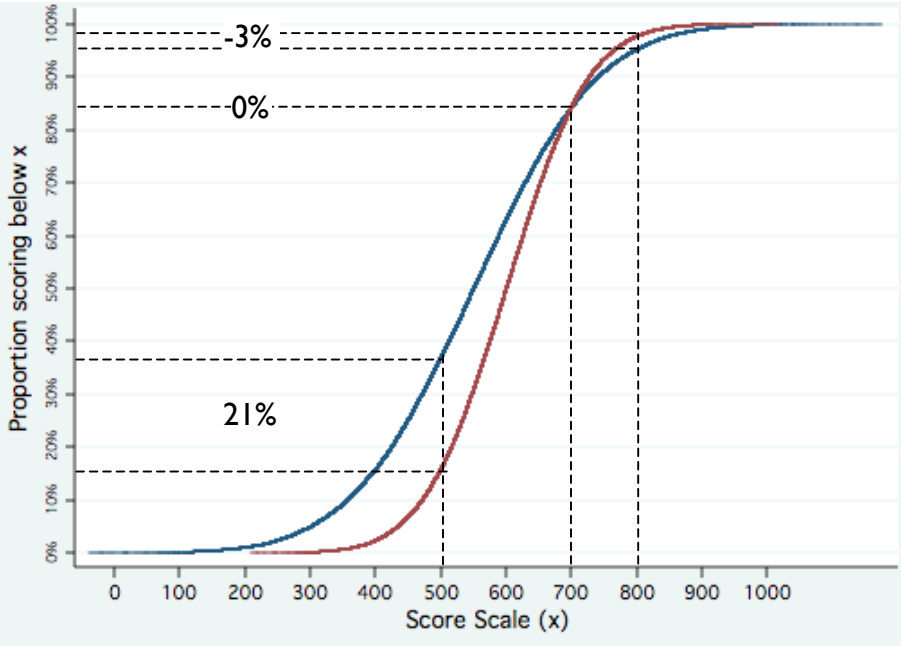


FIGURE 2. *Vertical gaps between CDFs represent score trends from Time 1 to Time 2.*

A P-P plot is, "a comparative plot of sample cumulative probabilities" (Fisher, 1983, p. 31) "constructed solely from vertical slices across CDFs" (Ho, 2007, p. 13). Thus, P-P plots display the vertical gaps between CDFs of test scores administered at Time 1 and Time 2. As Ho (2007) notes, "a monotone transformation of scale may contort the CDFs horizontally, but will not change the vertical relationships between the cumulative proportions" (p. 13). Therefore, P-P plots, and any statistics derived from them, are invariant to transformations of the score scale.

P-P curves, which increase monotonically from the origin to the point (1,1), display the percentiles of one distribution versus the percentiles of another distribution (Holmgren, 1995). When the distributions represent scores from the same test administered twice, the P-P curve shows the proportion of students scoring at or below a given cut-score at each time. In other words, the P-P function displays:

$$p_1 = F_1\left[F_2^{-1}(p_2)\right],$$

or the proportion of Time 1 scores at or below given percentiles of the Time 2 distribution.

A P-P plot generated from the simulated distributions is shown in Figure 3. The diagonal line is shown as a reference point. A P-P function that lies on the diagonal would represent identical score distributions at Time 1 and Time 2, whereas a P-P function that lies mainly above the diagonal would indicate a positive score trend. The vertical lines displayed in Figure 3 show this. The point (.26, .37) on the P-P function shows that 26% of students at Time 2 scored below the 37th percentile from the Time 1 distribution (the same 21% "gain" displayed in Figures 1 and 2).

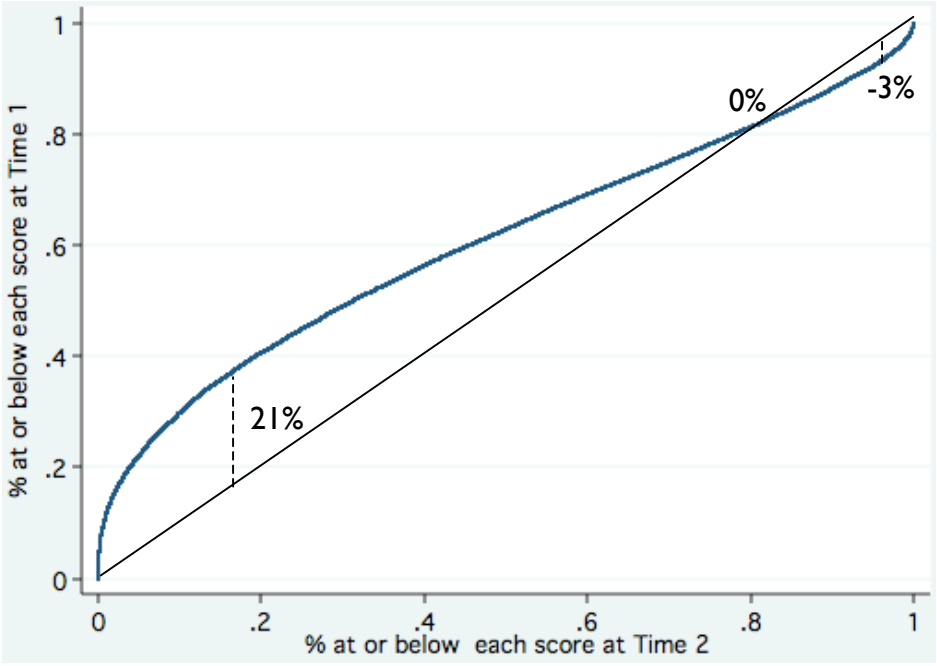


FIGURE 3. *Deviations from the P-P function to the diagonal represent score trends.*

The P-P plot is a scale-independent representation of the difference between two distributions, as horizontal scale distortions of the CDFs do not change either the P-P pairs or, therefore, the P-P plot. Statistics generated from P-P plots are likewise scale-invariant. Since vertical deviations from the P-P plots to the diagonal represent score trends, one useful and interpretable statistic of interest would be the area under the P-P curve.

The area under the P-P curve can be shown to be equivalent to the probability that a randomly drawn score from the Time 2 distribution is greater than a randomly drawn score from the Time 1 distribution. This may be designated as $P(X_2 > X_1)$. For identical distributions at Time 1 and Time 2, the area under the P-P curve (which would fall on the diagonal) would be 0.50. Thus the probability of randomly choosing a Time 2 score greater than a Time 1 score would be 50%. With positive score trends, the P-P curve would fall above the diagonal and the area would be greater than 50%. For the P-P plot in Figure 3, the area under the curve indicates that a randomly chosen Time 2 score has a 61% probability of being greater than a randomly chosen Time 1 score.

The $P(X_2 > X_1)$ statistic has ties to Receiver Operating Characteristic (ROC) Curves, Gini Coefficients and Lorenz Curves in economics, the Kolmogorov-Smirnov statistic, and the Mann-Whitney U statistic, where U is a simple linear transformation of $P(X_2 > X_1)$ (Ho, 2007; Wilk & Gnanadesikan, 1968). Another useful transformation of $P(X_2 > X_1)$ is found by taking an inverse normal transformation,

$$V = \sqrt{2}\Phi^{-1}\left(P(X_2 > X_1)\right),$$

which may be interpreted as the mean difference between two normal distributions with unit standard deviations that would generate a P-P plot with area $P(X_2 > X_1)$. In this sense, the V statistic may be interpreted as a scale-invariant effect size. Unlike regular effect sizes, it cannot be distorted by scale transformations, yet it may still be loosely interpreted as a distance in terms of standard deviation units.

For the P-P curve in Figure 3, $V \approx \sqrt{2}\Phi^{-1}(.611) \approx .40$. This indicates that the Time 2 scores increased by 0.40 standard deviation units over the Time 1 scores. This is supported by the fact that the distributions were simulated to represent an effect size of 0.40.

Methodology: State-NAEP Trend Comparisons

P-P plots may be estimated by the cumulative proportions of students scoring above cut-scores that are reported by many states. This information was collected for grades 4 and 8 in Reading and Mathematics in 2003, 2005, and 2007 corresponding with the grades, subjects, and years tested by NAEP. Table 1 displays a sample of the data collected for South Carolina.

TABLE 1
Example of state and NAEP test score data collected. Numbers represent cumulative percent of students scoring above each of three cut-scores.

South Carolina	State			NAEP		
	2003	2005	2007	2003	2005	2007
4th Grade Math	18.9	21.4	21.9	20.8	18.5	20.3
	65.6	59.5	58.6	68.2	64.1	64.2
	86.0	85.8	80.3	96.1	95.3	95.3
4th Grade Reading	23.6	20.4	17.3	40.6	42.6	41.1
	67.1	63.6	57.8	74.3	74.4	74.2
	97.7	97.1	95.9	94.7	94.2	94.6
8th Grade Math	32.9	33.7	32.1	32.2	28.6	29.1
	80.2	76.8	80.2	73.7	70.1	68.1
	93.7	92.0	93.2	95.2	93.3	92.6
8th Grade Reading	33.2	25.3	28.7	30.6	33.0	31.3
	79.3	70.3	75.4	75.8	75.3	75.4
	97.7	94.1	96.6	98.3	98.1	98.3

The data were then used to plot points on P-P curves representing score trends from 2003-05, 2005-07, and 2003-07. From these points, P-P curves were interpolated and numerical integration procedures were used to estimate the area under the curves. These area estimates

were then transformed to provide estimates of the scale-independent effect size, V , for trends on state and NAEP tests.

Figure 4 displays the interpolated P-P plots and estimated effect sizes for a sample of the data from Table 1. The figure shows that for 8th grade Math from 2003-05, state test trends indicate an increase of 0.25 standard deviation units, while NAEP trends indicate a decrease of 0.03 standard deviation units.

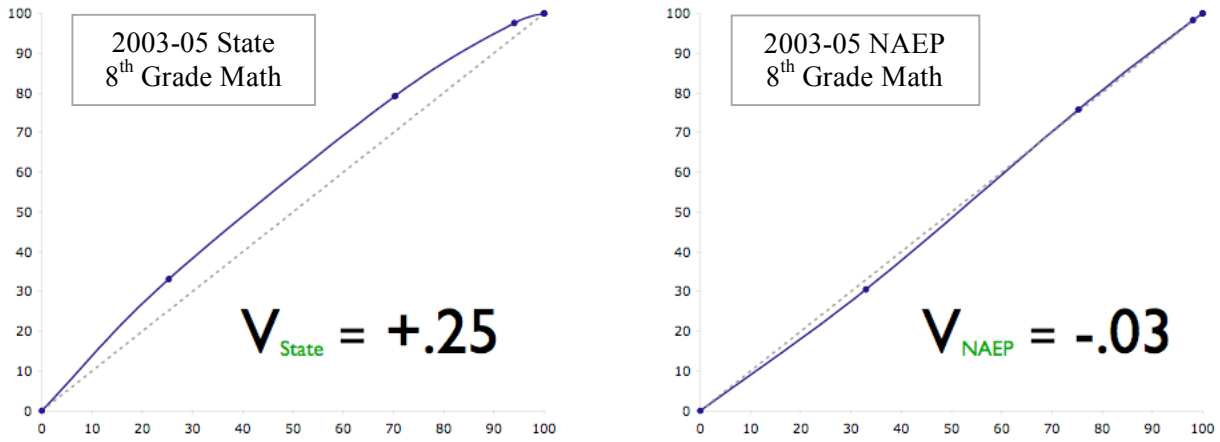


FIGURE 4. *Example of the interpolated P-P plots and estimated effect sizes for 2003-2005 8th grade Reading trends in South Carolina.*

Out of the 600 possible combinations of 50 states, 2 subjects, 2 grades, and 3 trends, the final sample contained 215 combinations from 32 different states. Table 2 displays the states included in the study along with the number of cut-scores used to interpolate P-P plots. Reasons for states being excluded from this study include too few cut-scores (<4) to estimate P-P plots, unreported data, untested grades, changes in tests, or changes in cut-scores across time points.

Results: Discrepancies in State-NAEP Trends

Figure 5 shows the 215 comparable State and NAEP trends for 2003-05, 2005-07, and 2003-07 using the scale-invariant effect size statistic, V . For Figures 5a, b, and c, the null hypothesis that the average State and NAEP trends are equal can be rejected ($p<0.001$), indicating that average state trends are significantly more positive than average NAEP trends for each of the time periods.

These average trends, and the discrepancies between State and NAEP trends, are displayed in Table 3. Table 3 shows that for 2003-05, the average State trend was +0.118 standard deviation units, while the average NAEP trend was +0.034 standard deviation units. Thus, trends reported from state tests over this time period were 3.462 times more positive than trends reported from NAEP.

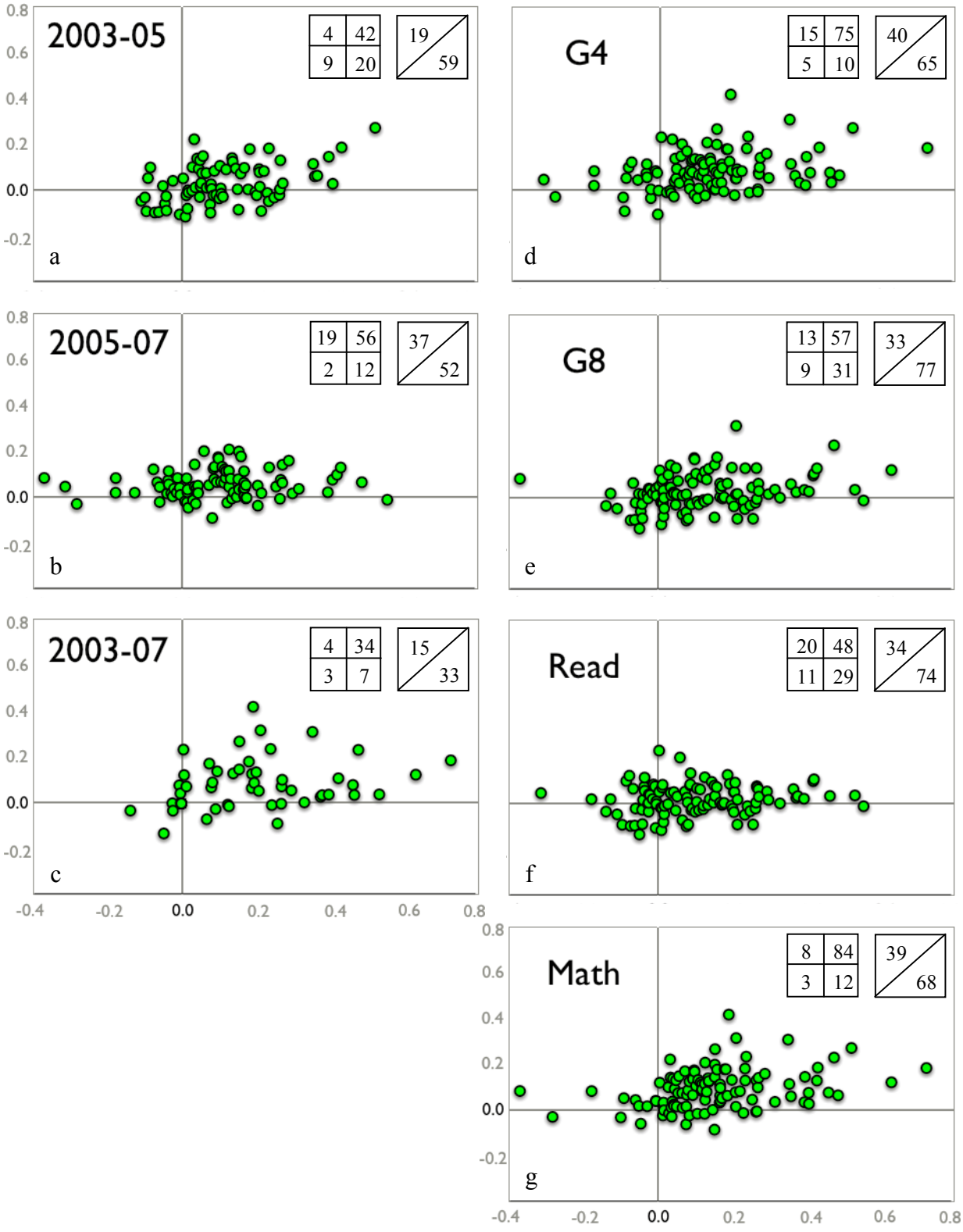
TABLE 2

The number of cut-scores reported by states included in the final sample of data.

	Number of cut-scores reported on state tests											
	Grade 4						Grade 8					
	Reading			Mathematics			Reading			Mathematics		
	03-05	05-07	03-07	03-05	05-07	03-07	03-05	05-07	03-07	03-05	05-07	03-07
Alabama		3	3		3	3		3	3		3	3
Alaska												
Arizona	3	3	3	3	3	3	3	3	3	3	3	3
Arkansas		3			3			3			3	
California	4	4	4	4	4	4	4	4	4	4		
Colorado	3	3	3		3		3	3	3	3	3	3
Connecticut	4			4			4			4		
Delaware							4	4	4	4	4	4
Florida	4	4	4	4	4	4	4	4	4	4	4	4
Georgia												
Hawaii		3			3		3	3	3	3	3	3
Idaho	3	3	3	3	3	3	3	3	3	3	3	3
Illinois							3			3		
Indiana												
Iowa												
Kansas				4	4	4	4	4	4			
Kentucky		3									3	
Louisiana	4	4	4	4	4	4	4	4	4	4	4	4
Maine	3	3	3	3	3	3	3	3	3	3	3	3
Maryland												
Massachusetts		3			3						3	
Michigan	3	3	3	3	3	3				3	3	3
Minnesota												
Mississippi	3	3	3	3	3	3	3	3	3	3	3	3
Missouri				3						3		
Montana	(3)	(3), 3	3	(3)	(3), 3	3	(3)	(3), 3	3	(3)	(3), 3	3
Nebraska												
Nevada												
New Hampshire												
New Jersey												
New Mexico												
New York	3			3			3			3		
North Carolina	3			3			3			3		
North Dakota		3			3			3			3	
Ohio		4						4			4	
Oklahoma		3			3		3	3	3	3	3	3
Oregon												
Pennsylvania							3	3	3	3	3	3
Rhode Island												
South Carolina	3	3	3	3	3	3	3	3	3	3	3	3
South Dakota												
Tennessee												
Texas	4			4			4			4		
Utah												
Vermont												
Virginia												
Washington	3	3	3	3	3	3						
West Virginia		4			4			4			4	
Wisconsin	3	3	3	3	3	3	3	3	3	3	3	3
Wyoming	3			3			3			3		

Notes: Blank cells represent data excluded from the analysis

Montana administered 2 tests in 2003 & 2005. The (ITBS) changed from high- to low-stakes in 2005.



FIGURES 5a-f. *State and NAEP trends by year, grade, and subject.*

TABLE 3
Summary of State-NAEP discrepancies by year, grade, and subject.

	Number of paired trends	Centroid (State, NAEP)	Discrepancy	State / NAEP	% of cases with State > NAEP
2003-05	78	(.118, .034)	.084	3.462	76%
2005-07	89	(.102, .062)	.040	1.647	58%
2003-07	48	(.210, .090)	.119	2.320	69%

Note: All discrepancies were found to be significant at $p<.01$

Figure 5 also shows the paired trend statistics by quadrant and by their location with respect to the diagonal. For 2003-05, more than three-fourths, or 59 out of 78 (76%), are below the diagonal. These are cases in which State trends are more positive than NAEP trends. 4 cases (5%), located in the first quadrant, showed positive NAEP trends and negative State trends, and 20 cases (26%) in the fourth quadrant showed positive State trends and negative NAEP trends. Thus, almost one-third of all cases showed a reversal in sign between State and NAEP trends.

As Ho (2007) notes, "None of these comparisons takes the measurement error of state tests into account. Disattenuation will spread the points horizontally away from the *y*-axis in approximate inverse proportion to the square root of the reliability of the state tests. Repeating the *t*-test assuming imperfect state test reliabilities would augment the overall finding of significantly different trends" (p. 17).

Results by Year

Figures 5a-c and the first three rows of Table 3 allow for comparisons among State-NAEP trend discrepancies for the periods of 2003-05, 2005-07, and 2003-07. While discrepancies between were statistically significant ($p<0.01$) for all three trends, the average discrepancy was twice as large for 2003-05 (0.084 standard deviation units) as it was for 2005-07 (0.040 standard deviation units). Likewise, while 76% of states reported higher State trends than NAEP trends from 2003-05, this fell to 58% of states for 2005-07. The 2005-07 period did show a greater percentage of states with sign reversals for State and NAEP trends, with 35% of cases falling in the first or fourth quadrants.

Of the 60 cases from which both 2003-05 and 2005-07 trend discrepancies could be estimated, 31 (52%) of the cases experienced bigger discrepancies in 2003-05 and 29 (48%) of cases experienced bigger discrepancies in 2005-07. Figure 6 shows the relationship between trend discrepancies for 2003-05 and 2005-07. The correlation between discrepancies for 2003-05 and 2005-07 was calculated to be -0.14. 5 cases (8%) reported NAEP trends more positive than State trends in both 2003-05 and 2005-07; 11 cases (18%) reported more positive NAEP trends in 2003-05 and more positive State trends in 2005-07; 18 cases (30%) reported more positive State trends in 2003-05 and more positive NAEP trends in 2005-07; and 26 cases (43%) reported State trends more positive than NAEP trends in both time periods.

Over the four-year period from 2003-07, the results show that State trends were 2.32 times more positive than NAEP trends, with 69% of states reporting State trends higher than

NAEP trends. Thus, these results show that while discrepancies were bigger for 2003-05 than for 2005-07, reported State trends were consistently larger than NAEP trends. The map in Figure 7 displays the average trend discrepancy over the 2003-07 time period. The map shows that only five states reported trends smaller than NAEP trends. The widespread display of positive trend discrepancies could provide possible evidence of State test score inflation.

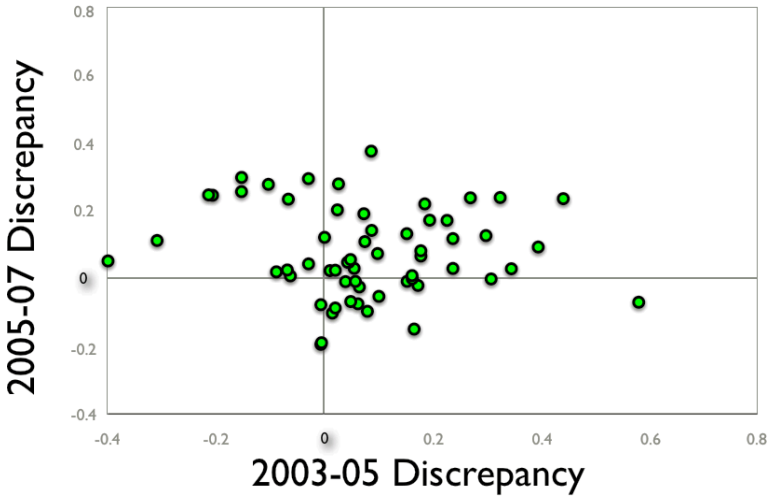


FIGURE 6. *Relationship between 2003-05 and 2005-07 State-NAEP trend discrepancies.*

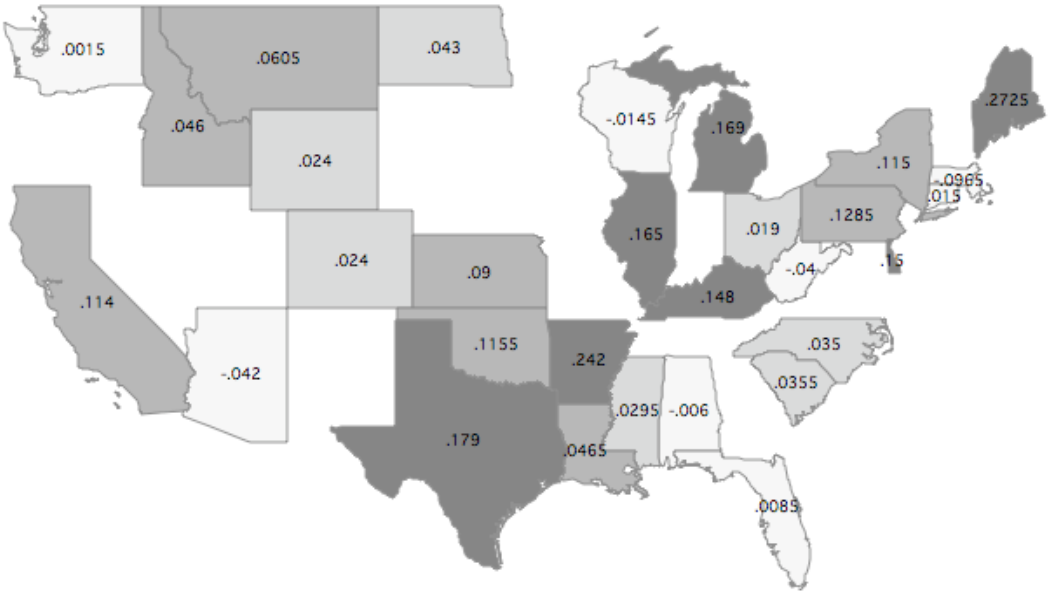


FIGURE 7. *Average State-NAEP discrepancies for each state.*

Results by Grade, Subject, and Year

Figure 5d-f and Table 3 also show the results disaggregated by subject and grade level. Looking at the table, discrepancies were larger for 8th grade (State gains 3.976 times higher than NAEP gains) than for 4th grade (State trends 1.611 times more NAEP trends). Discrepancies were also larger for reading (State trends 5.221 times greater than NAEP trends) than for mathematics (State trends 1.609 times greater than NAEP trends). The larger discrepancies in reading may be due to content differences. Some states included in this study only reported English Language Arts (ELA) test results, so these ELA tests would be expected to differ somewhat in content from the NAEP reading test

TABLE 4
Summary of State-NAEP discrepancies by year, grade, and subject.

	Number of paired trends	Centroid (State, NAEP)	Discrepancy	State / NAEP	% of cases with State > NAEP
4th Grade	105	(.139, .086)	.053	1.611	62%
8th Grade	110	(.125, .031)	.094	3.976	70%
Reading	108	(.110, .021)	.089	5.221	69%
Mathematics	107	(.154, .096)	.058	1.609	64%
4th Grade Read	53	(.108, .048)	.059	2.230	66%
4th Grade Math	52	(.171, .125)	.046	1.367	58%
8th Grade Read	55	(.112, -.005)	.117	n/a	71%
8th Grade Math	55	(.138, .068)	.070	2.027	69%
03-05 4th Read	18	(.107, .016)	.091	6.870	89%
05-07 4th Read	23	(.066, .065)	.001 (p=.49)	1.010	52%
03-05 8th Read	21	(.080, -.031)	.111	n/a	81%
05-07 8th Read	22	(.112, .021)	.091	5.314	52%
03-05 4th Math	18	(.173, .133)	.040 (p=.13)	1.302	63%
05-07 4th Math	22	(.128, .082)	.046 (p=.10)	1.559	64%
03-05 8th Math	21	(.118, .030)	.087	3.880	76%
05-07 8th Math	22	(.104, .079)	.025 (p=.14)	1.314	59%

Note: Unless otherwise noted, discrepancies were found to be significant at p<.01

Looking at subject-grade combinations, 8th grade reading trends showed the greatest average discrepancy. While states reported an average 8th grade reading trend of +0.112 standard deviation units, the average NAEP trend for these same states was -0.005 standard deviation units. The smallest State-NAEP discrepancy was found for 4th grade math, with an average State trend of 0.171 standard deviation units and an average NAEP gain of 0.125 standard deviation units.

For subject-grade-year combinations, the largest average State-NAEP discrepancy (0.111 units) was found for 8th grade reading in 2003-05, while the smallest average discrepancy (0.001 units) was found for 2005-07 4th grade reading. Furthermore, while three of the four average discrepancies on reading tests were found to be significant, only one of the four math

discrepancies (2003-05 8th grade math) showed a significant State-NAEP trend discrepancy. The lack of statistical significance may simply be due to the small sample sizes in these analyses.

Comparison to Other Trend Measures

If it were possible, it would have been preferred to estimate state-NAEP trend discrepancies by comparing traditional effect sizes, such as Cohen's d:

$$d = \frac{\bar{X}_2 - \bar{X}_1}{s_{\text{pooled}}} = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2 + s_2^2}{2}}},$$

which, in this study, would be defined as the difference between mean scores at Time 1 and Time 2 divided by a pooled standard deviation (Cohen, 1988). Unfortunately, many states do not report this information, so traditional effect sizes could not be calculated.

For the 15 cases in which traditional effect sizes (d) could be calculated, Figure 8 displays the relationship between these d values and the estimated V statistics. For these 15 cases, the average d effect size was found to be 0.0448 and the average V effect size was found to be 0.0618. Thus, the average traditional effect size was found to be smaller than the average V statistic by 0.0170.

The correlation between d and V estimates displayed in Figure 8 was found to be 0.729, indicating a strong positive linear relationship. The single obvious outlier, labeled on the figure, was for 2005-07 8th grade reading in Delaware. For this case, d was -0.028 and V was 0.209. Eliminating this outlier, the correlation increases to 0.971 and the difference between d and V estimates shrinks to 0.001. It is not known why Delaware is an outlier, although the means and standard deviations could have been misreported.

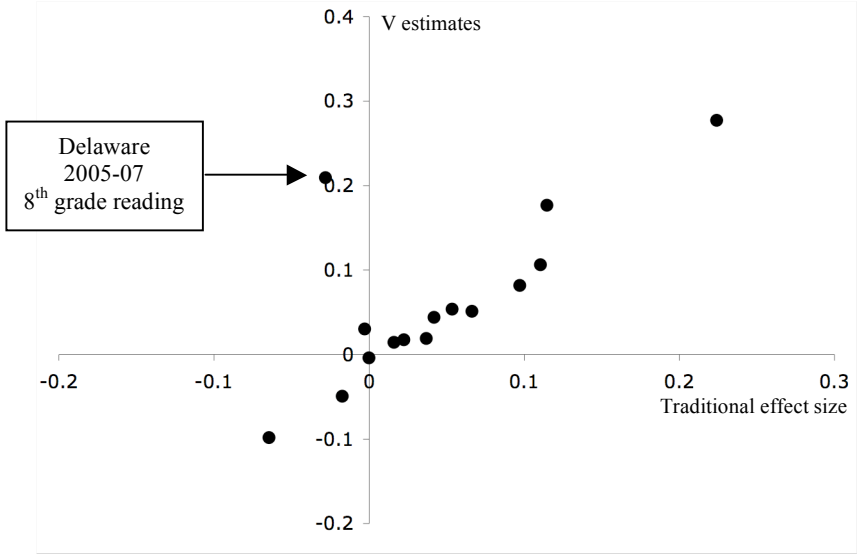


FIGURE 8. *Relationship between d and V trend effect size estimates for States.*

The V and d statistics can also be compared for NAEP trends. Figure 9 displays this relationship. Once again, a strong linear relationship exists, with a correlation of 0.977. The difference between the average d (0.0479) and the average V (0.0489) was only 0.001. This provides reassurance that the V estimates do, in fact, provide an effect-size measure of score trends.

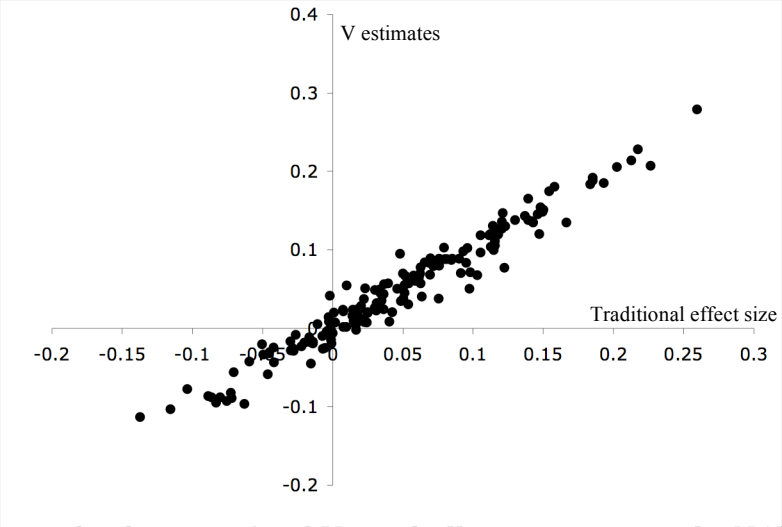


FIGURE 9. *Relationship between d and V trend effect size estimates for NAEP.*

The results from Figures 8 and 9 show that the V statistics used in this study are similar to traditional effect sizes. As was explained earlier, the V statistics should be expected to differ, somewhat, from the shortsighted PPS-base trend statistics. Figure 10 displays the relationship between the V statistics and the change in the percentage of students scoring proficient on state tests. The correlation between proficiency trends and V statistics in Figure 10 was found to be 0.880.

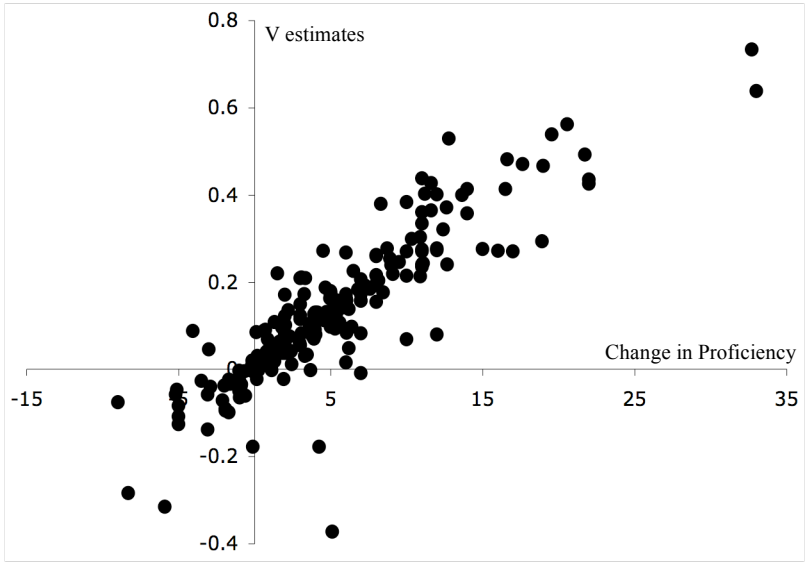


FIGURE 10. *Relationship between PPS-based and V trend statistics.*

Providing further evidence of the shortcomings of the change in proficiency statistics, the observations falling in the second and fourth quadrants of Figure 4.x represent cases with sign reversals in proficiency trends and V statistics. In the figure, 2% of the observations represent cases in which proficiency declined yet V statistics indicated a positive score trend. 4% of the observations represent cases in which proficiency increased yet V statistics indicated a negative score trend. This could provide possible evidence of states focusing instructional resources on students closest to proficiency (thus increasing proficiency in the face of negative score trend effect sizes), although other explanations are possible.

Discussion

Using a scale-invariant framework, State trends in 4th and 8th grade reading and mathematics for 2003-05, 2005-07, and 2003-07 were found to be significantly more positive than NAEP trends. While some researchers have suggested that discrepancies indicate state test results may be inflated due to manipulations (Hall & Kennedy, 2006; Jacob, 2007; Kleine, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Lee, 2006; Peterson & Hess, 2005, 2006; Ravitch, 2005), it must be noted that the existence of these discrepancies does not necessarily mean state test scores have been inflated through manipulation.

Jacob (2007) notes, "… there has been little research on reasons why student performance differs between NAEP and local assessments" (p. 11). This research is important because, as Koretz (2001) states, in order to conclude that a discrepancy between state and NAEP results, "reflects specific policies or practices, one needs to be able to reject with reasonable confidence other plausible explanations…" (p. 20).

Hill (1998), Ho and Haertel (2007), the Iowa Department of Education (2007), Jacob (2007), and Koretz (1999) all address plausible rival hypotheses that may explain any discrepancies between state and NAEP results. Synthesizing this research, some of these plausible rival hypotheses include:

- Differences in content coverage or sequence or opportunity to learn
- Differences in item formats or administration mode (paper- or computer-based)
- Differences in test difficulty
- Differences in score standards or standard-setting procedures
- Differences in test administration procedures/environment or administration date
- Differences in accommodations allowed during testing
- Differences in examinee populations or subgroup definitions
- Differences in examinee motivation or effort

The first four plausible rival hypotheses address differences between state and NAEP tests and scoring procedures. If a state test differs from NAEP in content coverage or sequence, then it would be expected that students would score higher on the state test (due to educators focusing on state content standards). Likewise, differences in item formats, test administration mode, or test difficulty may have a significant impact on score discrepancies between state tests and NAEP.

The next two rival hypotheses for score discrepancies address differences in test administration procedures. If state test administration procedures significantly differ from NAEP procedures (in terms of testing time, use of accommodations, or use of materials such as calculators during testing), then discrepancies in results between the two tests would not be completely unexpected.

The final three possible explanations for score discrepancies deal with potential differences in the examinees being tested under state tests and NAEP. While NCLB requires at least 95% of students to be tested annually and NAEP sets its standard at 85% (Hill, 1998, p. 3), this means that up to 20% of examinees could have been excluded from at least one of the tests. Clearly, these potential differences in the examinee pool could impact discrepancies between results from the two tests. Also, since state tests under NCLB are high-stakes and NAEP remains a relatively low-stakes test, differences in examinee motivation or effort could have an impact on discrepancies.

These plausible rival hypotheses are not exhaustive, but they do provide a reminder that discrepancies between state and NAEP score trends could be due to a combination of many factors. In order to attempt to show that state test scores have been inflated through manipulation, strong assumptions must be made that the discrepancies are not due to the above plausible rival hypotheses. While several studies have concluded that differences in test content (Wei, Shen, Lukoff, Ho, & Haertel, 2006), examinee motivation (Klein, Hamilton, McCaffrey, & Stecher, 2000; Linn, Baker, & Betebenner, 2002), examinee demographics, test item formats, and test administration time limits (Jacob, 2007) cannot explain the discrepancies between state test and NAEP results, the existence of these differences should at least temper expectations about the comparability of state test and NAEP score trends.

Another assumption implicitly made in comparing state and NAEP score trends is that results from NAEP are somehow the "gold standard." While NAEP may not be the gold standard, it may be the only available standard with which to compare the performance of states in reading and mathematics achievement. While NAEP scores have been more robust against educator manipulations, Hill (1998) notes, "As more and more states see the need for increased NAEP scores, practices will evolve that will virtually ensure gains on NAEP" (p. 10). Thus, Hill suggests that if NAEP results are used to validate state test results, NAEP results will become high-stakes and NAEP will become subject to the same manipulations as state tests.

REFERENCE LIST

Braun, H. & Qian, J. (2007). Mapping 2005 state proficiency standards onto the NAEP scales. U.S. Department of Education, NCES 2007-482. Retrieved September 6, 2007 from: http://nces.ed.gov/nationsreportcard/pdf/studies/2007482.pdf

Bush, G. (2000). Speech delivered at the 91st annual convention of the National Association for the Advancement of Colored People. Baltimore, Maryland. July 10, 2000. Retrieved September 3, 2007 from Washington Post web site: http://www.washingtonpost.com/wp-srv/onpolitics/elections/bushtext071000.htm

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum

Education Week (2006). Quality Counts At 10: A Decade of Standards-Based Education. Vol. 25, Issue 17.

Fisher, N. I. (1983). Graphical methods in nonparametric statistics: A review and annotated bibliography. *International Statistical Review*, 51(1), 25-58.

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: how to judge No Child Left Behind? *Educational Researcher*, 26(5): 268-278.

Hall, D., & Kennedy, S. (2006). Primary Progress, Secondary Challenge. Report. Retrieved on September 12, 2007 from http://www2.edtrust.org/NR/rdonlyres/15B22876-20C8-47B8-9AF4-FAB148A225AC/0/PPSCreport.pdf

Haertel, E., Thrash, W., & Wiley, D. (1978). Metric-free distributional comparisons. Report. ML-Group for Policy Studies in Education, Chicago, IL.

Hill, R. (1998). Using NAEP to compare states' data – while it's still possible. Paper presented at the 1998 Annual Meeting of the National Council on Measurement in Education, San Diego.

Ho, A.D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice*, 26(4), 11-20.

Ho, A.D. & Haertel, E.H. (2006a). Metric-free measures of test score trends and gaps with policy-relevant examples. CSE Technical Report #665. University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

Ho, A.D. & Haertel, E.H. (2007). *Apples to apples? The underlying assumptions of state-NAEP comparisons.* CCSSO Policy Brief. Retrieved February 23, 2008 from: http://www.ccsso.org/content/PDFs/Ho%20Haertel%20CCSSO%20Brief2%20Final.pdf

Holland, P.W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, 27(1): 3-17.

Holmgren, E. B. (1995). The p-p plot as a method for comparing treatment effects. *Journal of the American Statistical Association*, 90(429), 360-365.

Improving America's Schools Act of 1994, Pub. L. No. 103-382. (1994). Retrieved September 3, 2007 from The U.S. Department of Education web site: http://www.ed.gov/legislation/ESEA/toc.html

Iowa Department of Education (2007). Comparing NAEP and ITBS results. Retrieved September 18, 2007 from the Iowa Department of Education web site: http://www.iowaccess.org/educate/ecese/nclb/doc/comparing_naep_itbs.pdf

Jacob, B.A. (2007). Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments. Working paper 12817, National Bureau of Economic Research. Retrieved September 1, 2007 from NBER web site: http://www.nber.org/papers/w12817

Klein, S. Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What do test scores in Texas tell us? Report, The RAND Corporation, Santa Monica, CA. Retrieved September 5, 2007 from: http://www.rand.org/pubs/issue_papers/IP202/index2.html

Koretz, D. (1999). Limitations in the use of achievement tests as measures of educators' productivity. Presentation at the *Devising Incentives to Promote Human Capitol National Academy of the Sciences Conference*. Irvine, CA, December 18, 1999.

Koretz, D. (2001). State comparisons using NAEP: large costs, disappointing benefits. *Educational Researcher,* 20(3), 19-21.

Koretz, D., & Barron, S. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS). Report, The RAND Corporation, Santa Monica, CA.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), Educational Measurement (4th ed., pp. 531-578). American Council on Education and Praeger Publishers. Westport, Connecticut.

Laffont, J.J. & Martimort, D. (2001). The theory of incentives: The principal-agent model. Princeton: Princeton University Press.

Lee, J. (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. Civil Rights Project. Harvard University, Cambridge, MA.

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(4): 431-435.

Livingston, S. A. (2006). Double p-p plots for comparing differences between two groups. *Journal of Educational and Behavioral Statistics*, 31(4), 431-435.

McLaughlin, D., Bandeira de Mello, V. (2005). How to compare NAEP and state assessment results. Presented at the 35th Annual National Conference on Large-Scale Assessment. Retrieved April 18, 2007 from web site: http://38.112.57.50/Reports/LSAC_20050618.ppt

National Center for Education Statistics (NCES) (2007). Item development process. Retrieved September 12, 2007 from the NCES web site: http://nces.ed.gov/nationsreportcard/contracts/item_dev.asp

No Child Left Behind Act of 2001, Pub. L. No. 107-110. (2001). Retrieved September 3, 2007 from The U.S. Department of Education web site: http://www.ed.gov/policy/elsec/leg/esea02/index.html

Peterson, P. & Hess, F. (2005). Johnny can read… in some states. Retrieved June 22, 2007 from *Hoover Institution* web site: http://www.hoover.org/publications/ednext/3219636.html

Peterson, P. & Hess, F. (2006). Keeping an eye on state standards. Retrieved June 22, 2007 from *Hoover Institution* web site: http://www.hoover.org/publications/ednext/3211601.html

Ravitch, D. (2005). Every state left behind. *The New York Times*, November 7, 2005. Retrieved September 13, 2007 from The Brookings Institution web site: http://www.brookings.edu/views/op-ed/ravitch/20051107.htm

Rotherham, A. (1999). Toward performance-based federal education funding: Reauthorization of the Elementary and Secondary Education Act. Progressive Policy Institute Policy Report, April 1, 1999. Retrieved September 3, 2007 from the Democratic Leadership Council web site: http://www.ndol.org/documents/ESEA.pdf

Spencer, B. (1983). On interpreting test scores as social indicators: Statistical considerations. *Journal of Educational Measurement*, 20: 317-333.

Thiessen, B.A. (2008). The effectiveness of test security policies on deterring educators from manipulating test scores. Dissertation proposal; University of Iowa. Retrieved February 17, 2008 from web site: http://homepage.mac.com/bradthiessen/pubs/proposal.pdf

Thomas B. Fordham Foundation (2005). Gains on State Reading Tests Evaporate on 2005 NAEP. Report. Downloaded in November, 2005 from http://www.edexcellence.net/foundation/about/press_release.cfm?id=19

Wei, X., Shen, X., Lukoff, B., Ho, A.D., & Haertel, E.H. (2006). Using test content to address trend discrepancies between NAEP and California State Tests. ERIC # ED491544. Retrieved September 6, 2007 from: http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/29/00/1f.pdf

Wilk, M.B. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1), 1-17.