FREQUENCY OF SIMPSON'S PARADOX IN NAEP DATA

James Terwilliger, NAEP Coordinator, Minnesota Department of Education
Milo Schield, Augsburg College, Director of the W. M. Keck Statistical Literacy Project

Abstract: Simpson's Paradox occurs for two states when their difference in scores has the opposite sign of the score differences for each of the state subgroups. Simpson's Paradox is a specific manifestation of statistical confounding. The paradox has been understood for many years but is usually regarded as simply a curious anomaly. The purpose of this paper is to show that Simpson's Paradox is not rare in NAEP data. NAEP public-school data are analyzed for 2000n Grade 4 Math and 2002 Grade 8 Reading. Conditions for a Simpson's reversal are presented. Approximately 100 instances of Simpson's Paradox are found per data set based on the influence of three confounders: family income, school location and race/ethnicity. In analyzing the influence of race/ethnicity two approaches are used. A straight forward approach generated 64 Simpson's reversals in the NAEP 2002 Grade 8 reading data of which 18 involve initial differences that are statistically significant. A more liberal approach generated 117 Simpson's reversals in the same data set of which 52 involve initial differences that are statistically significant. Either way these results support the claim that Simpson's Paradox is not rare in NAEP data. As a percentage of all pairs of state differences in the same data that are statistically significant, 4% are reversed using a conservative approach while 10% are reversed using a more liberal approach. All Simpson's reversals – whether statistically significant or not – are argued to have 'journalistic significance' because of their political significance. Recommendations include ordering the data by key confounders as an adjunct when reporting results. The failure to allow adjustments for confounders can lead to a serious misinterpretation of the results which in turn can lead to questionable policies.
Keywords: Confounding, Standardization

1. THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP)

NAEP is a unique large-scale assessment program. For over 30 years NAEP has collected data on national samples of 4th, 8th and 12th graders. In 1990 NAEP began a biennial state-level assessment program which yields average scores for individual states in mathematics and reading at the 4th and 8th grade levels. NAEP offers the most reliable and widely acknowledged measure of student achievement across states. It is often referred to as the 'Gold Standard' in assessment. This study reports on the analysis of NAEP public school data from two data sets:

1. NAEP 2000n Grade 4 Math: The 'n' in '2000n' refers to data from students that were not allowed any special accommodations.[1] Data are available for 41 jurisdictions.[2]
2. NAEP 2002 Grade 8 Reading: Use of accommodations as needed. Data available for 42 jurisdictions.[2]

To ensure the robustness of the results, data sets were chosen that involved different years (2002 vs. 2000), different tests (reading vs. math) and different grades (Grade 8 vs. Grade 4).

2. NAEP SCORES VERSUS PREVALENCE OF CONFOUNDERS

The NAEP 2000n Grade 4 math test is the basis for all the data in this section. Consider the influence of family income (school lunch payment status), school location (center city, urban-fringe and rural) and race/ethnicity (white, black, Hispanic and Asian) on the association between states and their NAEP state scores.[3] The following plots show these associations.

[1] Accommodations are any non-standard conditions involved in the testing, e.g., allowing extra time. In 2000 NAEP was making a transition from traditional assessment in which no accommodations were permitted to the use of accommodations. That year the samples were randomly split with half the students being permitted to use accommodations that were deemed necessary while the other half was NOT allowed to use accommodations.
[2] Only jurisdictions from the contiguous 48 states were analyzed plus the Department of Defense schools: DESS and ODDS.
[3] In the interest of brevity, we refer to the mean NAEP score for a state simply as 'the state score.'

Figure 1 plots state scores by the percentage of students who are not eligible for free or reduced-cost school lunch. Being non-eligible is based on a higher family income and it means having to pay full cost for school lunch. So, 'non-eligible', 'pay' and 'high income' are used interchangeably as are 'eligible,' 'non-pay' and 'low income.' The straight line models the relation between the state scores and the percentage of students who are non-eligible. The circles identify pairs of states that are examples of Simpson's reversals. These examples will be analyzed in detail in the next section.

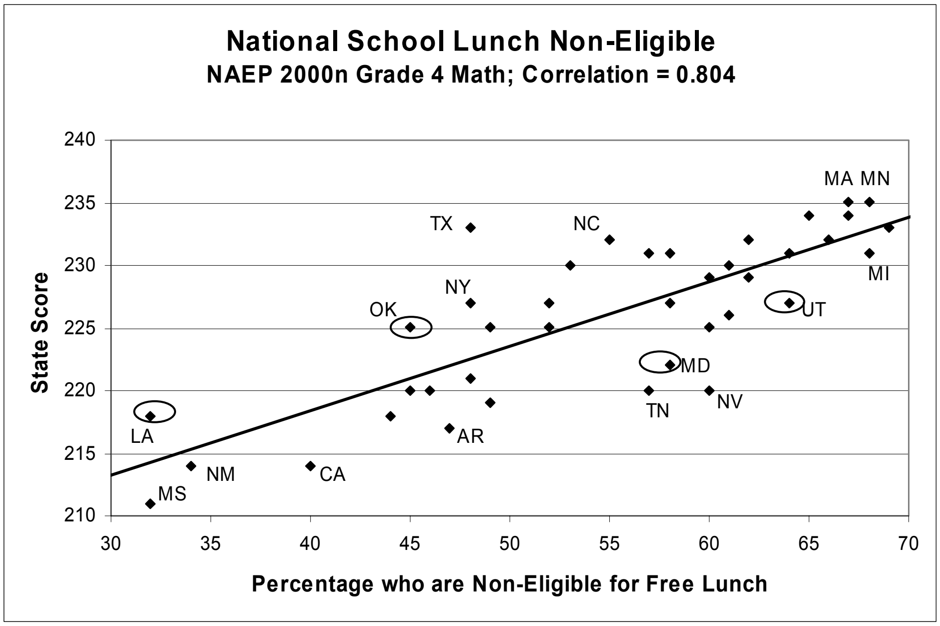

Figure 1: State Scores vs. Percentage who are Non-Eligible for Free Lunch

Figure 2 plots state scores by the percentage of students who attend a central city school (as opposed to a rural or an urban-fringe school). As before, the circles identify examples of Simpson's reversals.

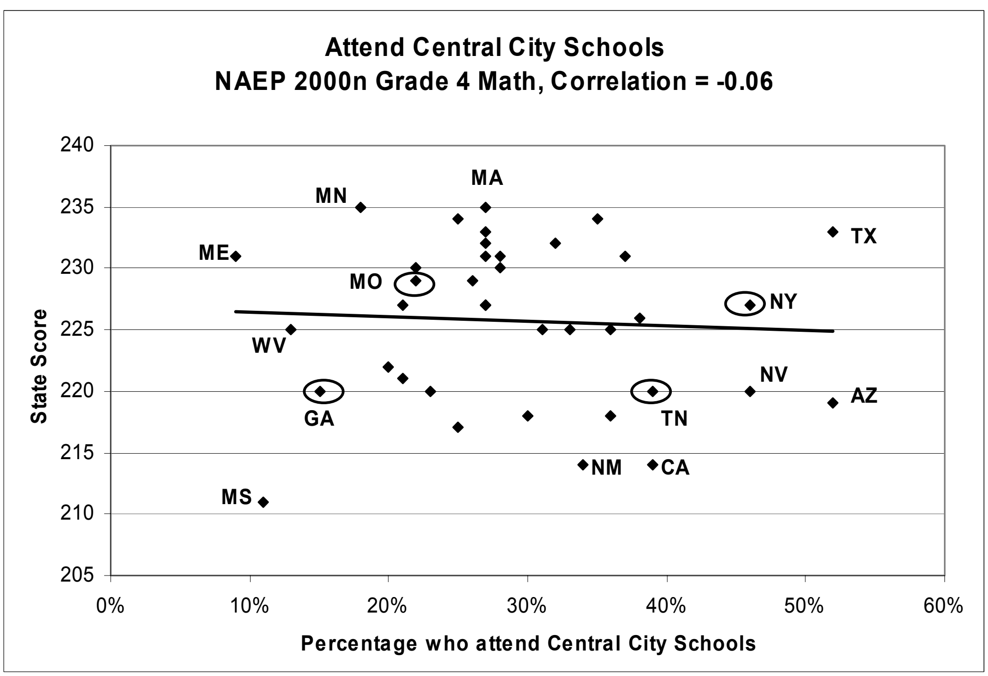

Figure 2: State Scores vs. Percentage who attend a Central City School

Figure 3 plots state scores by the percentage of students who are white (as opposed to non-white). Blacks, Hispanics and Asians are considered as non-whites in this case. Note that Simpson's reversals are not limited to those states circled. Those circled are just examples that will be analyzed in detail.

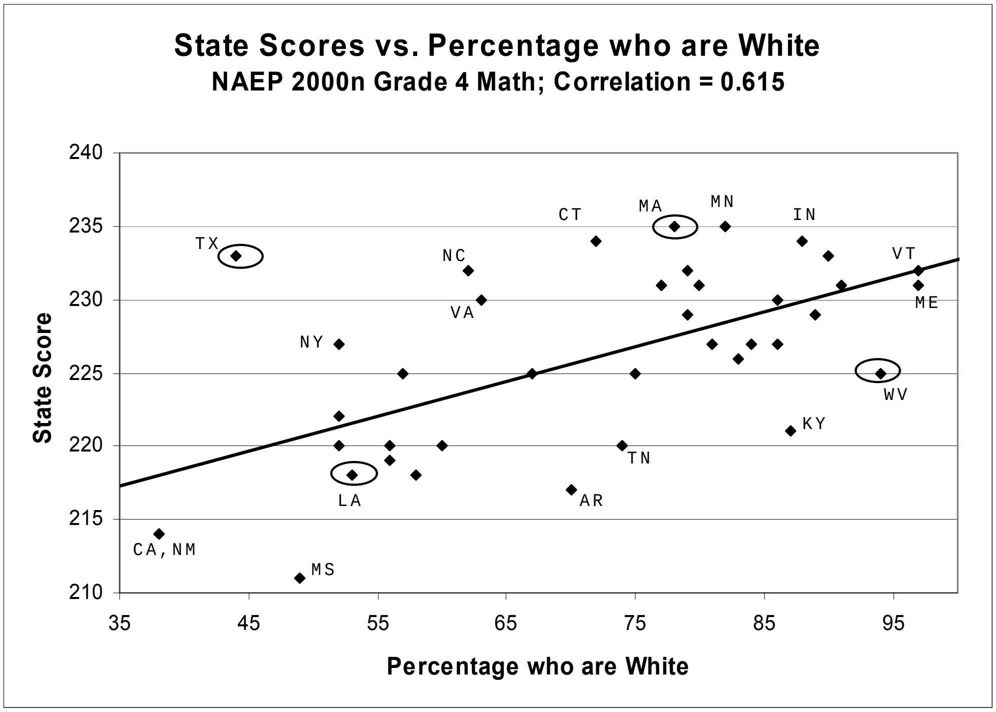

Figure 3: State Scores vs. Percentage of Students who are White

An obvious point in Figure 1 and Figure 3 is the strong association between state scores and the associated factor. The next step is to investigate specific examples of Simpson's reversals.

3. EXAMPLES OF REVERSALS AND CHANGES IN NAEP DATA

The NAEP 2000n Grade 4 math test is the basis for all the data in this section. Consider the influence of family income (school lunch payment status), school location (center city, urban-fringe and rural) and race/ethnicity (white, black, Hispanic and Asian) on the association between states and their NAEP state scores. The following tables present specific data for each of these confounders.

Table 1 shows state scores broken out by family income[4]. As shown in Table 1A, the state score is two points lower for Oklahoma (OK) than for Utah (UT). Yet when classified on family income (based on school lunch payment status), the state score for each subgroup score is higher for Oklahoma than for Utah. Note that the percentage of high income families is larger in Utah (64%) than in Oklahoma (45%) and students from high income families tend to score higher than those from low income families.

Table 1: State Scores Classified by Family Income

State	All	High $	Low $
UT	227	233	216
OK	↓225↓	↑234↑	↑218↑

Table 1A UT vs. OK

State	All	High $	Low $
MD	222	233	207
LA	↓218↓	233	↑211↑

Table 1B: MD vs. LA

[4] Federal guidelines identify a federal income-related criterion under which students in low-income families receive free or reduced-fee school lunches. Based on student responses, students were classified into four groups: Not eligible, eligible, 'Don't know' and 'No answer.' NAEP generated scores for the first three groups and the state average. To reduce the subgroups to just two categories, students in the last three subgroups were combined. Given the score and prevalence of those in the 'High Income Family' sub group plus the state average, the score for those in the 'Low-Income Family' subgroup was calculated.

As shown in Table 1B, the state score is four points lower for Louisiana (LA) than for Maryland (MD). Yet when classified on family income, the state score for each subgroup is at least as high for Louisiana as for Maryland. Note that the percentage of high income families is greater in Maryland (58%) than in Louisiana (32%), and students from such families tend to score higher.

Table 2 shows state scores broken out by school location[5]. As shown in Table 2A, the state score is two points lower for New York (NY) than for Missouri (MO). Yet when classified by school location, the state score for each subgroup is at least as high for New York as for Missouri. Note that the percentage of students who attend non-city schools is higher in Missouri (78%) than in New York (54%) and that those attending such schools tend to do better.

Table 2 State Scores Classified by School Location

State	All	City	Non-City
MO	229	216	233
NY	↓227↓	216	↑236↑

Table 2A: MO vs. NY.

State	All	City	Non-City
GA	220	208	222
TN	220	↑213↑	↑224↑

Table 2B: GA vs. TN

As shown in Table 2B, the state score is the same for Tennessee (TN) as for Georgia (GA). Yet when classified by school location, the state score for each subgroup is two to five points higher for Tennessee than for Georgia. Note that the percentage of students who attend non-city schools is higher in Georgia (85%) than in Tennessee (71%) and that the students who attend non-city schools tend to do better.

Table 3 shows state scores broken out by race/ethnicity. As shown in Table 3A, the state score is two points lower for Texas (TX) than for Massachusetts (MA). But when classified by race/ethnicity, the state score for each subgroup is two to 16 points higher for Texas than for Massachusetts.

Table 3 State Scores Classified by Race/Ethnicity

State	All	White	Black	Hisp.	Asian
MA	235	241	210	208	237
TX	↓233↓	↑243↑	↑220↑	↑224↑	↑247↑

Table 3A: MA vs. TX

State	All	White	Black
WV	225	226	203
LA	↓218↓	↑230↑+	↑204↑

Table 3B: WV vs. LA

How can it be that Texas students score higher than those in Massachusetts for every one of these four subgroups, yet those in Texas score lower overall? Some analysts find this puzzling and wonder if there is some error. Statisticians are well aware of this paradox. It can occur without any error in arithmetic.

The differences shown previously (zero to four points) may not seem that big. Consider a seven-point difference. As shown in Table 3B, the state score is seven points lower for Louisiana (LA) than for West Virginia (WV). Yet on each of the subgroups that were large enough[6] to give statistically reliable scores, Louisiana scores higher than West Virginia.

This kind of reversal is Simpson's Paradox: the direction of an association in the overall group is the reverse of that in each of the subgroups.[7] While statisticians know the conditions under which this reversal happens, this paradox is considered to be a fluke, an exception, an unlikely event. The purpose of this paper is to show that Simpson's Paradox is not rare in NAEP data.

The next step is to illustrate how Simpson's Paradox occurs.

[5] School locations are Central-City, Urban-Fringe and Rural. To reduce this to two categories, non-city scores were calculated based on the state score and the average score and prevalence of students at Central-City schools.
[6] NAEP does not report subgroup means for samples smaller than 60; these scores are not statistically 'reliable.'
[7] See Schield (1999) at www.StatLit.org/articles.

4. SIMPSON'S PARADOX

Simpson's Paradox involves confounding. Confounding occurs when two factors are mingled together. In a well-designed experiment with random assignment, the influence of confounding is minimized. In any observational study such as NAEP, confounding is always a concern – no matter how well designed the study. To understand how Simpson's Paradox can occur, consider the following figures.

Figure 4 illustrates the influence of family income in comparing Utah (UT) and Oklahoma (OK). The vertical axis is the NAEP score. The horizontal axis is the percentage of students who do not receive national school lunch subsidies (high income families). Students from high income families are on the right side; those from low income families are on the left. The scores are those shown in Table 1A. The scores for students from high income families (234 for OK, 233 for UT) are plotted on the right (100% high income families). The scores for students from low income families (218 for OK, 216 for UT) are plotted on the left (0% high income families).

The line connecting the two subgroup scores for a given state is a weighted average line. The state average NAEP score will lie on that line at a point determined by the percentage of families in the state who are high income: 64% in Utah and 45% in Oklahoma. The weighted average score for Utah is 227: the intersection of the vertical 64% line with the UT weighted-average line. The weighted average score for OK is 225: the intersection of the vertical 45% line with the OK weighted-average line. The state score is two points higher for Utah (227) than for Oklahoma (225).

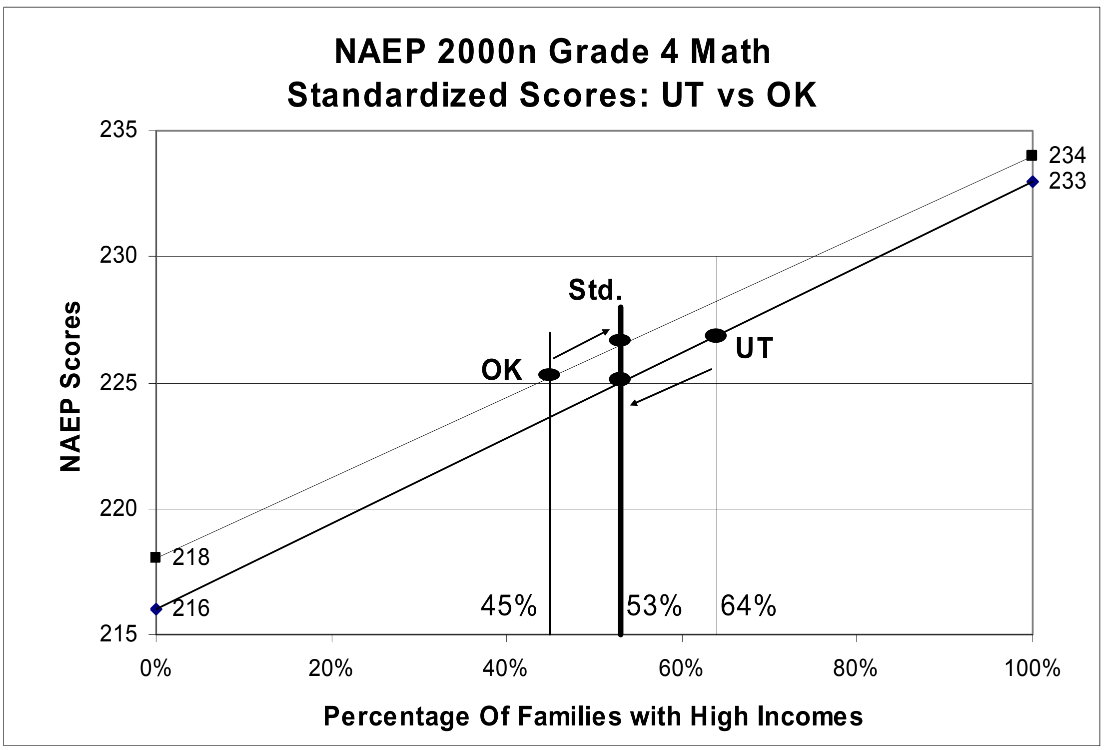

Figure 4: Simpson's Paradox: UT vs. OK

To take into account the difference in family incomes, standardized scores are calculated. Standardized scores are scores that would have been obtained if each state had the same mix of family incomes as they have collectively. If 58% of these students live in OK then 53% of all these students in both states taken collectively had high family incomes.[8] As shown in Figure 4, the standardized state score is two points

---

[8] According to NAEP in 1999 there were 447,906 public-school K-8 students in OK (328,522 in UT). So, 58% of these students are in OK. If so for grade 4, the combined percentage who are white is 53%: 58%(0.45) + 42%(0.64).

higher for Oklahoma (227) than for Utah (225). Adjusting for the influence of family income reversed the original association between state scores. Oklahoma has 'overtaken' Utah.

Figure 5 illustrates the influence of race/ethnicity in comparing Louisiana (LA) and West Virginia (WV). The generation of Figure 5 proceeds in the same manner as the generation of Figure 4. But now the horizontal axis is the percentage of students who are white. As separate groups, white students are on the right side; non-whites on the left. The scores shown in Table 3B are plotted and weighted average lines are generated. The weighted average score for West Virginia is 225 (95% are white) and for LA is 218 (53% are white). The state score is seven points higher for West Virginia (225) than for Louisiana (218).

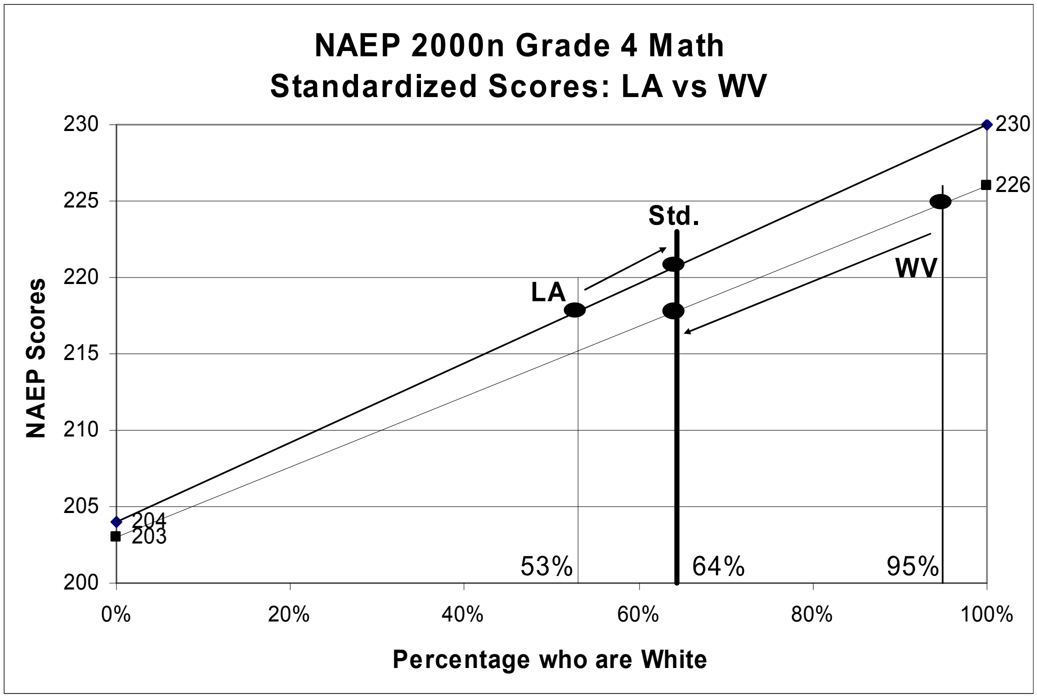

Figure 5: Simpson's Paradox: LA vs. WV

As before, standardized state scores are calculated using the confounder mixture that is found when both states are combined. If 73% of these 4th grade students are in LA, then 64% of the 4th grade students in both states are white.[9] The standardized state score is three points higher for Louisiana (221) than for West Virginia (218). Taking into account a relevant difference between two states (percentage of white) reversed the ranking between the states (LA and WV). West Virginia has been 'overtaken' by Louisiana.

For more on the nature and background of this type of graph, see Wainer (2002), Baker and Kramer (2001), Schield (2004) and Wainer (2004).

5. SUFFICIENT CONDITIONS FOR A REVERSAL OR CHANGE

Standardization is a process of generating new scores from the existing data that take into account the influence of a confounder. The graphical technique illustrated in Figure 4 and Figure 5 works well when the confounder has two values. See Schield (2004). But this technique does not work when the confounder has more than two values. Simpler sufficient conditions were used to identity a reversal or change that could handle multiple subgroups.

Reversal: a reversal of order in rank between two states after taking into account a confounder. State A has a lower score (higher rank number) than B, but after standardization A has a higher score

[9] According to NAEP in 1999 there were 558,743 public-school K-8 students in LA (205,840 in WV). So 73% of these students are in LA. If so for grade 4, the combined percentage who are white is 64%: 73%(0.53) + 27%(0.95).

(lower rank number) than B. Three conditions are jointly sufficient for a reversal. (1) The overall mean score for state A is lower than that for state B. (2) The mean score for each subgroup in State A is at or above that for the corresponding subgroup in State B. (3) The mean score for at least one subgroup in State A is above that of the corresponding subgroup in State B. These conditions define a reversal commonly referred to as Simpson's Paradox.[10]

Change: a change in rank of two states after taking into account a confounder. A change occurs when the first condition above is replaced with this: (4) The overall mean score for State A is lower than *or equal to* that for state B.[11] All reversals involve changes, but not all changes involve reversals.

Using these definitions the examples presented in Table 1 (A and B), Table 2A and Table 3 (A and B) involve Simpson's reversals. The example in Table 2B involves a non-reversing change.

6. RESULTS

The following tables summarize the number of Simpson's reversals and changes obtained when applying the aforementioned conditions to the NAEP data for various confounders. Technical details are shown in Appendix B and in the appendices listed in the following tables. 'Pairs' indicates the number of state pairs compared.[12] 'Change' and 'Reverse' indicates the number of changes and Simpson's reversals.

'Statistically Significant' indicates an initial difference in a pair of state NAEP scores that is statistically significant at the 5% level as determined by the NAEP data tool.[13] A reversal was considered statistically significant if the initial difference was statistically significant. Note that difference being tested is not the difference in standardized scores: scores generated by giving two units a standard mixture of a given confounder; the difference is that in the original scores.

		ALL			Statistically Significant			Source
Confounder	**States**	**Pairs**	**Change**	**Reverse**	**Pairs**	**Reversals**	**%**	**Appendix**
School Lunch: Two groups	39	741	23	15	504	0	0%	E
School Location: All[14]	38	703	4	1		0	0%	
Race/ethnicity: All	41	820	123	97	548	43	8%	F
Race: White vs. Non-white	39	741	62	43	504	11	2%	G

Table 4 Results for NAEP 2000n Grade 4 Math

		ALL			Statistically Significant			Source
Confounder	**States**	**Pairs**	**Change**	**Reverse**	**Pairs**	**Reverse**	**%**	**Appendix**
School Lunch: Two groups	40	780	31	19	505	1	0.2%	H
School Location: All[14]	39	741	3	3		0	0%	
Race/Ethnicity: All	40	780	137	117	505	52	10%	I
Race: White vs. Non-white	40	780	82	64	505	18	4%	J

Table 5 Results for NAEP 2002 Grade 8 Reading

[10] This definition is a slight broadening or generalization of the definition advanced in Schield and Burnham (2003) that required the score for each subgroup in State A to be *above that* for the corresponding subgroup in State B.
[11] The case where two states start with different scores (different ranks) and end up with the same score (same rank) is not analyzed. State A must have at least one subgroup that is below its match in B and at least one that is above.
[12] If there are N states, there are N(N-1)/2 state pairs.
[13] To obtain data by state, complete first three steps in footnote 18. (4) Select the 'User Options' menu and select 'Check Significant Differences.' (5) In the popup window for the NAEP Data Tool Check Selection Criteria, select the 'Average Score Scale' option and press 'Continue.' The NAEP Data Tool then returns a table indicating whether the second state score is significantly higher (>), lower (<) or equal (=) to the score of the first state. For more information about the differences, click on the '**Show Details**' button at the bottom.
[14] Vermont was excluded since it does not have enough students in the Rural and Urban Fringe categories to give results that are statistically reliable so an exact match comparison is not possible.

Note that comparisons involving race/ethnicity are approached in two ways. The white vs. non-white approach always involves comparable subgroups; the 'all' approach involves comparisons between states with some missing subgroups: subgroups that are not comparable. The latter approach has some assumptions that may be disputable whereas the former is straightforward. See Appendix A for details.

These two approaches to handling race/ethnicity give different results:

- Using a straightforward white/non-white approach, there are 64 Simpson's reversals in the NAEP 2002 data of which 18 involve initial differences that are statistically significant. In the NAEP 2000n data there are 46 Simpson's reversals of which 11 involve initial differences that are statistically significant.
- Using the more disputable 'all' approach involving non comparable subgroups, there are 117 Simpson's reversals in the NAEP 2002 data of which 52 involve initial differences that are statistically significant. In the NAEP 2000n data there are 97 Simpson's reversals of which 43 involve initial differences that are statistically significant.

Either way, these results support the claim that Simpson's reversals are not rare in NAEP data.

In the NAEP 2002 Grade 8 Reading data, 505 of the differences in state scores are statistically significant according to the NAEP Data Tool. Of these, 18 (4%) are reversed using the white-nonwhite approach while 52 (10%) are reversed using the more liberal 'all' approach. Up to 10% of the statistically significant differences in the 2002 state NAEP Grade 8 Reading scores are reversed by taking into account race. This high prevalence of Simpson's Paradox among statistically significant differences in state scores indicates that all comparisons of state scores must be treated with caution – even those involving differences that are statistically significant. Statistical significance does not eliminate confounding.

7. 'JOURNALISTIC SIGNIFICANCE'

NAEP is very careful in noting only those differences that have statistical significance. This care is shown in the online warning, "*NOTE: Observed differences are not necessarily statistically significant.*" This care is also shown in only making comparisons (cross-sectional or longitudinal) that are statistically significant. But Education Departments and journalists may not be as careful. News stories involving NAEP data often compare state scores with previous scores and with those of others states regardless of their statistical significance.[15] This may be done using point differences or rank differences. Since these comparisons are being reported they obviously have 'journalistic significance' – perhaps because federal controls and monies give these comparisons political significance. Given that these kinds of changes and reversals are newsworthy regardless of their statistical significance, all the changes obtained (at least 150 per data set) are said to have 'journalistic significance.'

8. PRESENTING CONFOUNDING

Simpson's Paradox reversals are an extreme form of confounding. There are many instances of confounding that do not involve Simpson's Paradox. One way of presenting confounding without adjusting the data is to use rank-based measures. We recognize there are problems in using ranks. NAEP does not currently publish ranks with their online data tool. Ranks can obscure big difference and magnify small ones. Talking about a change in rank as an increase is confusing since a higher rank has a lower number. Graphing ranks requires that changes in ranks may have journalistic significance without being statistically significant, much less statistically important. The following are three specific pitfalls associated with using ranks:

(1) Equal Scores Problem. When multiple states have the same NAEP score, they have the same rank. To illustrate this point, suppose that of the 48 states, 46 had the same score, with one state higher and one lower. Any of the 46 states in the middle could truthfully say, there is only one state that did better. Yet, it could truthfully be said that each of these 46 states finished 2nd to last. A related problem

[15] A convenience survey of five press releases by state education departments found that all of them included comparisons involving differences or changes that were not statistically significant.

is what rank should be assigned when states have equal scores. Same-ranked states get the highest rank possible in Excel but get the average rank in SPSS. In the 46-states tied example, Excel gives one 1st place, 46 in 2nd place and one in 48th place whereas SPSS gives one 1st place, 46 in 24.5th place and one in 48th place. However, it is awkward to speak of a state being in 24.5th place.

A simple solution to this equal-scores problem is to obtain greater precision. The results of using two digits after the decimal are shown in Appendix C and Appendix D. Although the increased precision certainly leads to 'hairsplitting', it does decrease other equally contentious problems.

(2) Different States Problem. Comparing the ranks of a state for two different sub groups is not meaningful if any state does not have data for both subgroups. This happens often with racial/ethnic subgroups as shown in Appendix D. The rankings of states for each racial/ethnic subgroup are not comparable because the number of jurisdictions for specific subgroups varies from 41 for white students to 13 for Asian students. Obviously it is misleading to say that Utah ranks higher among Asians (13th) than among whites (27th) since there are only 13 states with a statistically reliable prevalence of Asians. Utah is last among Asians.

A simple solution for this 'different numbers of states' problem is to use percentiles. Appendix C presents ranks and percentiles as generated from scores using Excel for white vs. non-white students. Percentiles for state scores by race are shown in Appendix D.

(3) Rank Explanation Problem. Explaining a state's overall rank in terms of the state's rank for each subgroup can be misleading. This is shown in Appendix C. Note that Iowa ranks 5th overall, 14th among white students and 5th among non-whites. It is tempting to say that Iowa is 'pulled up' from the 14th place rank of its white students to 5th place overall by the 5th place rank of the non-white students. Similarly Vermont ranks 9th overall, 20th among white students and 6th among non-whites. It is tempting to say that Vermont is 'pulled up' from 20th place among whites to 9th place overall by the 6th place rank of the non-white students.

While this 'pulled-up' style of explanation is totally appropriate for a weighted average, it is not appropriate for ranks or percentiles. Unlike weighted averages, the rank of the overall group is not algebraically determined by the ranks and proportions of the subgroups. It is difficult to explain New York's 21st place overall rank as being due to its 8th place rank among white and its 13th place rank among non-whites. It is difficult to explain Minnesota's 1st place rank overall as being due to its 7th place rank among whites and its 3rd place rank among non-whites. Since percentiles are just scaled ranks, the use of 'weighted-average' style explanations is inappropriate there also.

A solution for this 'rank-explanation problem' is to avoid such explanation for ranks or percentiles. .

In summary, while ranks and percentiles can be misleading and may reflect differences that are little more than 'splitting hairs', they can call attention to differences between subgroups and the overall group that would otherwise involve adjusting data for the influence of the confounder. So long as one doesn't try to explain the overall percentile by the percentiles of the subgroups, then comparing the overall percentile with a subgroup percentile can be very useful in demonstrating the influence of a confounder. Figure 6 shows 2000n NAEP Grade 4 Math percentiles by state overall and for just the white students. A higher percentile indicates a higher score.

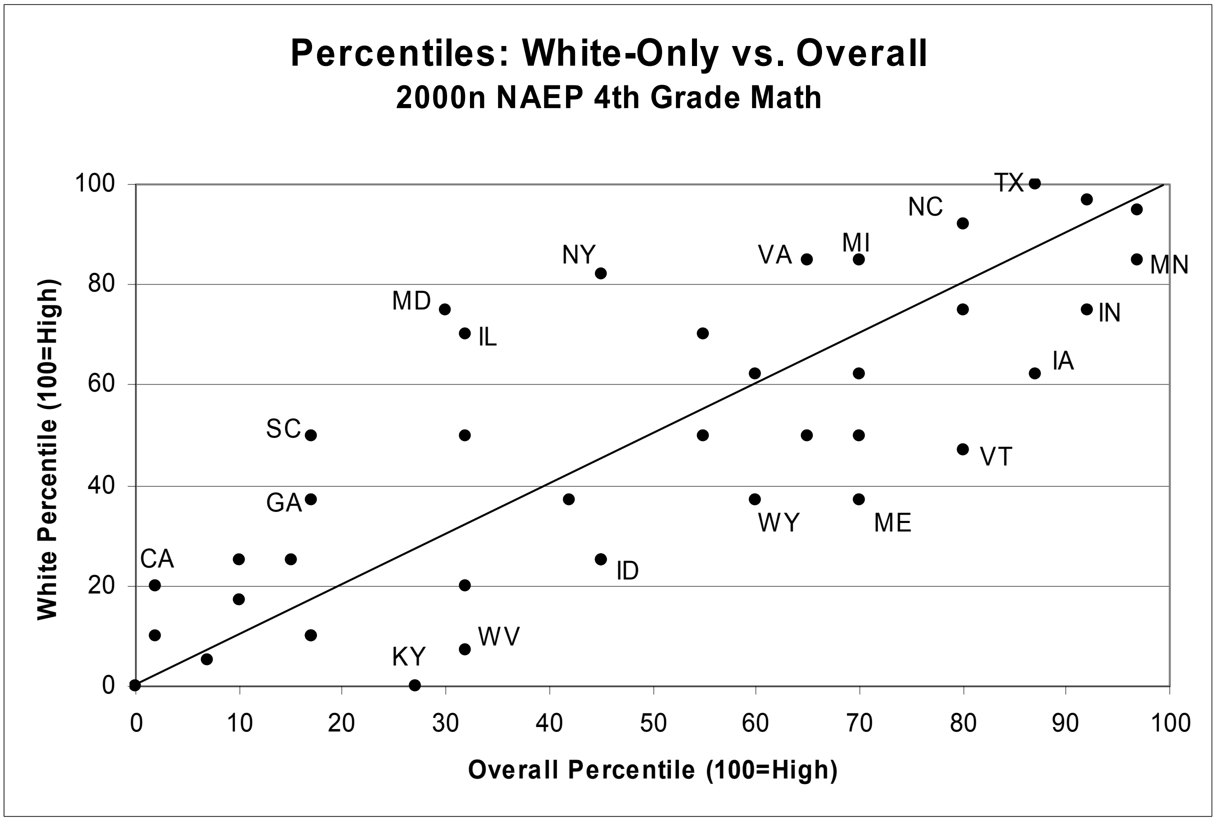

Figure 6: State Percentiles: White-only versus Overall

States below the line have percentiles that are lower for their white students than for their overall mix of students (e.g., Indiana, Iowa, Vermont and Maine). States above the line have percentiles that are higher for their white students than for their overall mix of students (e.g., Texas, Virginia and New York). This kind of comparison of percentiles between overall and white-only can readily indicate the influence of the non-white students. Appendix C shows the state scores, ranks and percentiles for whites and non-whites. Appendix D shows state ranks and percentiles by racial/ethnic group.

9. NAGB POLICY ON ADJUSTING DATA

Adjusting data for confounders is controversial. In 1994, the National Assessment Governing Board (NAGB) reviewed this matter. The Board noted that "*one of the methods being considered would provide for reporting across-state comparisons in an 'adjusted' or 'predicted' form based on ethnic and other demographic characteristics or on 'opportunity to learn' variables such as instructional approaches and time on task*." The board then reaffirmed its 1989 policy which stated that "*no levels of predicted or adjusted performance will be presented by NAEP for individual states*." The Board notes that "*any adjusted or predicted scores would be subject to serious methodological and political challenge and would be contrary to the strong national commitment to encouraging high standards for all children*." NAGB (1994). But there are political implications in not adjusting. It may be counterproductive to hold schools and states accountable for things not under their control. One way to handle this is to present adjusted scores as an adjunct to the actual scores. This would allow methodological and political issues to be discussed after data have been adjusted using techniques that are methodologically sound.

10. RACE

This preliminary investigation indicates that Simpson's reversals are more common when adjusting for race/ethnicity than when adjusting for family income or school location. Adjusting state scores for differences in race/ethnicity seems more contentious than adjusting for family income or school location. But adjusting for race does not imply that there are genetic differences between racial/ethnic groups or that these score differences are caused by genetic differences. Racial/ethnic differences may be related to differences in a host of factors (socio-economic status, parental education, reading materials in the home, culture, etc.) some of which may not be readily measured. The importance of analyzing education outcomes by race is shown in the requirement of the 'No Child Left Behind' Act of 2001 that data be disaggregated. As Mukhopadhyay and Henze (2003) note: "Without data broken out according to racial, gender, and ethnic categories, schools would not be able to assess the positive impact intervention programs have on different groups of students." Once data are obtained and presented for various subgroups, the need to take such factors into account becomes more obvious.

11. RECOMMENDATIONS

The general recommendation is to emphasize that state means are influenced by a variety of potential confounders. These confounders can influence the overall scores and ranks of states. Four specific actions are recommended.

(1) NAEP should investigate ways to facilitate confounder-related analysis. For example, NAEP could also publish state scores as percentiles within various subgroups to facilitate comparisons. The score data by subgroups are already published by NAEP. It is just a matter of generating percentiles when appropriate. Presenting percentiles of states within subgroups avoids methodological issues involved in standardizing (Section 4) or adjusting (Section 9) and facilitates seeing the influence of confounding (Section 8).

(2) NAEP should generate adjusted scores for states after taking into account the influence of factors other than race/ethnicity that are outside the school's control such as school location and socio-economic factors. Doing so will allow a discussion of many methodological issues without taking on race.

(3) NAEP should generate adjusted scores for states after taking into account the influence of student race/ethnicity. This will allow for a discussion of those issues unique to race and ethnicity.

(4) NAEP should increase the sample sizes so that a 'journalistically significant' difference of one or two points would have statistical significance. Implementing this recommendation will be expensive but it allows smaller differences to be meaningfully distinguished and this may have political benefits.

12. CONCLUSIONS

This study finds that Simpson's Paradox is not a rare event in comparing state NAEP scores. When comparing the scores of some 40 states on three confounders, at least 150 changes were identified per dataset with at least 110 of these involving Simpson's Paradox reversals. In the NAEP 2002 Grade 8 Reading data, 55 Simpson's reversals involve initial differences that are statistically significant. These Simpson's reversals are 10% of all statistically-significant difference in state scores. Statistical significance does not eliminate Simpson's Paradox reversals. All Simpson's reversals – whether statistically significant or not – are newsworthy, which gives them 'journalistic significance.' It is recommended that state scores and ranks be compared for key subgroups, that adjusted scores be calculated for factors outside the schools control, and that these adjusted state scores be employed as an adjunct when reporting results.

13. REFERENCES

Baker, S. G. & Kramer, B. S. (2001). "Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies." *Journal of Women's health and gender-based medicine*, 10, 867-872.

Mukhopadhyay, C. and Henze, R. "How Real is Race?" *Phi Delta Kappan*, May, 2003, p. 669-678.

NAGB (1994). "Resolution on Reporting State-Level NAEP Results." Photocopy of Resolution released by the National Assessment Governing Board (NAGB).

Schield, Milo (1999). "Simpson's Paradox and Cornfield's Conditions." *1999 ASA Proceedings of the Section on Statistical Education*, p. 106-111.[16]

Schield, Milo (2004). "Three Graphs For Statistical Literacy." International Conference on Mathematics Education (ICME-10).[16]

Schield, Milo and Thomas Burnham (2003). "Confounder-induced Spuriousity and Reversal: Algebraic Conditions Using a Non-Interactive Model for Binary Data." *2003 ASA Proceedings of the Section on Statistical Education*, p 3690- 3697. [16]

Wainer, Howard (2002). "The BK-Plot: Making Simpson's Paradox Clear to the Masses." *Chance Magazine* Vol. 15, No. 3, Summer 2002, pp. 60–62.

Wainer, Howard (2004). "Three Paradoxes in the Interpretation of Group Differences." Draft of a paper submitted to *The American Statistician*.[16]

Notes: This research was supported in part by a grant from the W. M. Keck Foundation to Augsburg College "*to support the development of statistical literacy as an interdisciplinary discipline*." This paper has not been reviewed or approved by the Minnesota State Department of Education, the U.S. Department of Education, the National Assessment of Educational Progress (NAEP) or any subunits thereof. This paper was presented at the annual meeting of the American Education Research Association (AERA) in San Diego on 16 April 2004.

Contacts: James Terwilliger is the NAEP Coordinator for Minnesota. His e-mail address is Jim.Terwilliger@state.mn.us. Milo. Schield is Director of the W. M. Keck Statistical Literacy project and Professor of Business Administration at Augsburg College. His e-mail address is Schield@Augsburg.edu. This paper is at www.augsburg.edu/ppages/~schield and at www.StatLit.org.

[16] For a copy of this paper, see 'Teaching Stat Lit' at www.StatLit.org.

Appendix A. MISSING DATA (LACK OF COMMON SUBGROUPS)

Missing data is a problem. In NAEP, the problem is that the amount of data obtained was too small to give statistically-reliable scores. Scores were not considered reliable if the associated subgroup had less than 60 subjects in the sample. Most of the instances involve race/ethnicity. This issue was addressed in two ways. The first involves grouping black, Hispanic and Asian into a non-white category so that all states had a non-white group that surpassed the NAEP minimum size requirement.[17] The second involves imputing scores for the missing subgroups to allow comparisons of all state scores. The second gives bigger numbers than the first, but it is a more disputable approach.

If two jurisdictions do not have common subgroups, it seems there is little that can be said. But if the jurisdiction having the lower NAEP score has subgroup scores that are higher or at least as high for each of the common subgroups, this gives some evidence for concluding that the scores in the missing subgroups would be at least as high as those in the jurisdiction having the higher NAEP score. And as long as at least one of the scores in the lower state is greater than that in the common subgroup in the higher state, then we have satisfied the sufficient condition mentioned earlier. Now this argument from near-ignorance is not very strong. A second argument is that the missing subgroups must be quite small as a percentage of students in that state. It is possible to have a reversal even if one of the subgroups in the lower state has a *lower* score than that of the common subgroup in the higher state. For these two reasons, the analysis of all four racial/ethnic groups assumes that if a reversal is justified by the common subgroups, it would not be contradicted by anything involving the missing groups. The extreme case involves Maine and Vermont (97% white) being compared against Mississippi (49%), Texas (44%), New Mexico (38%) and California (38%). Obviously this conclusion is very disputable. The reason for presenting this criterion is not to argue that it is true, but to argue that it a reasonable approach to handling the existing racial/ethnic differences between states.

Appendix B. TECHNICAL DETAILS

This appendix indicates how the data and formula were entered into the spreadsheets shown in subsequent appendices. First, obtain the desired NAEP data by jurisdiction/state and by subgroup from the web.[18] Copy the data to a spreadsheet and convert the state names to their two-character equivalents.[19] The ordering of the states is critical. To locate all the reversals and changes in the lower-triangle (e.g., Appendix E), the sort order must be state average (descending) and the NAEP scores of all state subgroups must be ascending.[20] Since Excel handles a maximum of three sort groups, in the case of race (Appendix F) this means sorting first on Black (A), Hispanic (A) and Asian (A), and then sorting on State Average (Descending) and White (Ascending). After the data has been appropriately ordered in a column format, it must be transposed into a row format as shown on the bottom of the table in Appendix E.[21] Thus, states on the left are trying to overtake; states along the bottom are being overtaken.

---

[17] NAEP guidelines require that a subgroup have at least 60 students to be shown. If 3,000 students are tested in a state, then data for subgroups involving less than 2% (60/3000) of the students will not be shown.

[18] Go to http://nces.ed.gov/nationsreportcard/naepdata/search.asp. (1) Select a *subject* (Math), *grade* (4th), *jurisdiction* (National/public) and *category* (Major Reporting Groups); press '*Continue*.' (2) Check the year desired and select *Major Reporting Group* (1. All Students). This gives the national score. (3) From the *User Options* menu, select *Add/Deletee Jurisdictions* to obtain the score for each state. From the list of Jurisdictions, press *Select All* and then de-select those not desired (American Samoa, District of Colombia, Guam, Hawaii, Virgin Islands, etc.). Press the *Accept Changes* button. This gives NAEP scores by state. (4) From the *User Options* menu, select *Add/Delete Subgroups*. Select the subgroup desired (Race as Identified by School Records); Press *Accept Changes.*

[19] Recall that Arkansas is AR (not AK) and Arizona is AZ (not AR).

[20] A different sort order could change the location of results and of statistical significance (e.g., place some in the upper-right of the table), but a different sort order would not change the number of reversals or changes. The formula is copied throughout the entire table so two states are compared twice: once in lower left, once in upper right.

[21] To transpose data from columns to rows, copy the column data to the Clipboard and then place the cursor in the upper-left hand cell of the area being copied to. From the Edit menu, select the Paste-Special option. Check the Transpose box located near the bottom and press OK. (If formulas are involved, check the Values checkbox).

The conditions for a change were entered as a spreadsheet formula into the upper-left hand cell. Consider a formula involving two subgroups where there is no missing data (e.g., Appendix E).

```
G8: =IF(AND($B8<=G$48,$D8>=G$50,$E8>=G$51,OR($D8>G$50,$E8>G$51))=TRUE,G$48-$B8,"")
```

If the condition is true, then the difference in state scores (G48-B8) is shown; otherwise a blank ("") is shown. The condition for two subgroups with no missing data involves an AND of several conditions. First, the state score of the state on the left must be less than or equal to that of the state score below (B8<=G48). Second, for each of the two subgroups the subgroup score in the state on the left must be greater than or equal to the score of that subgroup in the state on the bottom (D8>=G50, E8>=G51). Third, at least one of the subgroup scores for the state on the left (those overtaking) must be greater than that of the state on the bottom (those overtaken): OR(D8>G50,E8>G51). The dollar signs are added to keep certain rows and columns fixed to facilitate copying the resulting formula to all cells in the table. If there are more sub-groups, additional items must be added. If there are missing values, then the formula becomes more complex. See the formula in the bottom line for Appendix F.

If the conditions for a change were satisfied, the cell showed the size difference between the two state scores. A value of zero indicates a non-reversing change. Any value greater than zero indicates a Simpson's Paradox reversal.[22] The counts of states were obtained using the CountA function in combination with the CountBlank function. The maximum difference in state scores was obtained using the MAX function. Care should be taken in how one describes these changes.[23]

Statistical significance in the conditional formatting (the shaded areas) was estimated if the 95% confidence intervals failed to overlap. The width of a 95% confidence interval was calculated based on the national mean of the standard errors for the states studied.[24] Since the sample sizes were similar for all the states, the differences in standard errors largely reflected differences in the standard deviations.[25]

In the body of this paper, statistical significance was always determined using the exact approach using the NAEP data tool. This explains the difference between Appendix H (507 statistically significant differences of which 55 are reversed for an 11% rate) versus Table 5 (505 statistically significant differences of which 52 are reversed for a 10% rate). In the spreadsheets in the following appendices, those differences that were not statistically significant using the NAEP Data Tool but were statistically significant using the short-cut formula were italicized. Thus, in Appendix H, note the italicized value for OH-VT, NY-OH, and LA-TN. These three values are the difference between the 55 shown in the spreadsheet and the 52 shown in Table 5.

---

[22] If space is a problem, the top row of state data can be eliminated (it is impossible for the top state to pass anyone higher) and the right column of state data can be eliminated (it is impossible for any state to pass the bottom state).

[23] It is important to distinguish changing the data from calculating a new score based on a combination of existing data and hypothetical weights. Statisticians (almost) never change the data. Avoid saying, "The data was changed to give both states the same mix of students," even though that may be readily understood. Instead one should say, "Scores were calculated or constructed using the same mix for both states." Speaking about changes in scores (raw vs. calculated) may be technically correct, but can lead the unwary to conclude the data have been changed.

[24] In the NAEP 2000n Grade 4 math, the standard errors ranged from a minimum of 0.7 to a maximum of 1.9 with a mean of 1.3. Multiplying this by 1.96 and doubling it to get the full width gave a range of 5.096 which rounded to five. In the NAEP 2002 Grade 8 reading, the standard errors ranged from a minimum of 0.5 to a maximum of 1.8 with a mean of 1.13. Multiplying this by 1.96 and doubling it gave a range of 4.43 which rounded down to four. This approach has two weaknesses. (1) The range of standard errors is wide compared to the mean value. (2) This approach uses the same standard error for all states when in any given comparison of two states we need the unique standard error for just those two states. These weaknesses are somewhat mitigated since the goal is to indicate the general range where differences are statistically significant rather than to make precise measurements.

[25] For the NAEP 2000n Grade 4 math data, the associated width was about 5; for the NAEP 2002 Grade 8 reading data, the associated width was about 4. One reason for this decrease was that in 2000 the state samples were split between non-accommodations (2000n) and accommodations (2000) whereas in 2002 there was no split. Thus the relevant sample sizes in 2002 were about double those in 2000.

Appendix C. NAEP 2000n Grade 4 Math Ranks and Percentiles for Whites/Non-Whites

STATE	State Mean	White Mean	ST	White Pct	NonWhite Mean	Rank All	Rank White	Rank NonWhite	Percentile All	Percentile White
Minnesota	235.27	238.50	MN	82	220.56	1	7	3	100	84
Massachusetts	234.96	240.50	MA	78	215.32	2	3	10	97	94
Indiana	234.42	236.89	IN	88	216.31	3	10	7	94	76
Connecticut	234.24	242.11	CT	72	214.00	4	2	12	92	97
Iowa	232.90	234.66	IA	90	217.06	5	14	5	89	65
Texas	232.67	242.88	TX	44	224.65	6	1	1	86	100
N. Carolina	232.46	240.48	NC	62	219.37	7	4	4	84	92
Kansas	231.95	237.19	KS	79	212.24	8	9	16	81	78
Vermont	231.70	232.17	VT	97	216.50	9	20	6	78	50
North Dakota	230.89	232.75	ND	91	212.08	10	19	17	73	52
Michigan	230.89	238.53	MI	77	205.31	10	6	28	73	86
Maine	230.57	230.78	ME	97	223.78	12	23	2	68	42
Ohio	230.57	235.41	OH	80	211.21	12	13	19	68	68
Virginia	230.39	238.90	VA	63	215.90	14	5	9	65	89
Montana	229.81	232.81	MT	86	211.38	15	18	18	63	55
Wyoming	229.25	231.27	WY	89	212.91	16	21	15	60	47
Missouri	228.55	234.50	MO	79	206.17	17	15	26	57	63
Utah	227.29	230.39	UT	86	208.25	18	25	23	55	36
Idaho	226.89	229.51	ID	84	213.14	19	27	14	52	26
Oregon	226.63	229.51	OR	81	214.35	20	27	11	50	26
New York	226.56	238.49	NY	52	213.64	21	8	13	47	81
Nebraska	225.95	230.53	NE	83	203.59	22	24	34	44	39
Oklahoma	225.04	229.49	OK	67	216.01	23	30	8	42	23
Illinois	224.93	235.69	IL	57	210.67	24	12	21	39	71
West Virginia	224.85	225.79	WV	94	210.12	25	36	22	36	7
Rhode Island	224.63	232.96	RI	75	199.64	26	17	38	34	57
Maryland	222.31	236.84	MD	52	206.57	27	11	25	31	73
Kentucky	220.99	224.17	KY	87	199.71	28	38	36	28	2
S. Carolina	220.42	233.43	SC	56	203.86	29	16	32	26	60
Nevada	220.27	226.63	NV	60	210.73	30	33	20	23	15
Tennessee	219.84	226.56	TN	74	200.71	31	34	35	21	13
Georgia	219.56	231.16	GA	52	206.99	32	22	24	18	44
Arizona	218.77	230.19	AZ	56	204.24	33	26	30	15	34
Louisiana	217.96	229.51	LA	53	204.94	34	27	29	13	26
Alabama	217.94	228.18	AL	58	203.80	35	32	33	10	18
Arkansas	217.06	224.52	AR	70	199.65	36	37	37	7	5
New Mexico	213.87	226.51	NM	38	206.12	37	35	27	5	10
California	213.57	228.90	CA	38	204.17	38	31	31	2	21
Mississippi	210.97	223.66	MS	49	198.78	39	39	39	0	0

To obtain state scores with two-decimal accuracy, select 'Customize Table' under 'User Options.' Change 'Degree of Precision' from 'zero digits' to 'two digits.' Obtaining scores with two-digits gives much higher accuracy for non-white means in states having a very small non-white population (e.g., Vermont and Maine). It also eliminates most ties in state rankings but that is incidental.

To obtain percentiles in Microsoft Excel, use = 100*PercentRank(Array, Test, Significance) where Array is the complete list of scores, Test is the score of the state being tested and Significance is the digits of precision). If the values are non-matching, the PercentRank function returns 0% and 100% for two values, 0%, 50% and 100% for three values, 0%, 33%, 67% and 100% for four values, and 0%, 25%, 50%, 75% and 100% for five values. States having matching scores get assigned the highest rank but the lowest percentile possible in Excel.

The rounded values of these non-white scores were used for Vermont (97%, 217), Maine (97%, 224), West Virginia (94%, 210), North Dakota (91%, 212) and Iowa (90%, 217) in Appendix G.

Appendix D. State NAEP 2000n Grade 4 Math Ranks and Percentiles by Race/Ethnicity

	RANKS					PERCENTILES				
State	All	White	Black	Hispanic	Asian	All	White	Black	Hispanic	Asian
MN	1	7	11		8	100	85	67		41
MA	2	3	10	13	5	97	95	70	42	66
IN	3	10	7			95	77	80		
CT	4	2	8	11	3	92	97	77	52	83
IA	5	15	4			90	65	90		
TX	6	1	1	2	1	87	100	100	95	100
NC	7	4	2			85	92	96		
KS	8	9	18	6		82	80	45	76	
VT	9	22				80	47			
MI	10	6	26			75	87	19		
ND	10	21				75	50			
ME	12	25				70	40			
OH	12	14	12			70	67	64		
VA	14	5	9	1	2	67	90	74	100	91
MT	15	20				65	52			
WY	16	23		10		62	45		57	
MO	17	16	24			60	62	25		
DESS	18	12	3	4		57	72	93	85	
ODDS	19	18	5	3	9	55	57	87	90	33
UT	20	27		19	13	52	35		14	0
ID	21	29		15		50	25		33	
OR	22	29		17	6	47	25		23	58
NY	23	8	6	12	4	45	82	83	47	75
NE	24	26	31	20		42	37	3	9	
OK	25	32	16	7		40	22	51	71	
IL	26	13	21	8		37	70	35	66	
WV	27	38	19			35	7	41		
RI	28	19	25	22		32	55	22	0	
MD	29	11	23	5	7	30	75	29	80	50
KY	30	40	27			27	2	16		
SC	31	17	22			25	60	32		
NV	32	35	13	16	11	22	15	61	28	16
TN	33	36	28			20	12	12		
GA	34	24	15	9		17	42	54	61	
AZ	35	28	14	18	10	15	32	58	19	25
LA	36	29	17			12	25	48		
AL	37	34	20			10	17	38		
AR	38	39	30			7	5	6		
NM	39	37		14		5	10		38	
CA	40	33	32	21	12	2	20	0	4	8
MS	41	41	29			0	0	9		

Appendix E. NAEP 2000n Grade 4 Math: Family Income Changes

ADJUST FOR FAMILY INCOME 741 **Pairs** **11 States Overtaking** **NAEP 2000n Grade 4 Math**
Based on National School Lunch Program **23 Changes** **13 States Overtaken** **Number shown is the difference in State scores**
Two Subgroups: Elgible vs. Non **15 Reverals** **0 Size > 4** **Shading indicates difference is more than 4**
Scores for Non-eligible are inferred

Sort: Ave(D),Pay(A), No(A)						1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	3	2	0	0	1	0	0	0	0	3	0	3	4	1	0	1	1	0	0	0	0	**13**	11	Max
ST	Ave	%P	Pay	No	DP	MN	MA	IN	CT	IA	TX	VT	KS	NC	ME	ND	OH	MI	MT	VA	WY	MO	UT	OR	ID	NY	NE	WV	OK	IL	RI	MD	KY	NV	TN	GA	SC	AZ	AL	LA	AR	NM	CA	ST	#	Size
MN	235	68	240	224	16																																							MN	0	
MA	235	67	243	219	24																																							MA	0	
IN	234	65	240	223	17																																							IN	0	
CT	234	67	242	218	24																																							CT	0	
IA	233	69	236	226	10																																							IA	0	
TX	233	48	242	225	17	**2**		**1**	**1**																																			TX	3	2
VT	232	66	237	222	15																																							VT	0	
KS	232	62	241	217	24																																							KS	0	
NC	232	55	241	221	20								**0**																															NC	1	0
ME	231	64	234	226	8																																							ME	0	
ND	231	58	235	225	10																																							ND	0	
OH	231	57	239	220	19																																							OH	0	
MI	231	68	240	212	28																																							MI	0	
MT	230	53	236	223	13																																							MT	0	
VA	230	61	237	219	18																																							VA	0	
WY	229	60	234	222	13																																							WY	0	
MO	229	62	237	216	21																																							MO	0	
UT	227	64	233	216	17																																							UT	0	
OR	227	58	234	217	17																		**0**																					OR	1	0
ID	227	52	234	219	15																		**0**	**0**																				ID	2	0
NY	227	48	239	216	23																																							NY	0	
NE	226	61	235	212	23																																							NE	0	
WV	225	49	232	218	14																																							WV	0	
OK	225	45	234	218	16																		**2**	**2**																				OK	2	2
IL	225	52	235	214	21																						**1**																	IL	1	1
RI	225	60	236	209	28																																							RI	0	
MD	222	58	233	207	26																																							MD	0	
KY	221	48	231	212	19																																							KY	0	
NV	220	60	228	208	20																																							NV	0	
TN	220	57	231	205	26																																							TN	0	
GA	220	45	233	209	24																											**2**		**0**	**0**									GA	3	2
SC	220	46	235	207	28																											**2**			**0**									SC	2	2
AZ	219	49	231	207	24																														**1**									AZ	1	1
AL	218	44	230	209	21																													**2**										AL	1	2
LA	218	32	233	211	22																											**4**		**2**	**2**	**2**		**1**	**0**					LA	6	4
AR	217	47	229	206	23																																							AR	0	
NM	214	34	227	207	20																																							NM	0	
CA	214	40	229	204	25																																							CA	0	
MS	211	32	226	204	22																																							MS	0	
39					STATE	MN	MA	IN	CT	IA	TX	VT	KS	NC	ME	ND	OH	MI	MT	VA	WY	MO	UT	OR	ID	NY	NE	WV	OK	IL	RI	MD	KY	NV	TN	GA	SC	AZ	AL	LA	AR	NM	CA	39	**23**	4
					Ave	235	235	234	234	233	233	232	232	232	231	231	231	231	230	230	229	229	227	227	227	227	226	225	225	225	225	222	221	220	220	220	220	219	218	218	217	214	214			
					DS	16	24	17	24	9.7	17	15	24	20	8.3	9.5	19	28	13	18	13	21	17	17	15	23	23	14	16	21	28	26	19	20	26	24	28	24	21	22	23	20	25			
					Pay	240	243	240	242	236	242	237	241	241	234	235	239	240	236	237	234	237	233	234	234	239	235	232	234	235	236	233	231	228	231	233	235	231	230	233	229	227	229			
					No	224	219	223	218	226	225	222	217	221	226	225	220	212	223	219	222	216	216	217	219	216	212	218	218	214	209	207	212	208	205	209	207	207	209	211	206	207	204			
					%Pay	68	67	65	67	69	48	66	62	55	64	58	57	68	53	61	60	62	64	58	52	48	61	49	45	52	60	58	48	60	57	45	46	49	44	32	47	34	40			

G8 '=IF(AND($B8<=G$48,$D8>=G$50,$E8>=G$51,OR($D8>G$50,$E8>G$51))=TRUE,G$48-$B8,"")

Appendix F. NAEP 2000n Grade 4 Math: Race/Ethnicity Changes (All four subgroups)

ADJUST FOR RACE/ETHNICITY **820 Pairs** **123 Changes** **NAEP 2000n Grade 4 Math** **Number shown is difference in State Scores**

W=White,B=Black,H=Hisp.,A=Asian **97 Reversals** **39 Size > 4** **27 States are Overtaking** **Shading indicates difference in scores is > 4**

43 Statistically significant **24 States are Overtaken** **Bold indicates statsitical significance using NAEP data tool**

Sort: Ave(D),W(A),B(A),H(A),A(A)

						MN	MA	IN	CT	IA	TX	VT	KS	NC	ME	ND	OH	MI	MT	VA	WY	MO	ODDS	DESS	UT	OR	ID	NY	NE	WV	OK	RI	IL	MD	KY	TN	NV	GA	SC	AZ	AL	LA	AR	NM	CA	24	#	Max
Ave	W	B	H	A	ST	5	2	5	1	3	0	15	2	0	13	8	3	1	6	0	6	4	0	0	8	5	5	0	4	9	0	3	0	0	7	6	0	0	0	0	1	0	1	0	0	123	27	Size
235	239	209		232	MN																																									MN	0	
235	241	210	208	237	MA	0																																								MA	1	0
234	237	211			IN																																									IN	0	
234	242	211	210	242	CT	1	1	0																																						CT	3	1
233	235	213			IA																																									IA	0	
233	243	220	224	247	TX	2	2	1	1	0																																				TX	5	2
232	232				VT																																									VT	0	
232	237	204	215		KS							0																																		KS	1	0
232	240	217			NC	3		2		1		0	0																																	NC	5	3
231	231				ME																																									ME	0	
231	233				ND							1			0																															ND	2	1
231	235	207			OH							1			0	0																														OH	3	1
231	239	199			MI							1			0	0																														MI	3	1
230	233				MT							2			1																															MT	2	2
230	239	211	225	244	VA	**5**		**4**				2	2		1	1	1	1	0																											VA	9	5
229	231		212		WY																																									WY	0	
229	235	201			MO							3			2	2			1		0																									MO	5	3
228	233	212	224	232	Odds							**4**			3						1																									Odds	3	4
228	236	216	221		Dess					**5**		4			3	3	3		2		1	1																								Dess	8	5
227	230		204	217	UT																																									UT	0	
227	230		207	237	OR																				0																					OR	1	0
227	230		208		ID																				0	0																				ID	2	0
227	238	212	209	242	NY			**7**				**5**			**4**	**4**	**4**		3			2			0	0	0																			NY	10	7
226	231	196	203		NE																																									NE	0	
225	226	203			WV																																									WV	0	
225	229	205	215		OK																									0																OK	1	0
225	233	200	194		RI							**7**			**6**																															RI	2	7
225	236	203	215		IL							**7**			**6**	**6**			5		4	4			2	2	2		1	0		0														IL	12	7
222	237	202	216	234	MD							**10**			**9**	**9**			**8**		**7**	**7**			**5**		**5**		4			3														MD	10	10
221	224	199			KY																																									KY	0	
220	227	198			TN																																									TN	0	
220	227	207	208	225	NV																									**5**					1	0										NV	3	5
220	231	205	212		GA																				**7**	**7**	**7**		**6**	**5**					1	0										GA	7	7
220	233	203			SC							**12**			**11**						**9**				**7**	**7**	**7**		**6**	**5**		**5**			1	0										SC	11	12
219	230	207	205	231	AZ																				**8**					**6**					2	1										AZ	4	8
218	228	203			AL																									**7**					3	2										AL	3	7
218	230	204			LA																									**7**					3	2					0					LA	4	7
217	225	197			AR																																									AR	0	
214	227		208		NM																									**11**					**7**								3			NM	3	11
214	229	191	200	225	CA																																									CA	0	
211	224	198			MS																																									MS	0	
41					ST	MN	MA	IN	CT	IA	TX	VT	KS	NC	ME	ND	OH	MI	MT	VA	WY	MO	DDS	DESS	UT	OR	ID	NY	NE	WV	OK	RI	IL	MD	KY	TN	NV	GA	SC	AZ	AL	LA	AR	NM	CA		27	12
					Ave	235	235	234	234	233	233	232	232	232	231	231	231	231	230	230	229	229	228	228	227	227	227	227	226	225	225	225	225	222	221	220	220	220	220	219	218	218	217	214	214			
					W	239	241	237	242	235	243	232	237	240	231	233	235	239	233	239	231	235	233	236	230	230	230	238	231	226	229	233	236	237	224	227	227	231	233	230	228	230	225	227	229			
					B	209	210	211	211	213	220		204	217			207	199		211		201	212	216				212	196	203	205	200	203	202	199	198	207	205	203	207	203	204	197		191			
					H		208		210		224		215							225	212		224	221	204	207	208	209	203		215	194	215	216			208	212		205				208	200			
					A	232	237		242		247									244			232		217	237		242						234			225			231					225			

G7: '=IF(AND(G$48>=$A6,OR($B6="",G$49="",$B6>=G$49),OR($C6="",G$50="",$C6>=G$50),OR($D6="",G$51="",$D6>=G$51),OR($E6="",G$52="",$E6>=G$52),OR(AND(G$49>0,$B6>G$49),AND(G$50>0,$C6>G$50),AND(G$51>0,$D6>G$51),AND(G$52>0,$E6>G$52))),G$48-$A6,"")

Appendix G. NAEP 2000n Grade 4 Math: Race Changes (White vs. Non-white)

GROUPS: WHITE & NON-WHITE **741 Pairs** **62 Changes** **NAEP 2000n Grade 4 Math** Number shown is difference in state scores

Non-White Scores Inferred **43 Reversals** **9 Size > 4** **20 States Overtake** Shaded cells have a difference of 5 or more

2 digit for VT, ME, WV, ND and IA **11 Statistically Significant** **23 States Overtaken** Bold indicates statistical significance

Sort: Ave(D),W(A), NW(A)

DS = W - NW					**23**	2	1	6	1	2	0	1	3	0	0	3	1	1	2	0	1	2	4	2	1	0	4	2	0	3	0	0	8	8	0	0	0	0	2	0	2	0	0		20	Max
ST	Ave	%W	W	NW	DS	MN	MA	IN	CT	IA	TX	VT	KS	NC	ME	ND	OH	MI	MT	VA	WY	MO	UT	ID	OR	NY	NE	WV	OK	RI	IL	MD	KY	TN	NV	GA	SC	AZ	AL	LA	AR	NM	CA	ST	#	Sz
MN	235	82	239	217	22																																							MN	0	
MA	235	78	241	214	27																																							MA	0	
IN	234	88	237	212	25																																							IN	0	
CT	234	72	242	213	29			0																																				CT	1	0
IA	233	90	235	217	18																																							IA	0	
TX	233	44	243	225	18	2	2	1	1	0																																		TX	5	2
VT	232	97	232	217	15																																							VT	0	
KS	232	79	237	213	24			2																																				KS	1	2
NC	232	62	240	219	21	3		2		1		0	0																															NC	5	3
ME	231	97	231	224	7																																							ME	0	
ND	231	91	233	212	21																																							ND	0	
OH	231	80	235	215	20											0																												OH	1	0
MI	231	77	239	204	35																																							MI	0	
MT	230	86	233	212	21																																							MT	0	
VA	230	63	239	215	24			4					2			1		1	0																									VA	5	4
WY	229	89	231	213	18																																							WY	0	
MO	229	79	235	206	29																																							MO	0	
UT	227	86	230	209	21																																							UT	0	
ID	227	84	230	211	19																		0																					ID	1	0
OR	227	81	230	214	16																		0	0																				OR	2	0
NY	227	52	238	215	23			**7**					**5**			4	4		3		2	2	0	0	0																			NY	10	7
NE	226	83	231	202	29																																							NE	0	
WV	225	94	226	210	16																																							WV	0	
OK	225	67	229	217	12																							0																OK	1	0
RI	225	75	233	201	32																																							RI	0	
IL	225	57	236	210	26																	4	2				1	0		0														IL	5	4
MD	222	52	237	206	31																						4			3														MD	2	4
KY	221	87	224	201	23																																							KY	0	
TN	220	74	227	200	27																																							TN	0	
NV	220	60	227	210	18																												1	0										NV	2	1
GA	220	52	231	208	23																						**6**						1	0										GA	3	6
SC	220	56	233	203	30																						**6**			**5**			1	0										SC	4	6
AZ	219	56	230	205	25																												2	1										AZ	2	2
AL	218	58	228	204	24																												3	2										AL	2	3
LA	218	53	230	204	26																												3	2					0					LA	3	3
AR	217	70	225	198	27																																							AR	0	
NM	214	38	227	206	21																												**7**	**6**							3			NM	3	7
CA	214	38	229	205	24																												**7**	**6**					4		3			CA	4	7
MS	211	49	224	199	25																																							MS	0	
39					STATE	MN	MA	IN	CT	IA	TX	VT	KS	NC	ME	ND	OH	MI	MT	VA	WY	MO	UT	ID	OR	NY	NE	WV	OK	RI	IL	MD	KY	TN	NV	GA	SC	AZ	AL	LA	AR	NM	CA	39	**62**	7
					Ave	235	235	234	234	233	233	232	232	232	231	231	231	231	230	230	229	229	227	227	227	227	226	225	225	225	225	222	221	220	220	220	220	219	218	218	217	214	214			
					%W	82	78	88	72	90	44	97	79	62	97	91	80	77	86	63	89	79	86	84	81	52	83	94	67	75	57	52	87	74	60	52	56	56	58	53	70	38	38			
					White	239	241	237	242	235	243	232	237	240	231	233	235	239	233	239	231	235	230	230	230	238	231	226	229	233	236	237	224	227	227	231	233	230	228	230	225	227	229			
					Non-White	217	214	212	213	217	225	217	213	219	224	212	215	204	212	215	213	206	209	211	214	215	202	210	217	201	210	206	201	200	210	208	203	205	204	204	198	206	205			

G8 '=IF(AND($B7<=G$47,$D7>=G$49,$E7>=G$50,OR($D7>G$49,$E7>G$50))=TRUE,G$47-$B7,"")

Appendix H. NAEP 2002 Grade 8 Reading: Family Income Changes

ADJUST FOR FAMLY INCOME 780 Unique Pairs 31 Changes **NAEP 2002 Grade 8 Reading** Numbers shown are differences in state scores

EL=Elgible; NEL = Not Elgible 19 Reversals 1 Size > 3 20 States Overtaking Shading indicated difference is at least 4

1 Statistically significant 13 States Overtaken Bold indicates statistical significance

Sort: Ave(D), EL(A), NEL(A)

																																											13		Max
Ave	EL	NEL	ST	0	0	2	0	0	0	0	0	2	0	0	0	3	0	0	2	5	0	0	0	0	0	0	5	2	0	1	0	0	1	0	3	2	2	0	0	0	0	1	31	20	Size
272	257	276	VT																																								VT	0	
271	253	278	MA																																								MA	0	
270	260	273	ME																																								ME	0	
270	260	275	NE			0																																					NE	1	0
270	261	274	MT			0																																					MT	1	0
269	251	276	KS																																								KS	0	
269	256	274	VA																																								VA	0	
268	254	274	WA																																								WA	0	
268	257	272	OR																																								OR	0	
268	257	273	MO									0																															MO	1	0
268	257	273	OH									0																															OH	1	0
268	261	270	ND																																								ND	0	
267	247	275	CT																																								CT	0	
267	253	275	DE													0																											DE	1	0
266	259	270	ID																																								ID	0	
265	246	274	PA																																								PA	0	
265	253	269	IN																																								IN	0	
265	253	273	KY																	0																							KY	1	0
265	253	273	NC																	0																							NC	1	0
265	257	270	MI																	0																							MI	1	0
265	258	268	WY																																								WY	0	
264	250	275	NY													3			1																								NY	2	3
264	255	269	WV																	1																							WV	1	1
263	248	269	MD																																								MD	0	
263	249	269	UT																								0																UT	1	0
262	248	275	TX													**5**			3								1																TX	3	5
262	249	270	RI																								1	1															RI	2	1
262	253	270	OK																	3							1	1		0													OK	4	3
261	249	269	FL																								2																FL	1	2
260	246	268	TN																																								TN	0	
260	250	268	AR																														0										AR	1	0
258	245	267	GA																																								GA	0	
258	245	268	SC																																0								SC	1	0
257	242	266	AZ																																								AZ	0	
256	246	268	LA																																2	2	1						LA	3	2
255	246	268	MS																																3	3	2						MS	3	3
254	245	265	NM																																								NM	0	
253	240	264	AL																																								AL	0	
251	240	256	NV																																								NV	0	
250	240	262	CA																																							1	CA	1	1
40			ST	VT	MA	ME	NE	MT	KS	VA	WA	OR	MO	OH	ND	CT	DE	ID	PA	IN	KY	NC	MI	WY	NY	WV	MD	UT	TX	RI	OK	FL	TN	AR	GA	SC	AZ	LA	MS	NM	AL	NV	31	20	5
			Ave	272	271	270	270	270	269	269	268	268	268	268	268	267	267	266	265	265	265	265	265	265	264	264	263	263	262	262	262	261	260	260	258	258	257	256	255	254	253	251			
			EL	257	253	260	260	261	251	256	254	257	257	257	261	247	253	259	246	253	253	253	257	258	250	255	248	249	248	249	253	249	246	250	245	245	242	246	246	245	240	240			
E6			NEL	276	278	273	275	274	276	274	274	272	273	273	270	275	275	270	274	269	273	273	270	268	275	269	269	269	275	270	270	269	268	268	267	268	266	268	268	265	264	256			

'=IF(AND(E$47>=$A6,OR($B6="",E$48="",$B6>=E$48),OR($C6="",E$49="",$C6>=E$49),OR(AND(E$48>0,$B6>E$48),AND(E$49>0,$C6>E$49))),E$47-$A6,"")

Appendix I. NAEP 2002 Grade 8 Reading: Race/Ethnicity Changes (All four subgroups)

ADJUST FOR RACE/ETHNICITY
W=White,B=Black,H=Hisp.,A=Asian

137 Changes
117 Reversals
55 Size > 3

780 Unique Pairs
507 Stat. Significant differences
11 % of S/S differences are reversed

NAEP 2002 Grade 8 Reading
29 States Overtaking
25 States Overtaken

Number shown is the difference in state scores
Shading indicates a difference of at least 4
Bold indicates statistical significance

Sort: Ave(D),W(A),B(A),H(A),A(A)

Ave	W	B	H	A	ST	12	0	14	4	7	3	0	12	5	3	1	5	0	0	6	4	1	10	4	4	0	11	1	9	0	4	3	0	0	7	5	1	0	0	0	0	0	1	0	25 137	#	Max Size
272	272				VT																																								VT	0	
271	278	246	246	270	MA	1																																							MA	1	1
270	270				ME																																								ME	0	
270	273	246	251		NE	2		0																																					NE	2	2
270	273				MT	2		0																																					MT	2	2
269	273	244	253		KS	3		1																																					KS	2	3
269	275	252	261	279	VA	3		1	1	1	0																																		VA	5	3
268	269				ND																																								ND	0	
268	270		249	275	OR								0																																OR	1	0
268	271	247	247	272	WA			2					0																																WA	2	2
268	271	250			MO			2					0	0	0																														MO	4	2
268	273	246			OH	*4*		2			1		0	0																															OH	5	4
267	275	252	250	282	DE	**5**		3		3			1	1	1	1	1																												DE	8	5
267	277	240	239	265	CT	**5**		3		3			1																																CT	4	5
266	269		247		ID																																								ID	0	
265	267	247			IN																																								IN	0	
265	267	248			KY																0																								KY	1	0
265	267		249		WY																																								WY	0	
265	270	242			MI								3							1			0																						MI	3	3
265	271	236	241	253	PA			**5**					3																																PA	2	5
265	274	247	252		NC	**7**		**5**	**5**	**5**			3	3	3		3			1	0		0	0	0																				NC	13	7
264	264	242			WV																																								WV	0	
264	274	246	251	261	NY	**8**		**6**	**6**	**6**			**4**				*4*			2			1	1	1		0																		NY	11	8
263	267		238	254	UT																						1																		UT	1	1
263	274	246	253	284	MD	**9**		**7**	**7**	**7**	**6**		**5**	**5**			**5**			3			2	2	2		1	1	0																MD	**15**	9
262	268	238	251		OK																		3						1																OK	2	3
262	268	243	240	251	RI																						2																		RI	1	2
262	276	247	250	271	TX	**10**		**8**		**8**			**6**				**6**			**4**	3		3	3	3		2		1			0													TX	13	10
261	269	244	252		FL															**5**			**4**				3		2		1	1													FL	6	5
260	265	240			TN																																								TN	0	
260	267	238			AR																																								AR	0	
258	268	243			SC																		**7**				**6**		**5**		**4**				2	2									SC	6	7
258	268	246	242	265	GA																						**6**		**5**			**4**			2	2	0								GA	6	6
257	267	250	242		AZ																**8**	**8**					**7**		**6**						3	3									AZ	6	8
256	268	240			LA																		**9**						**7**		**6**				*4*	**4**									LA	5	9
255	268	240			MS																		**10**						**8**		**7**				**5**	**5**									MS	5	10
254	266		247		NM																						**10**								**6**										NM	2	10
253	264	234			AL																																								AL	0	
251	259	234	237	258	NV																																								NV	0	
250	265	242	238	257	CA																						**14**								**10**								3		CA	3	14
40					ST	VT	MA	ME	NE	MT	KS	VA	ND	OR	WA	MO	OH	DE	CT	ID	IN	KY	WY	MI	PA	NC	WV	NY	UT	MD	OK	RI	TX	FL	TN	AR	SC	GA	AZ	LA	MS	NM	AL	NV	29	137	14
					Ave	272	271	270	270	270	269	269	268	268	268	268	268	267	267	266	265	265	265	265	265	265	264	264	263	263	262	262	262	261	260	260	258	258	257	256	255	254	253	251			
					W	272	278	270	273	273	273	275	269	270	271	271	273	275	277	269	267	267	267	270	271	274	264	274	267	274	268	268	276	269	265	267	268	268	267	268	268	266	264	259			
					B		246		246		244	252			247	250	246	252	240		247	248		242	236	247	242	246		246	238	243	247	244	240	238	243	246	250	240	240		234	234			
					H		246		251		253	261		249	247			250	239	247			249		241	252		251	238	253	251	240	250	252				242	242			247		237			
G6:					A		270					279		275	272			282	265						253			261	254	284		251	271					265						258			

'=IF(AND(G$47>=$A6,OR($B6="",G$48="",$B6>=G$48),OR($C6="",G$49="",$C6>=G$49),OR($D6="",G$50="",$D6>=G$50),OR($E6="",G$51="",$E6>=G$51),OR(AND(G$48>0,$B6>G$48),AND(G$49>0,$C6>G$49),AND(G$50>0,$D6>G$50),AND(G$51>0,$E6>G$51))),G$47-$A6,"")

Appendix J. NAEP 2002 Grade 8 Reading: Race Changes (White vs. Non-White)

ADJUST FOR RACE/ETHNICITY
W=White,NW = Non-White

780 Unique Pairs 81 Changes
64 Reversals 19 Size > 3
18 Statistically Significant

NAEP 2002 Grade 8 Reading
25 States Overtaking
21 States Overtaken

Number shown is the difference in state scores
Shading indicates a difference of at least 4
Bold indicates statistical significance

Sort: Ave(D),NW(A),W(A)																																												21	#	Max
%W	NW	W	Ave	ST	0	0	6	2	0	3	0	5	4	1	0	0	0	0	6	4	9	4	3	4	0	1	0	11	1	5	0	0	0	5	4	1	0	0	0	0	0	1	1	81	25	Size
96	272	272	272	VT																																								VT	0	
73	252	278	271	MA																																								MA	0	
87	250	273	270	MT																																								MT	0	
86	252	273	270	NE			0																																					NE	1	0
96	270	270	270	ME																																								ME	0	
82	251	273	269	KS			1																																					KS	1	1
66	257	275	269	VA			1	1		0																																		VA	3	1
81	247	273	268	OH																																								OH	0	
94	252	269	268	ND																																								ND	0	
81	255	271	268	MO									0																															MO	1	0
78	257	271	268	WA									0	0																														WA	2	0
82	259	270	268	OR									0																															OR	1	0
70	244	277	267	CT																																								CT	0	
63	253	275	267	DE			3	3		2		1	1																															DE	5	3
89	242	269	266	ID																																								ID	0	
81	239	271	265	PA																																								PA	0	
90	247	267	265	KY																																								KY	0	
77	248	270	265	MI															1		0																							MI	2	1
64	249	274	265	NC								3							1	0	0	0																						NC	5	3
88	250	267	265	WY																	0																							WY	1	0
86	253	267	265	IN																	0			0																				IN	2	0
57	251	274	264	NY			**6**					*4*							2	1	1	1	1	1																				NY	8	6
95	264	264	264	WV																																								WV	0	
86	238	267	263	UT																																								UT	0	
55	250	274	263	MD								**5**							3	2	2	2	2					0																MD	7	5
76	243	268	262	RI																								1																RI	1	1
44	251	276	262	TX			**8**			**7**		**6**							**4**	3	3	3	3	3		2		1	1	0														TX	**13**	**8**
62	252	268	262	OK																	3			3				1		0														OK	4	3
58	250	269	261	FL															**5**		**4**							2		1														FL	4	5
75	239	267	260	AR																								3																AR	1	3
77	243	265	260	TN																																								TN	0	
56	245	268	258	SC																								**5**		**4**				2	2									SC	4	5
54	246	268	258	GA																								**5**		**4**				2	2	0								GA	5	5
56	244	267	257	AZ																								**6**						3	3									AZ	3	6
55	241	268	256	LA																								**7**						**4**										LA	2	7
53	240	268	255	MS																								**8**						**5**										MS	2	**8**
38	247	266	254	NM																															**6**									NM	1	6
61	236	264	253	AL																																								AL	0	
60	239	259	251	NV																																								NV	0	
35	242	265	250	CA																																						3	1	CA	2	3
40				ST	VT	MA	MT	NE	ME	KS	VA	OH	ND	MO	WA	OR	CT	DE	ID	PA	KY	MI	NC	WY	IN	NY	WV	UT	MD	RI	TX	OK	FL	AR	TN	SC	GA	AZ	LA	MS	NM	AL	NV		81	8
				Ave	272	271	270	270	270	269	269	268	268	268	268	268	267	267	266	265	265	265	265	265	265	264	264	263	263	262	262	262	261	260	260	258	258	257	256	255	254	253	251			
				W	272	278	273	273	270	273	275	273	269	271	271	270	277	275	269	271	267	270	274	267	267	274	264	267	274	268	276	268	269	267	265	268	268	267	268	268	266	264	259			
				NW	272	252	250	252	270	251	257	247	252	255	257	259	244	253	242	239	247	248	249	250	253	251	264	238	250	243	251	252	250	239	243	245	246	244	241	240	247	236	239			
				%W	96	73	87	86	96	82	66	81	94	81	78	82	70	63	89	81	90	77	64	88	86	57	95	86	55	76	44	62	58	75	77	56	54	56	55	53	38	61	60			

F6 '=IF(AND(F$47>=$D6,$C6>=F$48, $B6>=F$49,OR($B6>F$49, $C6>F$48)),F$47-$D6,"")