THOSE MISLEADING SAT AND NAEP TRENDS:
SIMPSON'S PARADOX AT WORK

Gerald W. Bracey

Each year when it releases the latest SAT results, the College Board trumpets the percentage of minority kids taking the test. That percent is now about one third of all takers. The Board does not tell people, and ought to, that the growth in minority testtakers renders the longitudinal trends of the national SAT average difficult to interpret.

The Board's omission is a bit difficult to understand given the 1977 report of the panel it convened to look at what was then a 14-year decline in the SAT national average. That panel attributed most of the change to changes in the demographics of who was taking the SAT--more women, more minorities, more students with mediocre high school records, more low family income students. All of those changes are associated with lower scores.

Even the "national average" in any one year is somewhat iffy since it currently includes 4% of the seniors in Mississippi, 23% of the seniors in Montana, 52% of the seniors in California and 83% of the seniors in Connecticut. Other states cover the range between the extremes of Mississippi and Connecticut. What such a hodge podge of percentages means is not easy to decipher. Even the College Board seems to recognize this. Through 2001 it called its release the "National Report;" in 2002 it changed the label to "Total Group Report." The iffiness of the trends increases when we look at changes in the proportion of the total testtaking pool represented by different ethnic groups.

Consider first, though, the overall trend for SAT scores and the trend for the various ethnic categories used by the College Board.

	1981*	2002	Gain	1981	2002	Gain
		Verbal		Mathematics		
White	519	527	+8	509	533	+24
Black	412	431	+19	391	427	+36
Asian	474	501	+27	512	569	+57
Mexican	438	446	+8	447	457	+10
Puerto Rican	437	455	+18	428	451	+23
American Indian	471	479	+8	463	483	+20
All	504	504	0	494	516	+22

What on earth is going on here? The increase in math scores for most groups exceeds, and sometimes far exceeds the gain for all students. The Verbal scores show an even more paradoxical outcome: All groups show an increase, but the gain for the whole group is exactly zero. Nil.

The operative word above is "paradoxical." What we have here is an instance of a paradoxical phenomenon so common in research it has a name: Simpson's Paradox. A google search on "Simpson's Paradox" results in 2800 hits.

Bracey, G.W. (2004). Simpson’s paradox and other statistical mysteries. American School Board Journal, 191 (2), 32–34.

To understand the paradox we must first look at changes in the ethnic composition of the SAT testtaking group over time. This is given below.

	1981		2002	
	#	%	#	%
White	719,383	85	698,659	65
Black	75,434	9	122,684	11
Asian	29,753	3	103,242	10
Mexican	14,405	2	48,255	4
Puerto Rican	7,038	1	14,273	1
American Indian	4,655	0	7,506	1
Total		100		92

(2002 percentages do not sum to 100% because of 8 percent responding "Latin American" or "Other," categories not used in 1981).

The source of the paradox is the changing composition of the SAT testtakers. Minorities now comprise a much larger proportion of the total than they did 20 years ago. And, except for the Mathematics scores of Asians, all minority scores, while rising, remain below the overall average. Adding more and more of these improving, but still low, scores attenuates the rise of the overall average. In the case of the verbal score, it attenuates it to zero.

Simpson's Paradox is stated in many ways. They all convey the idea that when subgroups' scores on a variable are aggregated into a single total, the total might show a relationship that is the reverse of the relationship seen in the subgroups. Hence, the paradox.

To assist understanding of Simpson's Paradox, let's examine a medical example uncovered by the google search the on Net. It showed the proportion of patients who survived or died during their hospital stay. Overall, the results looked like this:

	Survived	Died	Total	Survival Rate
Hospital A	800	200	1000	80%
Hospital B	900	100	1000	90%

Hospitals are dangerous places generally, but it looks like if you must check into one, Hospital B is you medical facility of choice. But what if we divide the patients into those who were in good condition prior to treatment and those who were in poor condition?

Bracey, G.W. (2004). Simpson’s paradox and other statistical mysteries. American School Board Journal, 191 (2), 32–34.

Good Condition Patients

	Survived	Died	Total	Survival Rate
Hospital A	590	10	600	98%
Hospital B	870	30	900	97%

Poor Condition Patients

Hospital A	210	190	400	53%
Hospital B	30	70	100	30%

Thus while both hospitals had the same survival rate for all patients, Hospital A treated a higher proportion of those who were in bad shape to start with and managed to keep a higher proportion alive. Hospital A is the place for you whether you are in good or poor condition or arrival there.

This example lets us see that Simpson's paradox can affect measures taken at just one time and aggregated for groups that differ in some important way (good condition, poor condition). It and can affect as well measures taken over time when the composition of the subgroups changes over that time (the changing ethnic makeup of the SAT testtaking sample).

Back in education, we see Simpson's paradox at work in NAEP trends as well in SAT trends.

Reading	1971	1999
Age 17	285	288
Age 13	255	259
Age 9	208	212

Over a period of 28 years, there is little overall change. Some commentators have used these numbers to criticize schools: Spending has increased ("soared," "skyrocketed," "mounted" are words commonly used by the critics), but test scores are "flat." ("stagnant," "sluggish," "static," choose your term). As with the SAT, though, looking at trends by ethnic group reveals something different:

Reading	White		Black		Hispanic	
	1971	1999	1971	1999	1975!	1999
Age 17	291	295	238	264	252	271
Age 13	261	267	222	238	232	244
Age 9	212	221	170	186	183	193

! Hispanics constituted too small a sample to generate a reliable estimate in the 1971 assessment.

Bracey, G.W. (2004). Simpson’s paradox and other statistical mysteries. American School Board Journal, 191 (2), 32–34.

The changes for white students pretty much mirror the changes for the whole sample.

The gains for black and Hispanic students, though, are much larger than for the entire group. However, their scores remain lower than whites and, by Simpson's paradox, because they are now a larger proportion of the total group, attenuate the gains seen when all groups are combined.

The proportion of whites in the sample falls from roughly 80% to roughly 70% (it varies slightly for different ages). The proportion of the entire group made up of blacks changes over time from about 14 percent to about 16%, while the proportion of Hispanics doubles from about five percent to about 10 percent). Asians were not represented as a separate group until the science assessment of 1996 and even in that year there was concern about the accuracy of the estimated scores.

Similar results are seen for NAEP assessments in mathematics and science as well as reading.

It sometimes appears as if test scores are falling when, in fact, test scores for all groups are rising at the same time as lower scoring groups are making up a larger proportion of the total. This, it should be obvious, does not mean the same thing as falling test scores due to declining achievement. It *should* be obvious, but it is often conveniently overlooked by school critics. Indeed, since some of these critics are statistically sophisticated, one must conclude that they overlooked Simpson's Paradox not only conveniently, but also deliberately.

-----

*(1981 is used as a starting point because it was the first year the Board published a document showing SAT data by gender and ethnicity. 1981 also marked the lowest point of the decline of average SAT scores that had begun in 1963. The Board category, Latin American, which covers Central and South American students, was not in use in 1981 and currently accounts for four percent of all SAT testtakers. They scored 458 on the Verbal in 2002 and 464 on the Math. Another four percent now check "other," also not used in 1981 and also account for 4 percent of the total. They scored 502 on the Verbal and 514 on the Math.).

Bracey, G.W. (2004). Simpson's paradox and other statistical mysteries. American School Board Journal, 191 (2), 32–34.