

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 are statistically significant. A more disputable approach generated 117 Simpson's reversals in the same data set of which 52 are statistically significant. Either way these results support the claim that Simpson's Paradox is not rare in NAEP data. 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 EDUCATION 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 1992 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.

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,'

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."

'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.

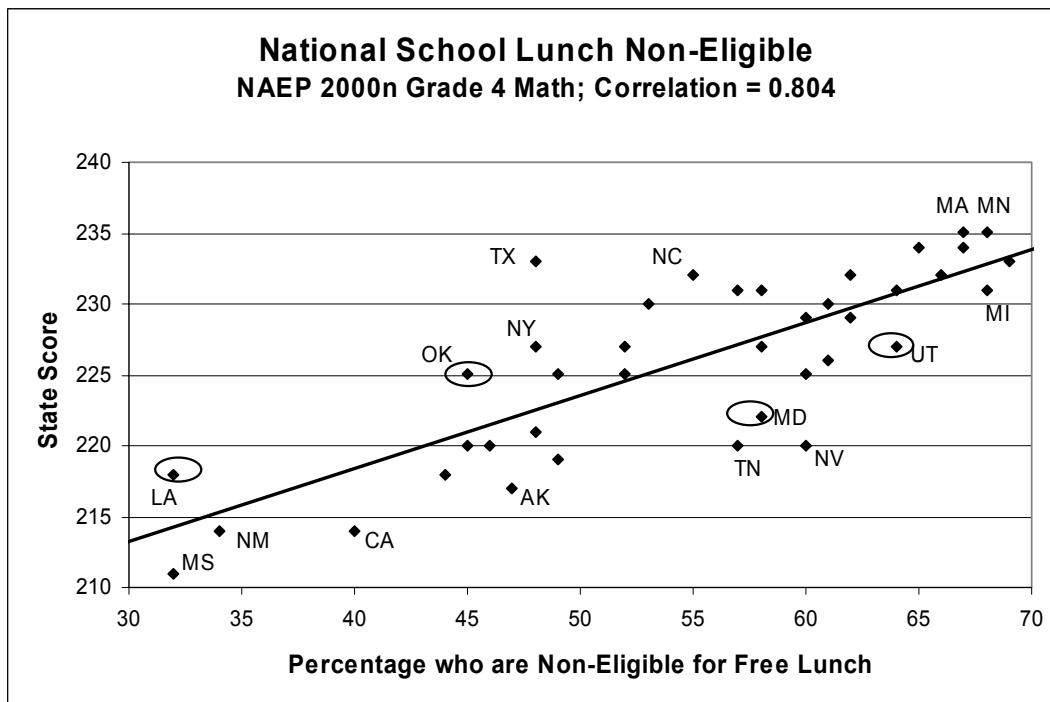


Figure 1: State Scores vs. Percentage who are Non-Eligible for Free Lunch

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.

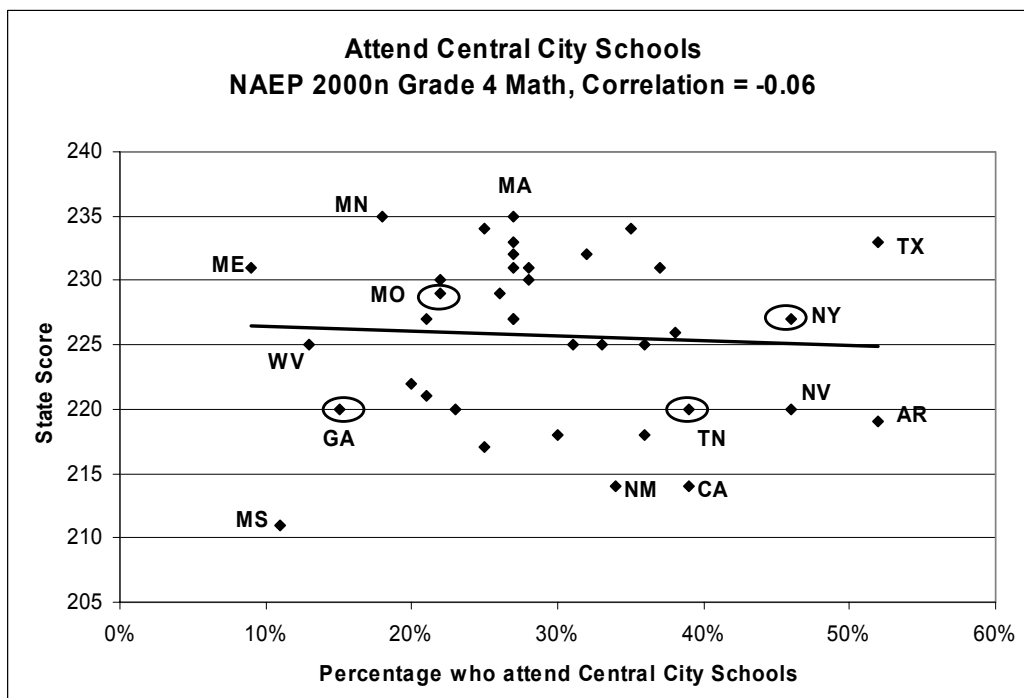


Figure 2: State Scores vs. Percentage who attend a Central City School

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.

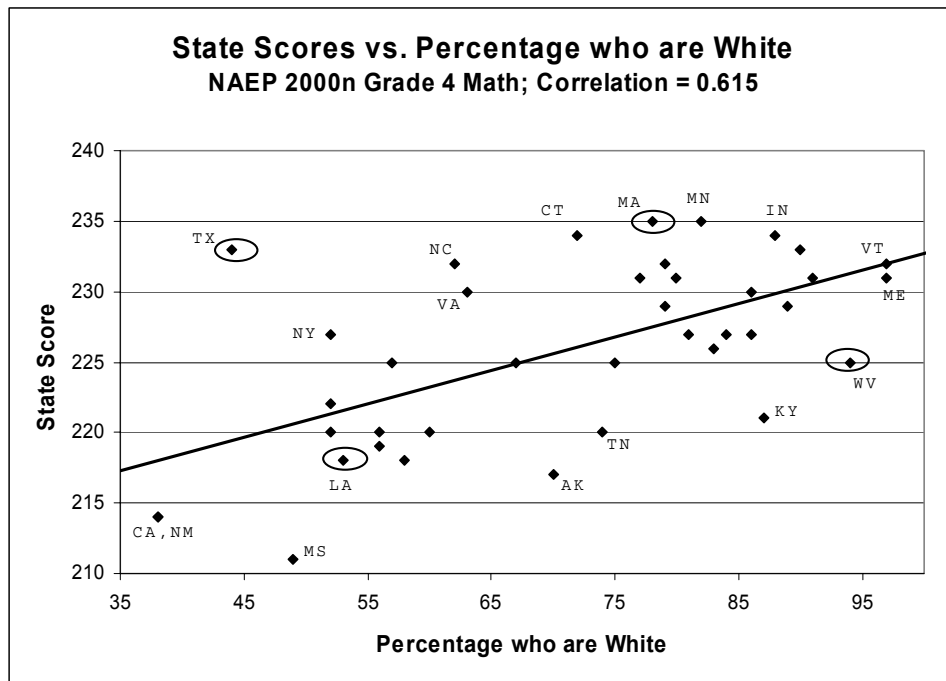


Figure 3: State Scores vs. Percentage of Students who are White

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 income⁴. 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” sub group 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 location⁵. 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 enough⁶ 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.

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).

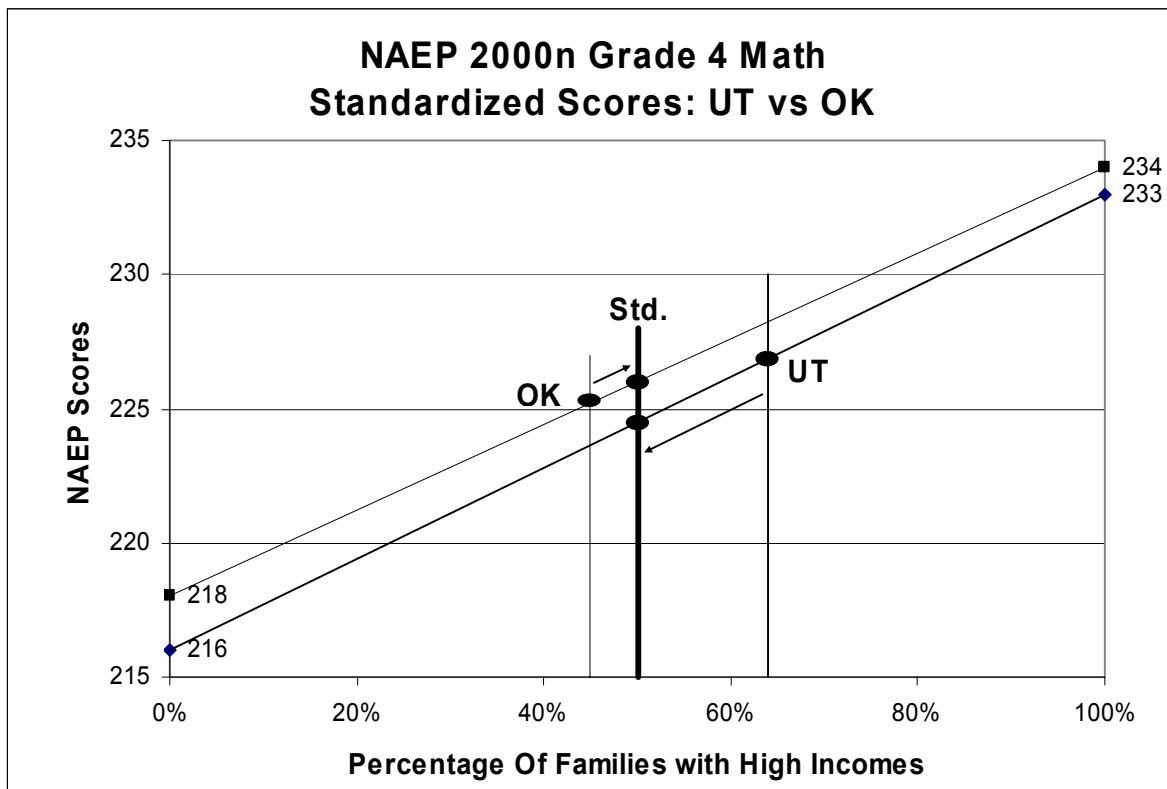


Figure 4: Simpson's Paradox: UT vs. OK

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. In this case, 50% of students in both states taken collectively had high family incomes. As shown in Figure 4, the standardized state score is 226 for Oklahoma and 224 for Utah. Thus the standardized score is two points higher for Oklahoma (226) than for Utah (224). 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; blacks 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).

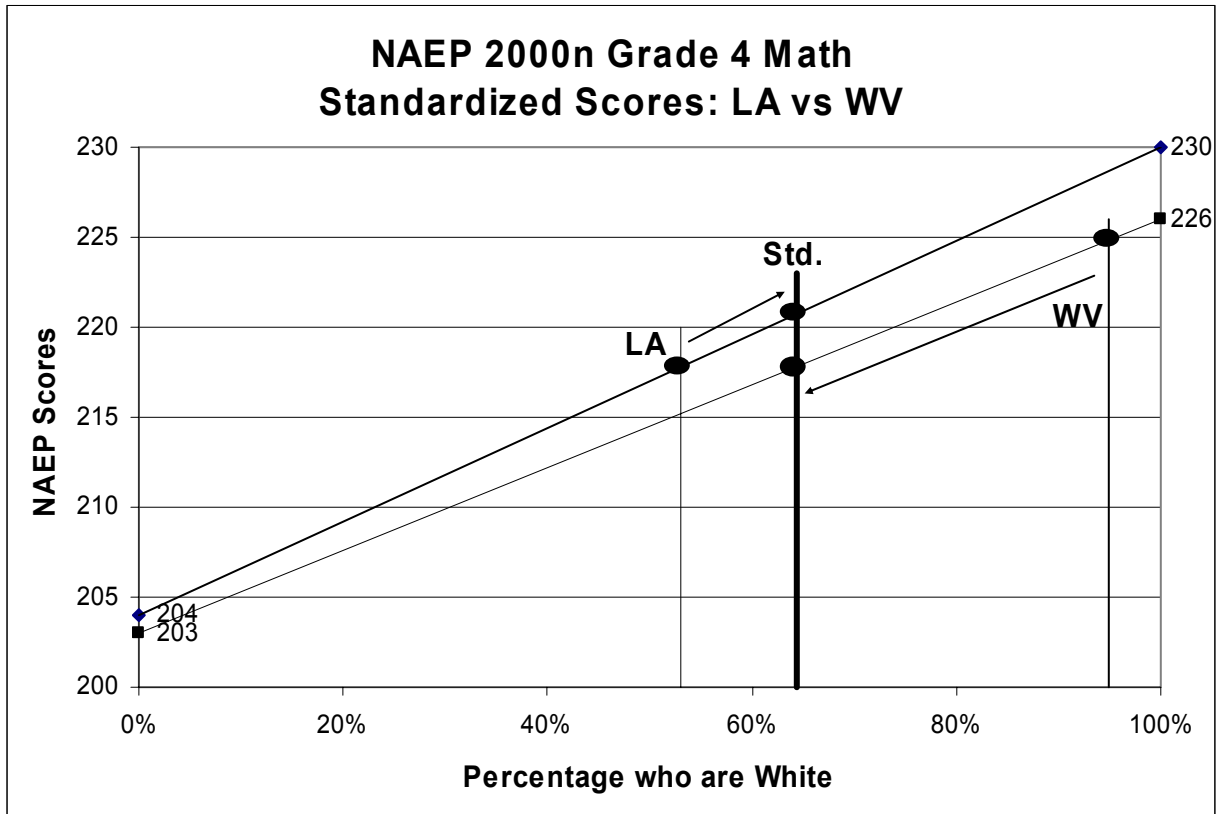


Figure 5: Simpson's Paradox: LA vs. WV

As before, standardized state scores are calculated using the confounder mixture that is found when both states are combined. The standardized state score is three points higher for Louisiana (221) than for West Virginia (218). Once again, 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 identify 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 (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.⁸

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.⁹ 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 being compared.¹⁰ 'Change' and 'Reverse' indicates the number of changes and Simpson's reversals. 'Stat Sig' indicates the number of these reversals that are statistically significant at the 5% level.

Note that statistical significance as used in this context does not mean the statistical significance of the difference in standardized scores: scores generated by giving two units a standard mixture of a given confounder. It means that the difference in the original scores is statistically significant. And this means that the associated Simpson's reversal is statistically significant.

Confounder	States	Pairs	Change	Reverse	Stat Sig	Source
School Lunch: 2 groups	39	741	23	15	0	Appendix D
School Location: All	38	703	4	1	0	
Race/ethnicity: All	41	820	123	97	43	Appendix E
Race: White vs. Non-white	39	741	64	46	11	Appendix F

Table 4 Changes and Reversals for NAEP 2000n Grade 4 Math

Confounder	States	Pairs	Change	Reverse	Stat Sig	Appendix
School Lunch: 2 groups	40	780	31	19	1	Appendix G
School Location: All	39	741	3	3	0	
Race/Ethnicity: All	40	780	137	117	52	Appendix H
Race: White vs. Non-white	40	780	82	64	18	Appendix I

Table 5 Changes and Reversals for NAEP 2002 Grade 8 Reading

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 are initially statistically significant. In the NAEP 2000n data there are 46 Simpson's reversals of which 11 are initially statistically significant.

⁸ 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.

⁹ 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.

¹⁰ If there are N states, there are N(N-1)/2 state pairs.

- Using the more disputable “all” approach involving incomparable subgroups, there are 117 Simpson's reversals in the NAEP 2002 data of which 52 are initially statistically significant. In the NAEP 2000n data there are 97 Simpson's reversals of which 43 are initially statistically significant.

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

7. STATISTICAL SIGNIFICANCE

Statistical significance for a single pair-wise comparison of state means was determined two ways.

- Using a single estimate. The width of a 95% confidence interval was calculated based on the mean of the standard errors for the states studied.¹¹ This approach was used to determine the shaded areas in the tables in the Appendices. Since the sample sizes were similar for all the states, the differences in standard errors largely reflected differences in the standard deviations. 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.
- Using the NAEP data tool for a given pair of states to see whether that difference in scores was statistically significant. This approach was used to count those Simpson's reversals that are statistically significant as shown in Table 4 and Table 5.

The weakness of the first approach is the assumption that the confidence interval for the difference in state means is the same for any two states (which it is not); the strength is the ease of calculation. The strength of the second is the accuracy of the result; the weakness is the manual labor necessary to obtain the results. The actual differences between these approaches were small: less than 5 shifts out of 50.

8. '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 making a cross-sectional comparison among states and in making a longitudinal comparison for a given state. But journalists are not generally as careful. News stories involving NAEP data often compare state scores with previous scores and with those of others states.¹² This may be done using point differences or rank differences. These differences are often small and therefore lack statistical significance. But they are being reported so 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) have ‘journalistic significance.’

9. NAEP POLICY ON ADJUSTING DATA

Adjusting data for confounders is controversial. In 1994, the NAEP Board of Directors 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 character-*

¹¹ 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 insignificant rather than to make precise measurements.

¹² A convenience survey of five press releases by state education departments found that all of them included comparisons that were not statistically significant.

istics 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."* NAEP (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 the associated 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." And 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 potentially confounding factors. These confounders often have a sufficiently strong influence on results that the rank order of states based on the NAEP scores will be affected. Three specific activities are recommended.

- (1) State scores should be listed by various confounder values to allow comparisons between states and comparisons over time. For example, for just those students who receive reduced fee school lunches, how do state scores compare with each other? For just Hispanics, how do the scores for a given state in 2002 compare with those for the same state in 1990? This data is already published by NAEP. It is just a matter of disaggregating the data and presenting it separately. This avoids any methodological issues involved in standardizing. Examples of this are shown in Appendix C.
- (2) Generate adjusted scores for states after taking into account the influence of factors outside the school's control: school location and socio-economic factors. Doing so will allow a discussion of many methodological issues.
- (3) Generate adjusted scores for states after taking into account the influence of student race/ethnicity. This will allow for a more complete discussion of the influence of race and associated factors.
- (4) 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 when pair-wise comparisons are made between states using NAEP data. 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. Of these Simpson's reversals, 55 are statistically significant in the NAEP 2002 Grade 8 Reading data. All Simpson's reversals – whether statistically significant or not – are newsworthy, which gives them 'journalistic significance.' It is recommended that state scores 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.

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

Schild, Milo (1999). "Simpson's Paradox and Cornfield's Conditions." *1999 ASA Proceedings of the Section on Statistical Education*, p. 106-111.¹³

Schild, Milo (2004). "Three Graphs That Can Change Statistical Education." International Association of Statistical Educators (IASE) 2004 Roundtable on Curriculum Development in Statistical Education.¹³

Schild, 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- 3697.¹³

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*.¹³

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 Education 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. Schild is Director of the W. M. Keck Statistical Literacy program and Professor of Business Administration at Augsburg College. His e-mail address is Schild@Augsburg.edu. This paper is at www.augsburg.edu/ppages/~schild and at www.StatLit.org.

¹³ These papers are at www.StatLit.org/Articles.

Appendix A. MISSING DATA (LACK OF COMMON SUBGROUPS)

Missing data is a problem. The problem is not that needed data was not obtained; 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. Almost all 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.¹⁴ 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.

As mentioned previously, the state scores are for NAEP 2000n Grade 4 Math. There are 64 Simpson's reversals among these 82 changes when 26 states overtake 22 states after adjusting for race using white and non-white groups. Of the 820 matches, 10% (82) changed rank while 8% (64) reversed ranks.

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 (white, black, Hispanic and Asian) 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 is a reasonable approach to handling the existing racial/ethnic differences between states.

Appendix B. TECHNICAL DETAILS

This appendix identifies how the data and formula were entered into the spreadsheets shown in subsequent appendices. While NAEP data is readily obtainable on the web, obtaining the data for all jurisdictions in one list requires extra steps.¹⁵ Once obtained, the ordering of the states is critical. To locate all the reversals and changes in the lower-triangle (e.g., Appendix D), the sort order must be state average (descending) and the NAEP scores of all state subgroups must be ascending.¹⁶ Since Excel handles a maximum of three sort groups, in the case of race (Appendix E) this means sorting first on Black (A),

¹⁴ 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.

¹⁵ Go to <http://nces.ed.gov/nationsreportcard/naepdata/search.asp>. Select the *Subject* (Mathematics), *Grade* (8th), *Jurisdiction* (National public) and *Category* (Major reporting groups). Press Continue. From the next screen, uncheck years not desired. Select *Major Reporting Group desired* (1. All students). This returns the related NAEP scores for the national public level. From the menu above, select 'User Options'. From the sub-menu, select 'Add/Remove Jurisdictions.' Press 'Select All' and then uncheck unwanted jurisdictions. This gives the NAEP score by state. All other choices of Reporting Group give just the scores within that subgroup.

¹⁶ 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.

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 B2. 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 any formulas were involved in the data being copied, the check box for Values would also need to have been checked). Thus, states on the left are trying to overtake; states along the bottom are being overtaken.

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 D).

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 Appendix E.

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. 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). 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.¹⁷

Statistical significance was obtained in two ways: an estimated approach and an actual approach. The estimated approach estimates the size of a single confidence interval to use in comparing all states. See Section 7. The statistical significance obtained using the estimated approach was used to determine the formula for conditional formatting which generated the shaded area in the tables in the subsequent appendices. Otherwise, statistical significance was always determined using the exact approach: an approach based on the results from the NAEP data tool.¹⁸

¹⁷ 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.

¹⁸ Select a grade, test, state and "Major Reporting Groups"; press "Continue." Check the year desired and select Major Reporting Group 1: "All Students." This gives the appropriate state score. Select "User Options" menu and select "Add/Delete Jurisdictions" to obtain the scores of other states. From the list of Jurisdictions, select the state or states against which the original state score is to be compared. Press the "Accept Changes" button. Select the "User Options" menu and select "Check Significant Differences." 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. This 'Difference Check Result' popup window shows the differences in question and the associated p-values.

Appendix C. STATE NAEP 2000n Grade 4 MATH RANKS SORTED BY SUBGROUPS**State Scores by Race/Ethnic subgroups**

State	ALL	White	Black	Hisp	Asian
TX	5.5	1	1	2.5	1
CT	3.5	2	8	11	3.5
MA	1.5	3	10	14.5	5.5
NC	8	4	2		
MN	1.5	6	11		8.5
MI	11.5	6	26.5		
VA	14.5	6	8	1	2
NY	21.5	8	5.5	12	3.5
IN	3.5	10	8		
KS	8	10	17.5	7	
MD	29	10	23	5	7
DESS	18.5	12.5	3	4	
IL	26.5	12.5	20.5	7	
IA	5.5	15	4		
OH	11.5	15	13		
MO	16.5	15	24		
ND	11.5	19			
MT	14.5	19			
ODDS	18.5	19	5.5	2.5	8.5
RI	26.5	19	25	22	
SC	32.5	19	20.5		
VT	8	22			
ME	11.5	24.5			
WY	16.5	24.5		9.5	
NE	24	24.5	31	20	
GA	32.5	24.5	15.5	9.5	
ID	21.5	29		14.5	
OR	21.5	29		17	5.5
UT	21.5	29	19	13	19
AR	35	29	13	18	10
LA	36.5	29	17.5		
OK	26.5	32.5	15.5	7	
CA	39.5	32.5	32	21	11.5
AL	36.5	34	20.5		
NV	32.5	36	13	14.5	11.5
TN	32.5	36	28.5		
NM	39.5	36		14.5	
WV	26.5	38	20.5		
AK	38	39	30		
KY	30	40.5	26.5		
MS	41	40.5	28.5		

State Scores by School Lunch Eligibility

STATE	OVERALL	Eligible	NonEligible
IA	5.5	1.5	15
DESS	18.5	1.5	32.5
IN	3.5	4.5	7
TX	5.5	4.5	2.5
ME	11.5	4.5	23
ODDS	18.5	4.5	37
ND	11.5	7	18.5
MN	1.5	9	7
NC	8	9	4.5
WY	16.5	9	23
KS	8	13.5	4.5
OH	11.5	13.5	9.5
MT	14.5	13.5	15
ID	21.5	13.5	23
OK	26.5	13.5	23
WV	26.5	13.5	30
CT	3.5	17.5	2.5
VT	8	17.5	12
UT	21.5	19	27.5
VA	14.5	20.5	12
NY	21.5	20.5	9.5
MA	1.5	23	1
MO	16.5	23	12
OR	21.5	23	23
MI	11.5	25	7
NE	24	27	18.5
KY	30	27	32.5
LA	36.5	27	27.5
IL	26.5	29	18.5
SC	32.5	30.5	18.5
NV	32.5	30.5	39
RI	26.5	33	15
AL	36.5	33	35
AK	38	33	37
AR	35	35	32.5
NM	39.5	35.5	40
MD	29	38	27.5
GA	32.5	38	27.5
TN	32.5	38	32.5
MS	41	40	41
CA	39.5	41	37

