Simpson's Paradox Favors Idaho's NAEP Results: 2003-2013

Bert Stoneberg, Ph.D.
K-12 Research Idaho

Abstract

We often hear that even though Idaho does not spend a lot of money on K-12 public education we get a "pretty big bang for our buck." This claim is typically supported by NAEP average scores for all Idaho students as compared with the average scores for all students in the nation's public schools. Simpson's Paradox is a phenomenon in which subgroups show one trend and the aggregate of all subgroups show another. NAEP scores were examined for three demographic groups: All Students, White Students, and Hispanic Students. First, the NAEP 2013 average score for each Idaho demographic group was statistically compared with the average score of their peer group from each of the states and from the nation's public schools. Then the percentile ranks for Idaho and peer national public demographic groups were compared for NAEP reading and mathematics in grades 4 and 8 from 2003 to 2013. For both approaches, Idaho's All Students results for both NAEP subjects in both grades where typically higher than the corresponding results for White Students and Hispanic Students.

Author Note

Bert Stoneberg was Idaho's NAEP State Coordinator from 2002 to 2012. He is retired, but continues independent consulting in educational assessment, evaluation and research. Visit his website at http://k12researchidaho.com

Address correspondence concerning this paper to Bert D. Stoneberg, P.O. Box 5912, Boise ID 83705, or email the author at bert@k12researchidaho.com

Suggested citation: Stoneberg, B.D. (2014). S*impson's Paradox Favors Idaho's NAEP Results: 2003-2013*. Available online: http://k12researchidaho.com/naep/2013/pub0004.pdf

Simpson's Paradox Favors Idaho's NAEP Results: 2003-2013

Bert D. Stoneberg[1]

We in Idaho often hear that even though we don't spend a lot of money on K-12 public education we get a "pretty big bang for our buck." This claim is typically supported by statistics about NAEP average scores for all Idaho students as compared with the average scores for all students in the nation's public schools.

Bracey (2004) in his article entitled *Simpson's Paradox and Other Statistical Mysteries*, pointed out that statements about student achievement that might begin with "Statistics show . . ." need to be carefully examined.

> Simpson's Paradox is a phenomenon in which subgroups show one trend and the aggregate of all subgroups show another. In other words, what is true for the parts is not necessarily true for the whole; hence the paradox. In standardized testing, the paradox frequently crops up when one tries to calculate "national average" scores or "state average" scores. [. . .]
>
> State-level participation on the NAEP was voluntary until the No Child Left Behind Act. Now, Secretary of Education Rod Paige has indicated that he will use the discrepancies between proficient as defined by the states and as defined by NAEP to shame states into doing better. These discrepancies are often quite large. [. . .]
>
> Statistics … is a tricky business, and as Simpson's Paradox suggests, things are not always as straightforward as them seem. That's worth bearing in mind when you hear [statements about] public education that begin, "Statistics show . . ." In fact, the statistics might show something entirely different. Paradoxical, isn't it? [. . .]

NAEP Achievement Level Scores

NAEP achievement level scores are not average scores. Secretary Page did the nation and the states a disservice when he elected to compare state "grade-level" proficient percentages with NAEP "above-grade-level" *Proficient* percentages. It is noteworthy that NAEP protocol uses a capital P and italics to indicate *Proficient* has a stipulated meaning, not the common language meaning.

Andrew Kolstad, senior technical advisor at the National Center for Education Statistics, has explained, "State assessments often define 'proficiency' as solid grade-level performance, often indicating readiness for promotion to the next grade. NAEP's policy

definition of its '*Proficient*' achievement level is 'competency over challenging subject matter' and is implicitly intended to be higher than grade-level performance."

The National Academy of Sciences (NAS) conducted a congressionally mandated external evaluation of NAEP, and published its findings and recommendations in 1998 (Pellegrino, Jones, Mitchell, 1998).

NAS found that the then current achievement-level-setting procedures were fundamentally flawed. The judgment tasks were difficult and confusing; rater's judgments of different item types were internally inconsistent; appropriate validity evidence for the cut-scores was lacking; and the process had produced unreasonable results. NAS recommended that the achievement levels be used on a developmental basis only.

NAS recommended that NAEP reports should focus on the change, from one administration of the assessment to the next, in the percentages of students in each achievement level (i.e., *Below Basic*, *Basic*, *Proficient*, and *Advanced*), rather than focusing on the percentages in each category in a single year. This recommendation challenges Secretary Page's push to compare NAEP and state "proficient" results. It is, however, entirely consistent with NAEP's mission, which is to measure student achievement and to report change in performance over time.

A National Center for Education Statistics (NCES) web page entitled *Status of Achievement Levels* (2013) provides this statement about using the NAEP achievement levels. The strongly edited statement is consistent with the findings and recommendations from the National Academy of Sciences' external evaluation report:

> Federal law requires NAEP achievement levels be used on a trial basis until the Commissioner of Education Statistics determines that the achievement levels are "reasonable, valid, and informative to the public." So far, no Commissioner has made such a determination. Thus, achievement levels should continue to be interpreted and used with caution. The National Assessment Governing Board and NCES believe that the achievement levels are useful for reporting trends in the educational achievement of students in the United States.

A recent statistical study focusing on NAEP and state achievement level percentages, however, has not supported their use for reporting trends. "Trend comparisons require both technical care and substantive consideration. As useful as PAC [percent above cut-score] statistics have been in communicating test results to the public, their properties as trend statistics render them ill-suited for trend comparison" (Ho, 2007).

Demographic Group Percentages and Performance

Simpson's Paradox, as applied to educational assessment, holds that the averages score for "all students" depends not only on the average score for each of the student subgroups but

also on the proportion of students in each subgroup. Exhibit 1 displays the percentage of three demographic groups in each state and the nation's public schools; Exhibit 2 shows the average scores for the demographic groups on NAEP 2013 reading and mathematics. Exhibit 1 shows that Idaho's proportion of White students in 2013 was higher than that of 38 states and the nation's public schools, but lower than only 5 states. Idaho's proportion of Hispanic students was higher than 26 states, but lower than 14 states and the nation. Idaho's proportion of Black students was not significantly different from 4 states, but lower than 45 states and the nation.

Exhibit 2 shows the average scores for all students in the nation's public schools on the NAEP 2013 reading and mathematics at grades 4 and 8.

Exhibit 1. Idaho's student ethnic/racial populations (e.g., White, Hispanic, and Black) statistically compared with student populations in the other states and the nation's public schools from the NAEP 2013 fourth-grade mathematics assessment. [Green = higher than Idaho; Red = lower than Idaho; Yellow = not significantly different from Idaho]

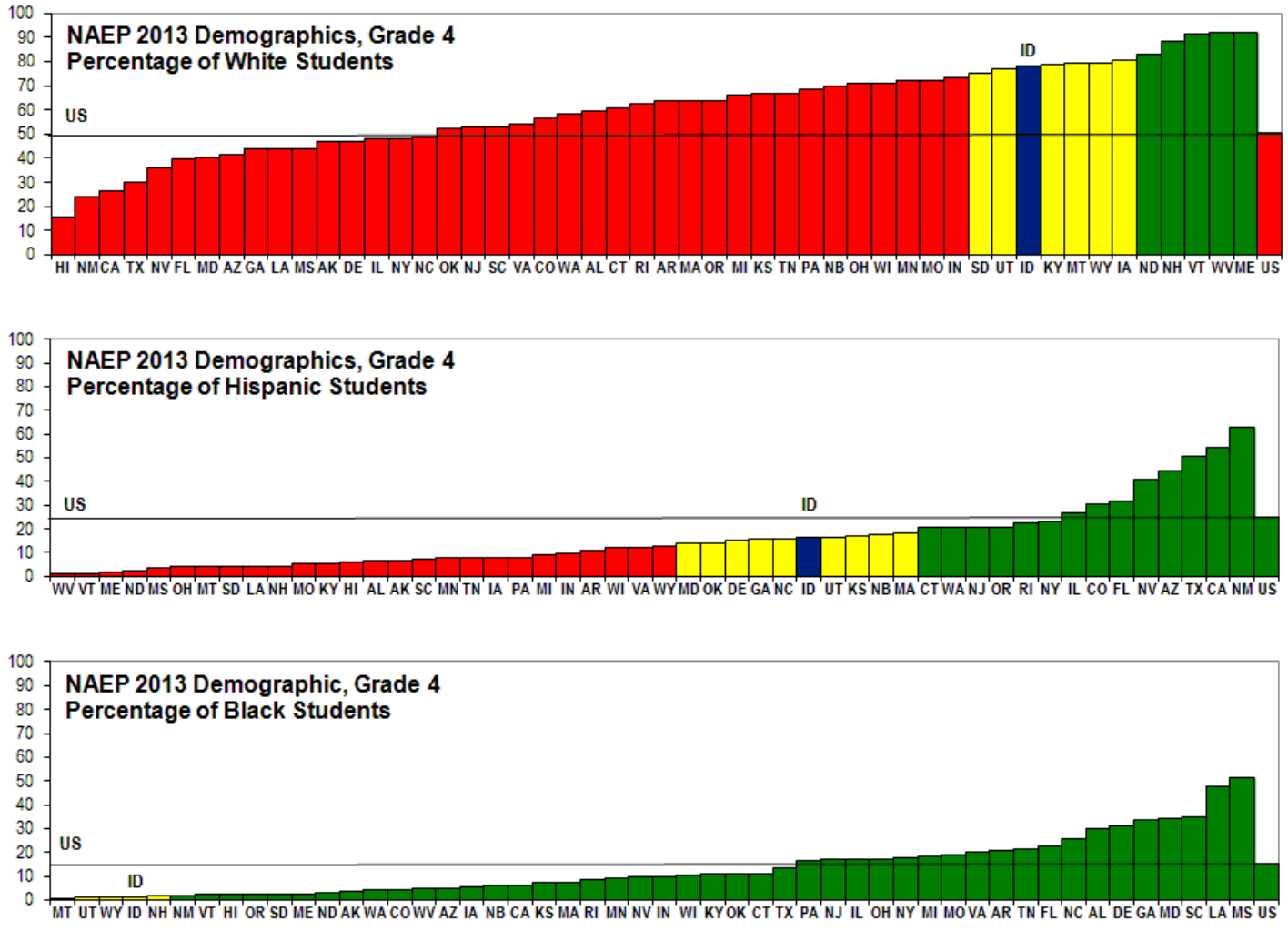


Exhibit 2. National public school averages for three student demographic groups (e.g., White, Hispanic and Black) on NAEP 2013 reading and mathematics for grade 4 and 8.

Grade 4	White	Hispanic	Black
• Reading	231	207	205
• Mathematics	250	230	224
Grade 8			
• Reading	275	255	250
• Mathematics	293	271	263

The "bubble graph" is an excellent tool to illustrate how Idaho benefits from Simpson's Paradox. It will not be used for this paper, but a good example from the Idaho NAEP Results page on the State Department of Education website can be seen in Appendix A.

Methodology

Two basic approaches were used to illustrate the reading and mathematics performance of Idaho's three student demographic groups that received average scores for each NAEP assessment from 2003 to 2013, namely All Students, White Students, and Hispanic Students. The first approach used average scores, the second used percentile ranks.

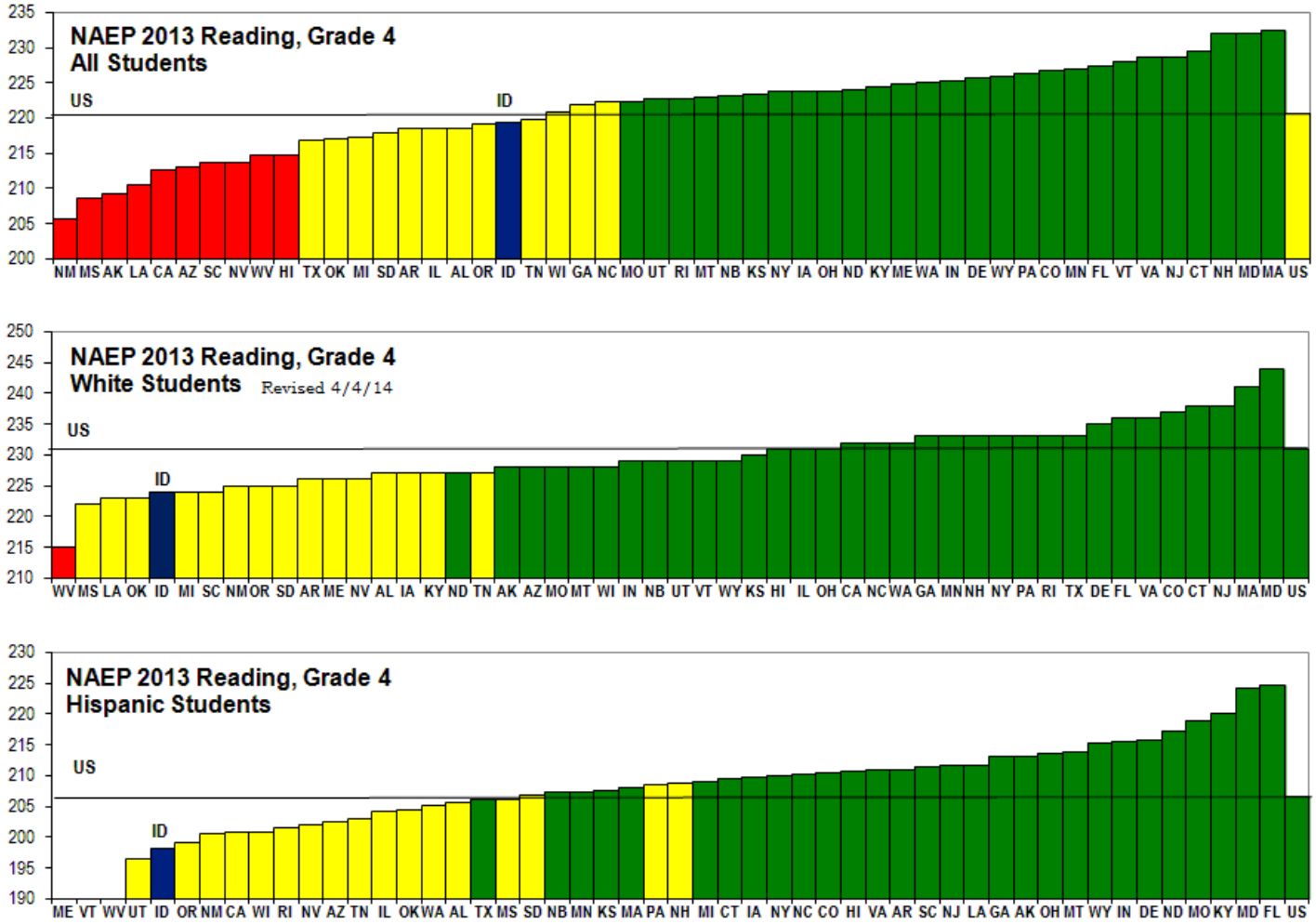
Approach One: Cross-State Comparisons on Average Scores

The NAEP 2013 average score for each Idaho demographic group was statistically compared with the average score of their peer group from each of the states and from the nation's public schools. The NAEP Data Explorer performed the statistical tests preserving a family probability level of 0.05. Then a histogram displaying an ordered list of the states by average score was prepared. The histogram also indicated which states' average scores were higher than Idaho, lower than Idaho, or not significantly different from Idaho. A histogram was made for each NAEP 2013 grade-subject assessment, i.e., grade 4 reading, grade 8 reading, grade 4 mathematics, and grade 8 mathematics.

The three histograms for All Students, White Students, and Hispanic Students each grade-subject were displayed together on one exhibit so the reader could see the place of each Idaho group when compared to their peer groups in the other jurisdictions.

Exhibits 3 through 6 display the results from this examination of NAEP average scores.

Exhibit 3. Cross-state comparisons histograms on average scores from the NAEP 2013 reading assessment for each of three fourth-grade student demographic groups: All Students, White Students, and Hispanic Students.



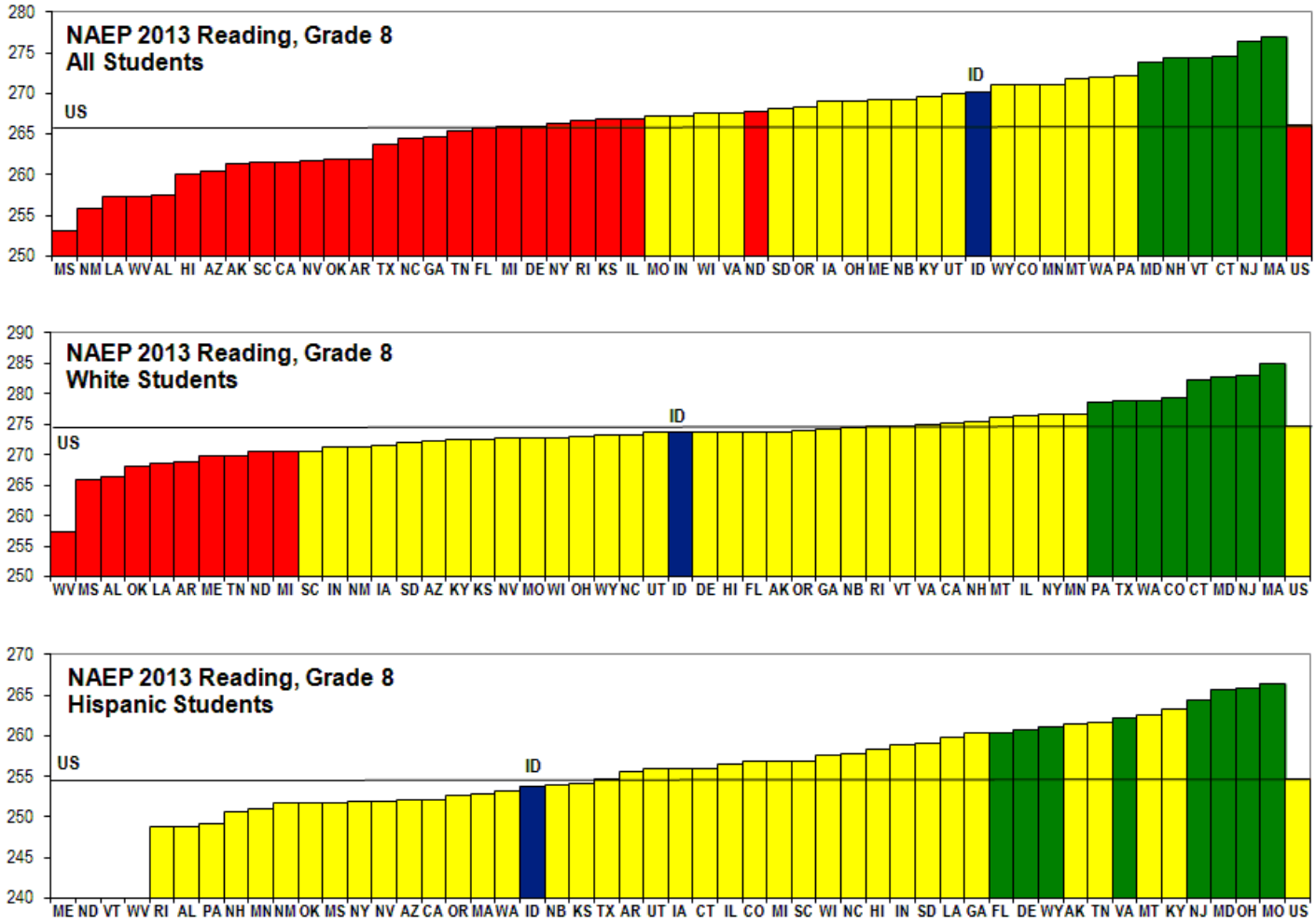
NAEP 2013 fourth-grade reading:

The All Student average score for Idaho was ***not significantly different*** from the All Student national public school average.

The White Student average score for Idaho was ***lower*** than the White Student national public school average.

The Hispanic Student average score for Idaho was ***lower*** than the Hispanic Student national public school average.

Exhibit 4. Cross-state comparisons histograms on average scores from the NAEP 2013 reading assessment for each of three eighth-grade student demographic groups: All Students, White Students, and Hispanic Students.



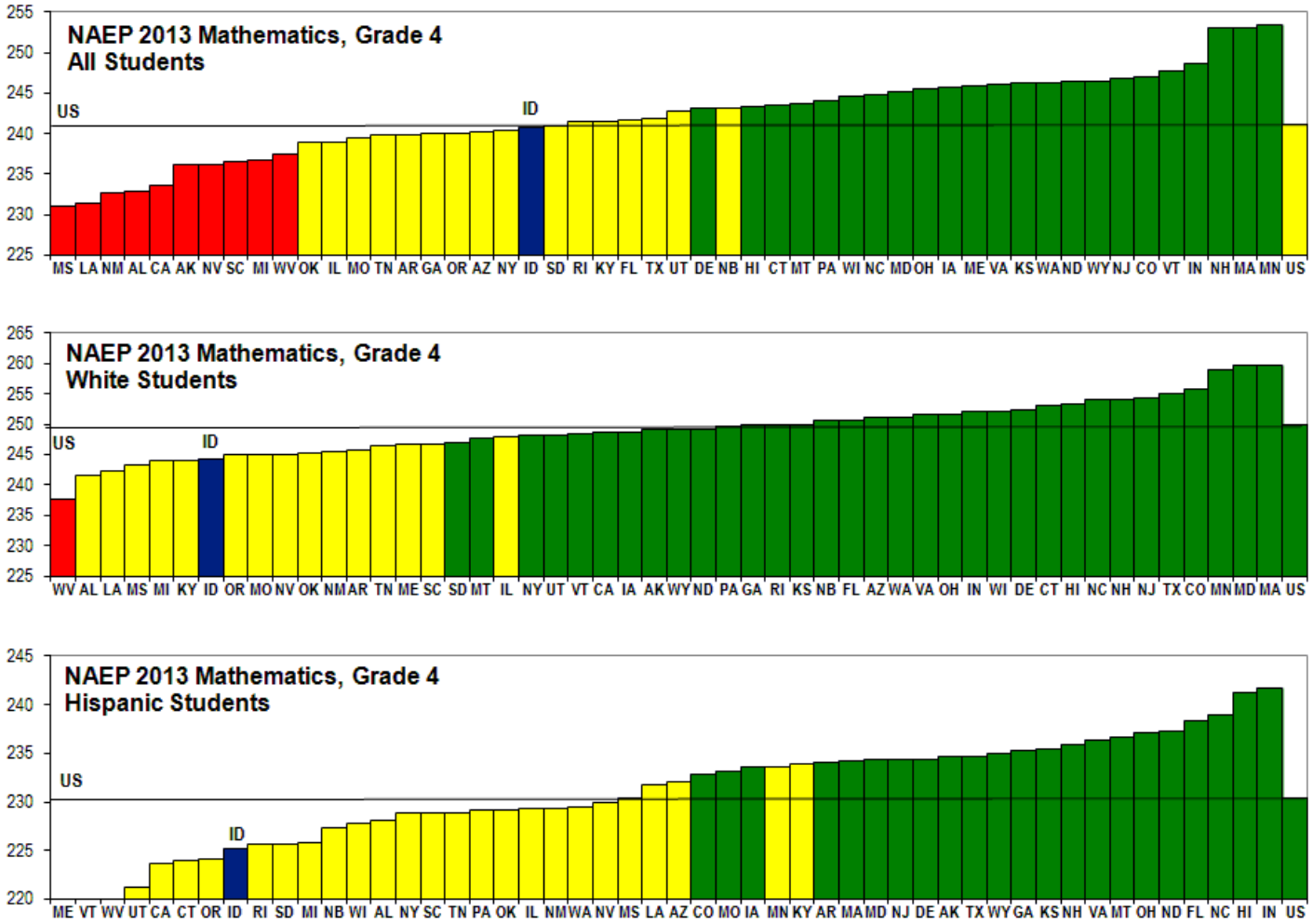
NAEP 2013 eighth-grade reading:

The All Student average score for Idaho was ***higher*** than the All Student national public school average.

The White Student average score for Idaho was ***not significantly different*** from the White Student national public school average.

The Hispanic Student average score for Idaho was ***not significantly different*** from the Hispanic Student national public school average.

Exhibit 5. Cross-state comparisons histograms on average scores from the NAEP 2013 mathematics assessment for each of three fourth-grade student demographic groups: All Students, White Students, and Hispanic Students.



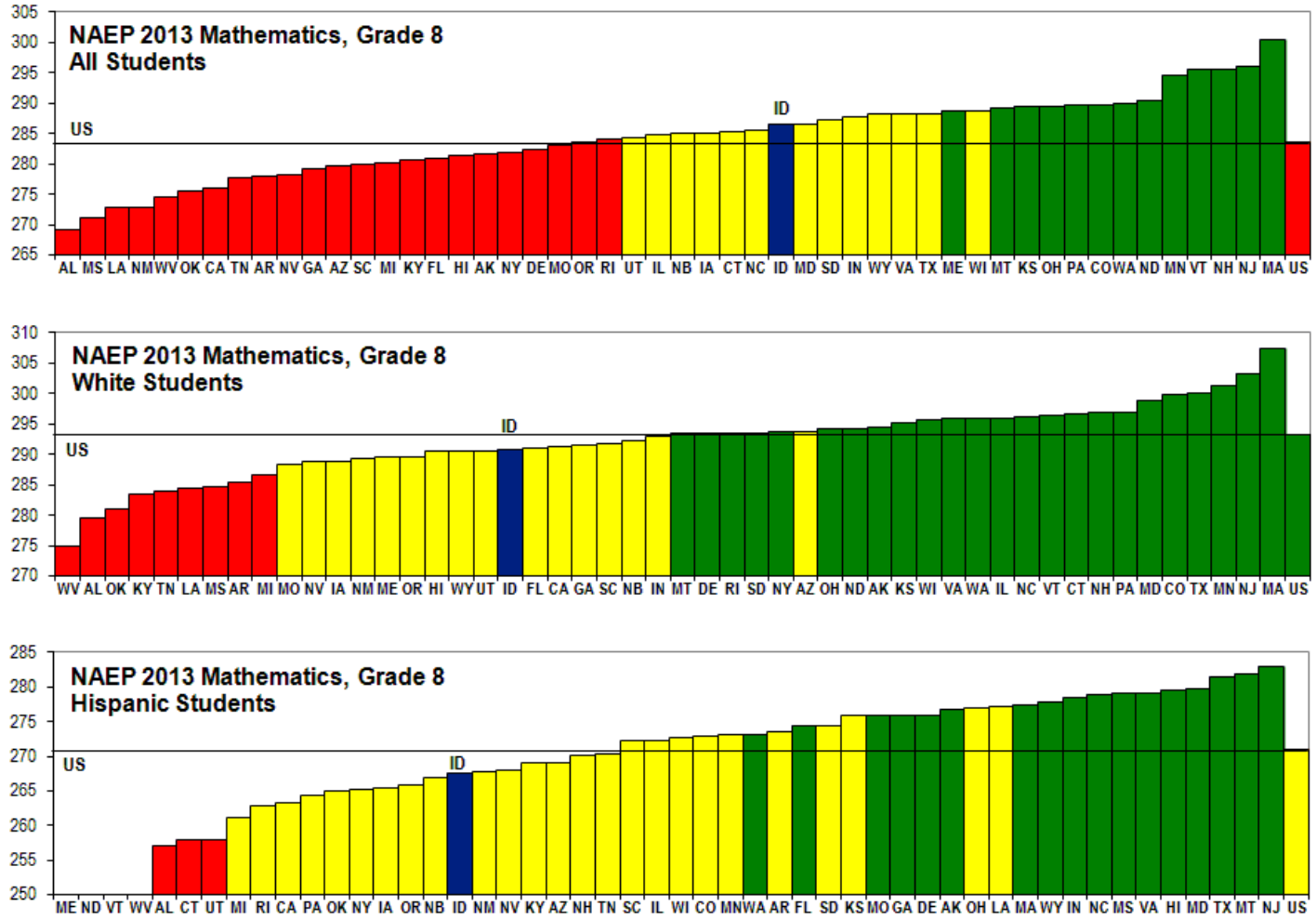
NAEP 2013 fourth-grade mathematics:

The All Student average score for Idaho was ***not significantly different*** from the All Student national public school average.

The White Student average score for Idaho was ***lower*** than the White Student national public school average.

The Hispanic Student average score for Idaho was ***lower*** than the Hispanic Student national public school average.

Exhibit 6. Cross-state comparisons histograms on average scores from the NAEP 2013 mathematics assessment for each of three eighth-grade student demographic groups: All Students, White Students, and Hispanic Students.



NAEP 2013 eighth-grade mathematics:

The All Student average score for Idaho was ***higher*** than the All Student national public school average.

The White Student average score for Idaho was ***lower*** than the White Student national public school average.

The Hispanic Student average score for Idaho was ***not significantly different*** from the Hispanic Student national public school average.

Approach Two: Idaho vs. National Public on Percentile Ranks

The percentile ranks for Idaho and national public demographic groups were compared for NAEP reading and mathematics in grades 4 and 8 from 2003 to 2013.

The students in the nation's public schools in NAEP 2003 were selected to serve as the norm groups for each grade-subject-demographic assessment. For example, there were three norm groups for fourth-grade reading, one for All Students, one for White Students, and one for Hispanic Students. In total, there were 12 grade-subject-demographic norm groups.

Percentile ranks were derived from using average scores and standard deviations in a two step process. The first step calculated a z-score for the reference or focus group using the equation:

$$\text{z-score} = \frac{focus\ group\ average - norm\ group\ average}{norm\ group\ standard\ devidation}$$

Then MS Office Excel spreadsheet functions calculated the percentile rank from the z-score using this equation:

$$\text{Percentile Rank} = TRUNC(100*NORMSDIST(z\text{-}score))$$

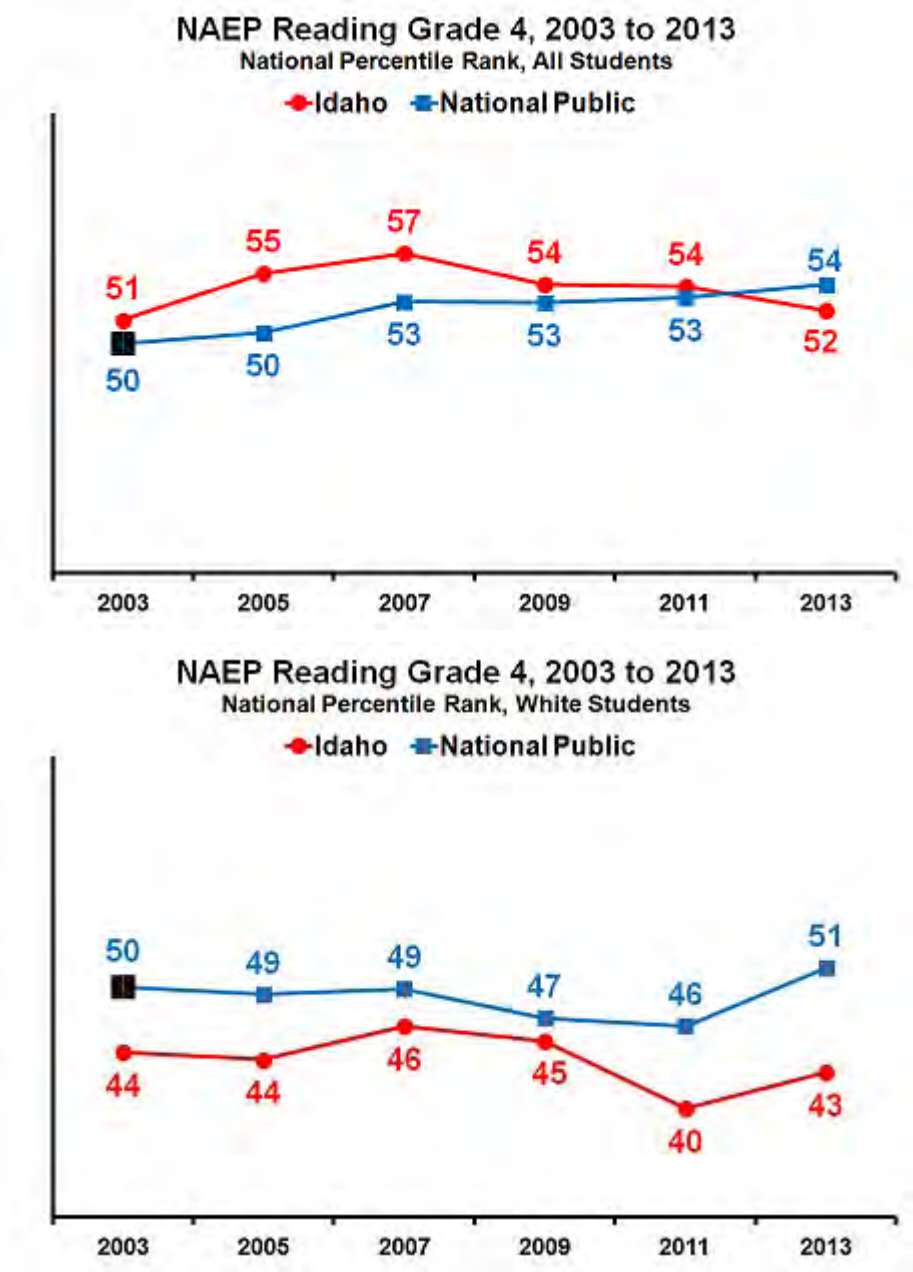
The 2003 norm groups, indicated by a black marker in the exhibits, all has a percentile rank = 50. The norm group percentile rank is the anchor against which all percentile ranks from 2003 through 2013 are compared, whether Idaho or national public. Percentile ranks are *effect size* statistics. As such, they were not submitted to statistical tests. The purpose of effect size statistics is to display the magnitude of differences.

A percentile rank of 65 tells us that the average student of this focus group scored higher than 65 percent of the students in the norm group. Narratives describing percentile ranks should always identify the grade, the subject, the year, and the demographic group. For example:

> On the NAEP 2007 fourth-grade reading test, Idaho's average White Student scored higher than 57 percent of the White Students in the 2003 norm group (i.e., White Students in the nation's public schools in 2003).

Exhibits 7 through 12 display the results from this examination of NAEP percentile ranks for Idaho and national public school students. Exhibits 7-10 display All Students and White Student results. Exhibits 11-12 display Hispanic Student results.

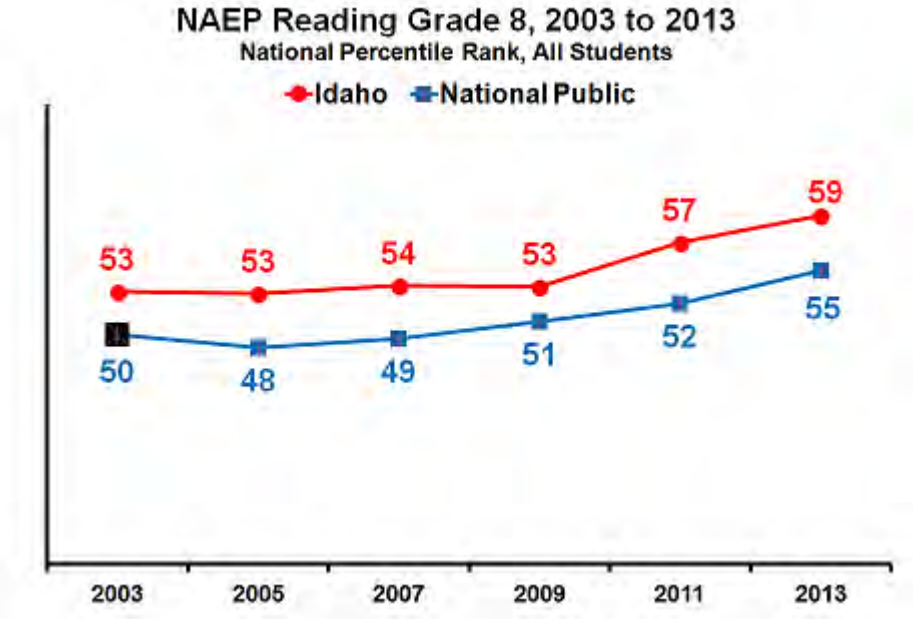
Exhibit 7. Idaho vs. National Public comparisons on percentile ranks from two student demographic group (All Students and White Students). NAEP assessment results from 2003-2013: NAEP Reading, Grade 4.

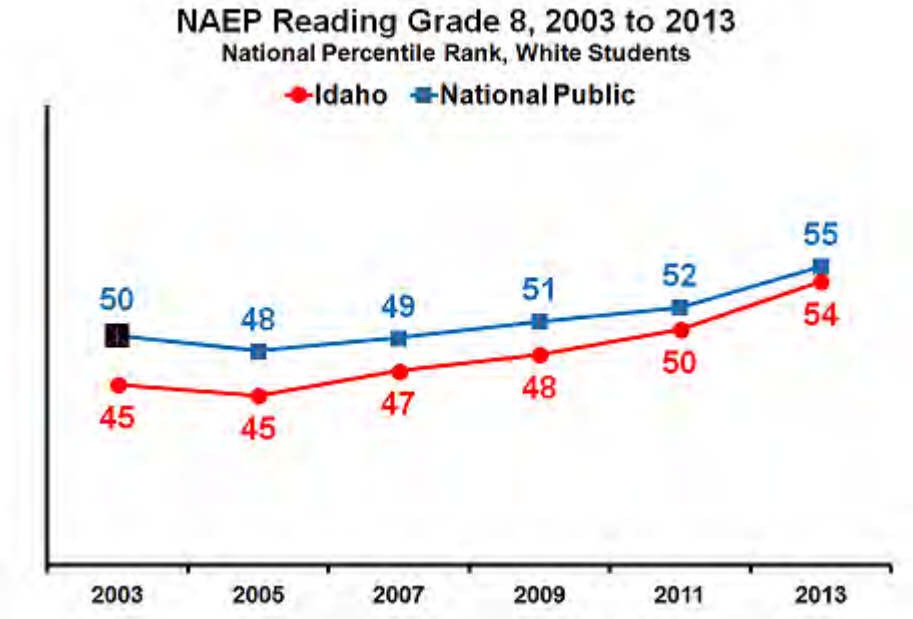


In the six grade 4 NAEP reading assessments from 2003 to 2013:

- All students in Idaho had a higher percentile rank than all students in the nation's public schools every year except 2013.
- White student in Idaho had a lower percentile rank than White students in the nation's public school all six assessments.

Exhibit 8. Idaho vs. National Public comparisons on percentile ranks from two student demographic group (All Students and White Students). NAEP assessment results from 2003-2013: NAEP Reading, Grade 8.

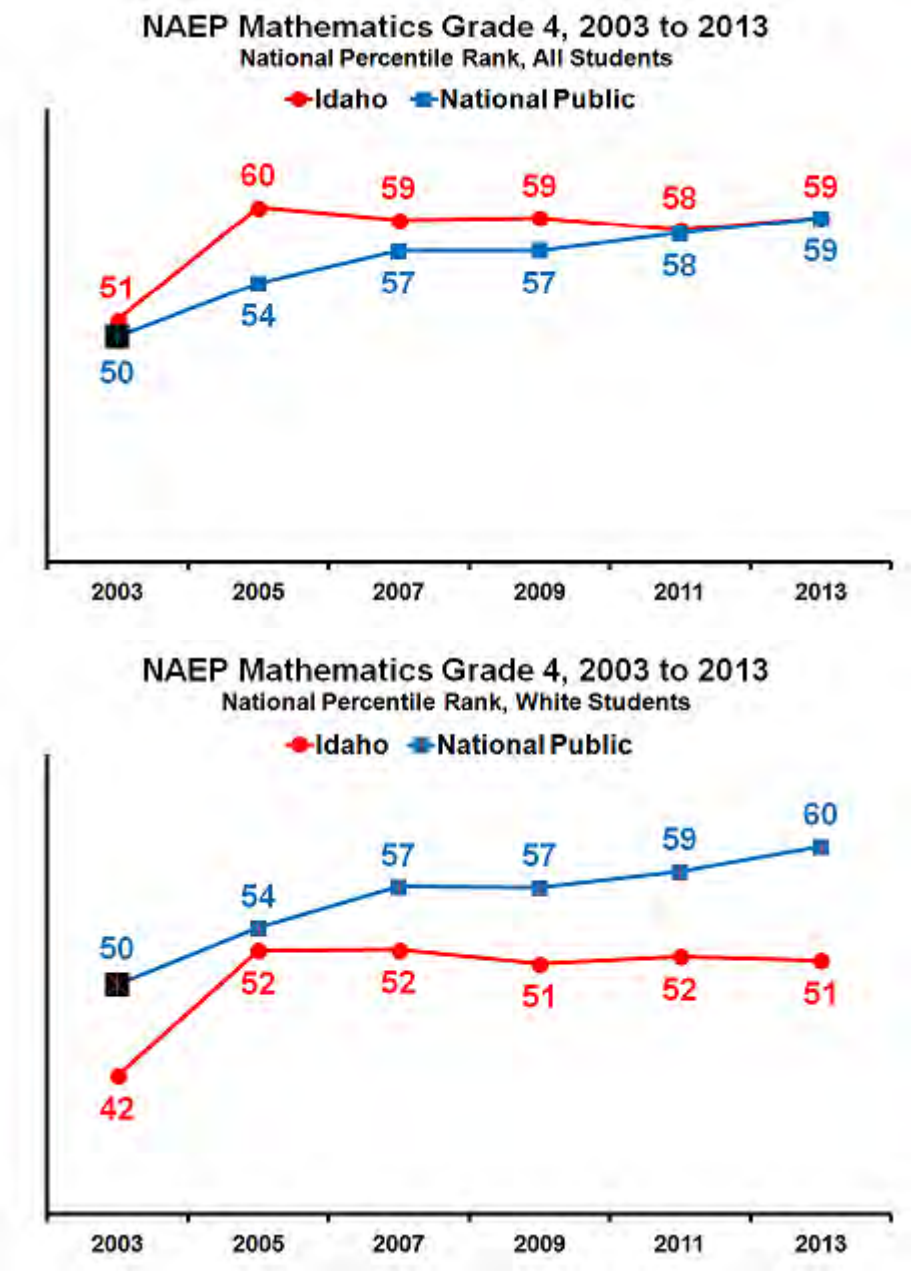




In the six grade 8 NAEP reading assessments from 2003 to 2013:

- All students in Idaho had a higher percentile rank than all students in the nation's public schools every assessment.
- White student in Idaho had a lower percentile rank than White students in the nation's public school all six assessments.

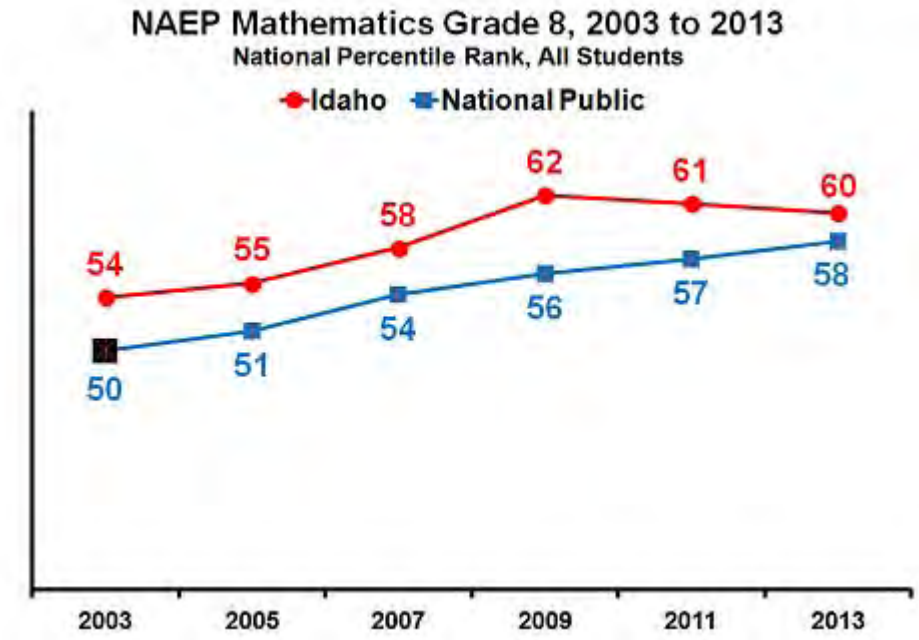
Exhibit 9. Idaho vs. National Public comparisons on percentile ranks from two student demographic group (All Students and White Students). NAEP assessment results from 2003-2013: NAEP Mathematics, Grade 4.

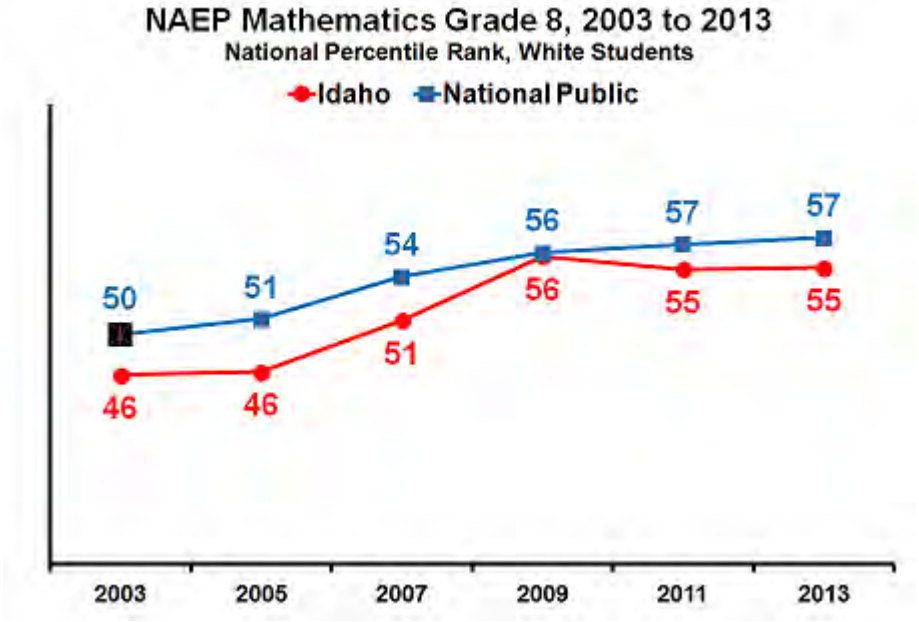


In the six grade 4 NAEP mathematics assessments from 2003 to 2013:

- All students in Idaho had a higher percentile rank than all students in the nation's public schools every assessment, except for "ties" in 2011 and 2013.
- White student in Idaho had a lower percentile rank than White students in the nation's public school all six assessments.

Exhibit 10. Idaho vs. National Public comparisons on percentile ranks from two student demographic group (All Students and White Students). NAEP assessment results from 2003-2013: NAEP Mathematics, Grade 8.

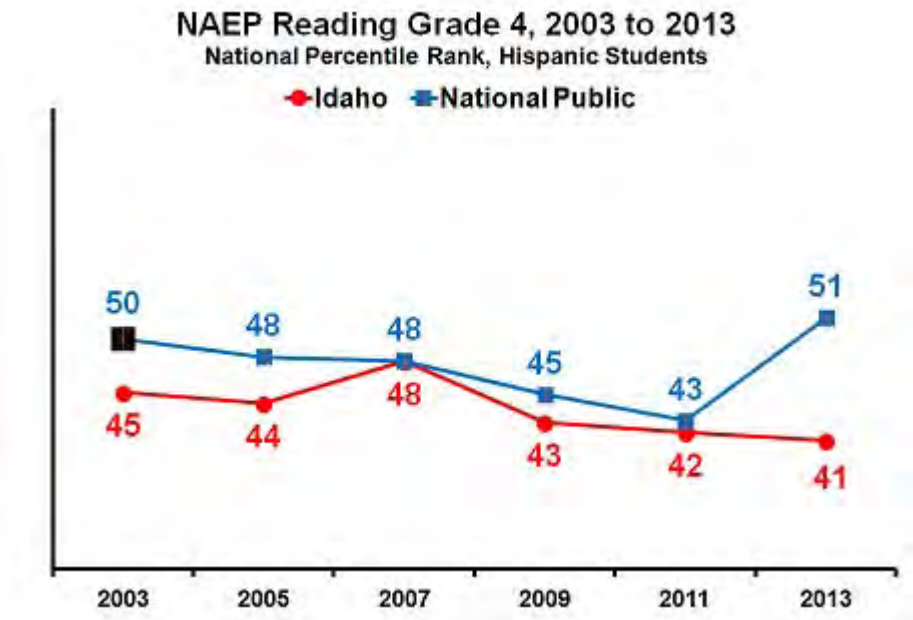


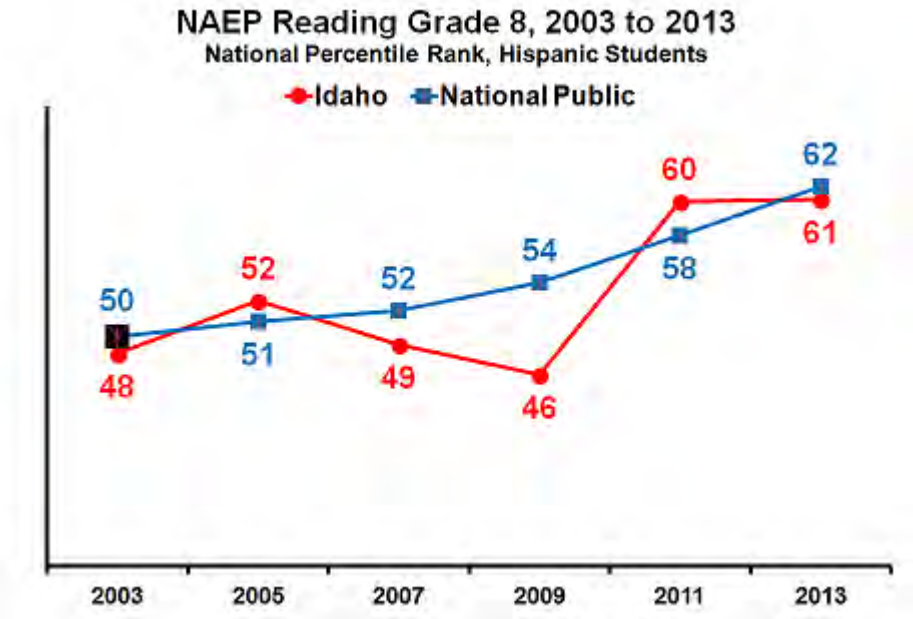


In the six grade 8 NAEP mathematics assessments from 2003 to 2013:

- All students in Idaho had a higher percentile rank than all students in the nation's public schools all six assessments.
- White student in Idaho had a lower percentile rank than White students in the nation's public school every assessment except 2009.

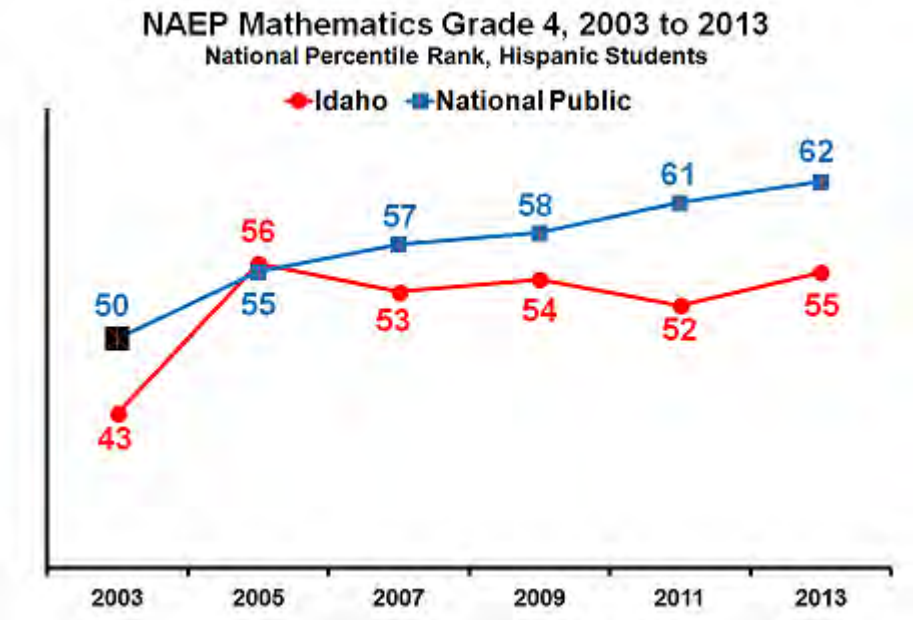
Exhibit 11. Idaho vs. National Public comparisons on national percentile ranks from the 2003-2013 NAEP assessments for Hispanic Students: NAEP Reading, Grades 4 and 8.

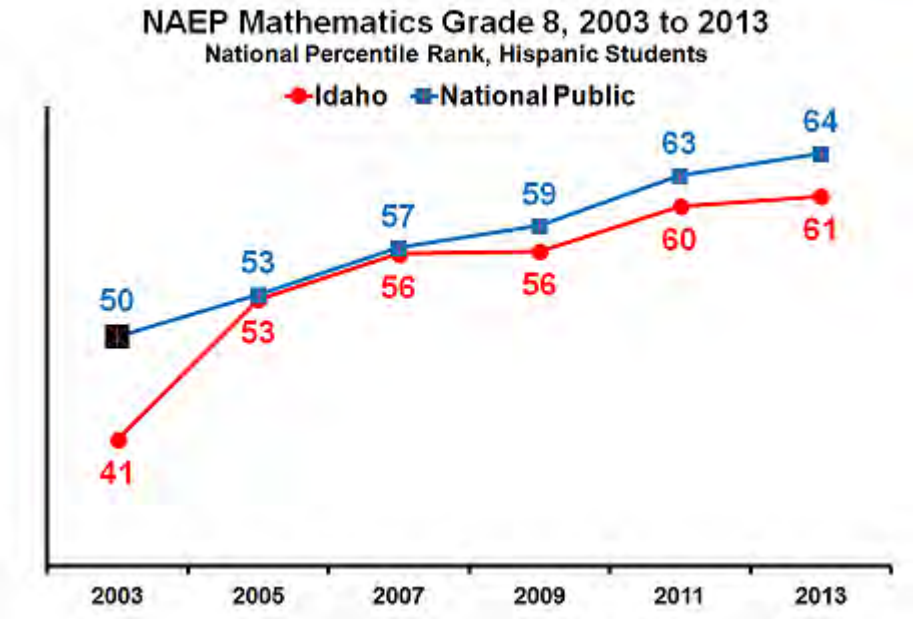




- On the six NAEP reading assessments for grade 4 from 2003 to 2013, Idaho Hispanic students had lower percentile ranks than Hispanic students in the nation's public schools every assessment except 2007, which was a "tie".
- On the six NAEP reading assessments for grade 8 from 2003 to 2013, Idaho Hispanic students had lower percentile ranks than Hispanic students in the nation's public schools on three assessments (2003, 2007, 2009 and 2013), but had higher percentile ranks on two assessments (2005 and 2011).

Exhibit 12. Idaho vs. National Public comparisons on national percentile ranks from the 2003-2013 NAEP assessments for Hispanic Students: NAEP Mathematics, Grades 4 and 8.





- On the six NAEP mathematics assessments for grade 4 from 2003 to 2013, Idaho Hispanic students had lower percentile ranks than Hispanic students in the nation's public schools every assessment except 2005, where Idaho Hispanics had a higher percentile rank.
- On the six NAEP mathematics assessments for grade 8 from 2003 to 2013, Idaho Hispanic students had lower percentile ranks than Hispanic students in the nation's public schools every assessment except for 2005, which was a "tie".

Note: A percentile rank is not the same thing as a percentage correct response. If on a 50 item test, 35 items had to be answered correctly, the passing score of 35 is 70 percent correct response. If 95 percent of the students scored at or above 35 on the test, then 5 percent scored below 35. Thus, for a score of 35, the percentile rank is 5 or the percent of students scoring below 35.

References

Bracey, G.W. (2004, February). Simpson's Paradox and Other Statistical Mysteries. *American School Board Journal*. Available at http://www.asbj.com/MainMenuCategory/Archive/2004/February

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

Pellegrino, J.W., Jones, L.R., and Mitchell, K.J. (Eds.). (1998). *Grading the Nation's Report Card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press. Available online: http://eric.ed.gov/?id=ED446096

Status of Achievement Levels. (2013). Washington D.C.: U.S. Department of Education, Institute for Education Science, National Center for Education Statistics. Retrieved December 6, 2013, from http://nces.ed.gov/nationsreportcard/achlevdev.aspx

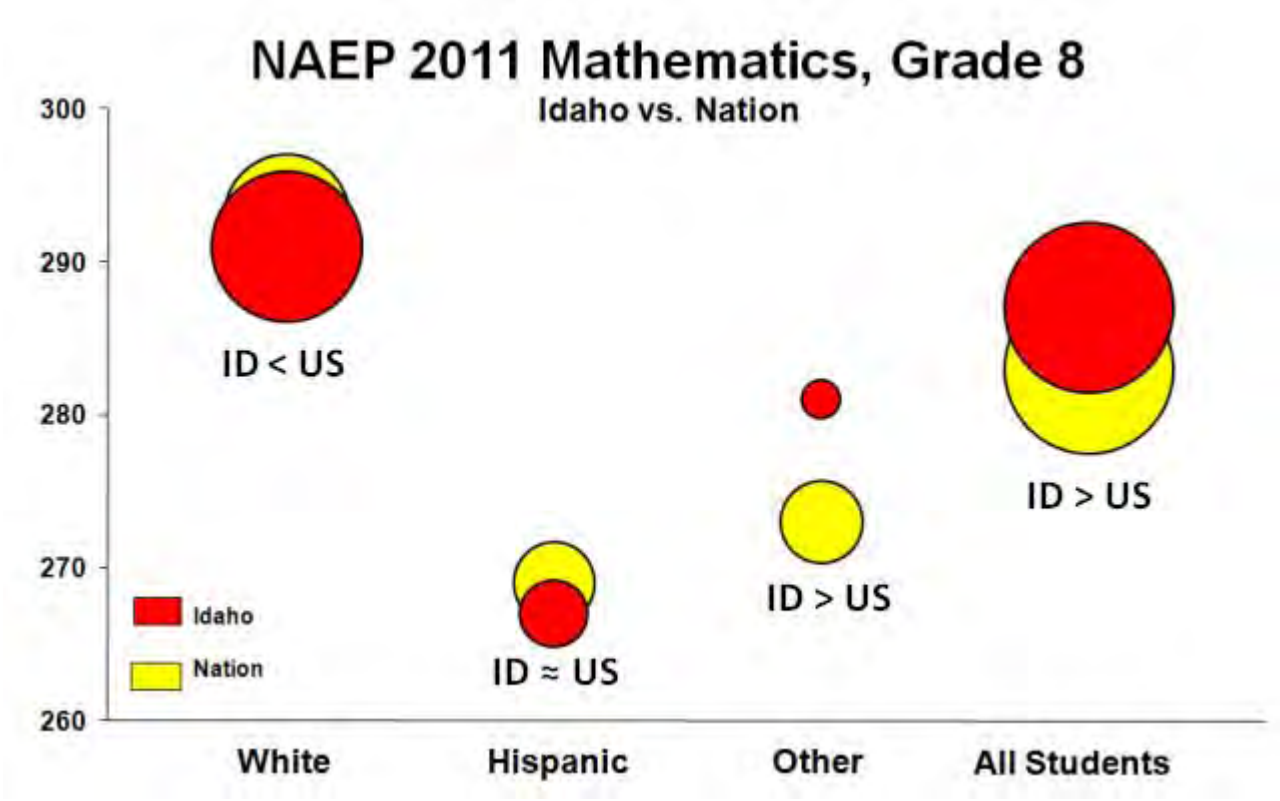
----- APPENDIX A -----

National Assessment of Educational Progress in Idaho

Simpson's Paradox in Idaho
NAEP 2011 Mathematics, Grade 8

Idaho Snapshot 2011

The Mathematics assessment of the National Assessment of Educational Progress (NAEP) used multiple-choice and constructed-response questions to examine student skills in number properties & operations, measurement, geometry, algebra & functions, and data analysis & probability. Mathematics scores range from 0 to 500, where 214 is *Basic* (meets grade 8 expectations), 249 is *Proficient*, and 282 is *Advanced*.



The "total score" for all students depends not only on the average score for each of the subgroups (center of circle), but also on the proportion of students in each subgroup (area of circle).

- On the Grade 8 NAEP mathematics test in 2011, Idaho's White students scored 291, lower than 293 for their national counterparts. White students made up 79 percent of Idaho's students, but only 54 percent of the nation's students.
- On the Grade 8 NAEP mathematics test in 2011, Idaho's Hispanic students scored 267, not statistically different from 269 for their national counterparts. Hispanic students made up 16 percent of Idaho's students, but 23 percent of the nation's students.
- On the Grade 8 NAEP mathematics test in 2011, Idaho's "Other" students scored 281, higher than 273 for their national counterparts. "Other" students made up 5 percent of Idaho's students, but 24 percent of the nation's students.
- On the NAEP mathematics test in 2011, Idaho's eighth-grade students (287) scored higher than eighth-graders in the nation's public schools (283).

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 NAEP Mathematics. Visit Idaho NAEP on the web at http://www.sde.idaho.gov/site/naep/