

THE CONTINUING USE AND MISUSE OF SAT SCORES

David W. Grissmer
RAND Corporation

The Scholastic Aptitude Test (SAT), the most widely publicized test taken by Americans, strongly influences public opinion about the quality of American schools. Unfortunately, the SAT scores not only have no statistical validity for tracking trends in the achievement of American students but actually show a perverse relationship to the trends in achievement as tracked by statistically valid scores. Thus the scores are quite misleading indicators of the quality of schools. The College Entrance Examination Board, a nonprofit institution established to foster excellence in education, clearly understands the statistical limitations of the SAT scores yet actively seeks annual publicity around the release of national, state, and school district SAT scores. This article questions whether the decision by the College Board to continue publishing aggregate SAT scores is in the public interest.

The Scholastic Aptitude Test (now called the Scholastic Assessment Test; SAT) occupies a unique place in our national consciousness. It is probably the most widely taken “high-stakes” test in the nation. It has been taken over the past 40 years by most students interested in applying for college admission—over one third of high school graduates. The score received has partially determined whether college admission is granted and which college one attends. Like many other high-stakes tests, the SAT has attracted criticism from many quarters. These criticisms focus on whether the test is equitable to all those taking it, whether it can validly predict success in college, and whether its use improves the college admission process (Crouse & Trusheim, 1983; Jacobs, 1995). All of these concerns focus on the appropriate interpretation and use of individual-level scores in the college admission process. Similar to other tests that are influential in critical decisions, remedies for issues concerned with test validity or perceived inequity are sought through the courts.

There is a unique issue in public policy associated with the publication and use of aggregate SAT test scores. The publication of aggregate score averages for schools, school districts, states, and the nation need not be done to carry out the primary purpose of the test: the sorting of individuals in the college admission process. However, annual SAT average scores have been routinely published at all levels since the 1960s. These scores receive widespread media coverage at every level because they provide easy and, in many cases, the only available comparative score data among local schools and school districts and among states. Because SAT scores are given at the end of K–12 education, they are often seen as providing indicators of the quality of the K–12 education system.

This widespread publicity and their interpretation mean that SAT scores frequently enter the public debate on education. Educational expenditures are the

Correspondence concerning this article should be addressed to David Grissmer, RAND Corporation, 1200 South Hayes Street, Arlington, Virginia 22202. Electronic mail may be sent to davidg@rand.org.

single largest public expenditure at the local and state level, and indicators such as the SAT can potentially influence elections, appointments of educational policymakers from school principals to boards of education, and educational policies. Perhaps as important, they can shape the attitude of taxpayers toward schools: Whether schools are improving or deteriorating, whether money is spent efficiently or inefficiently, whether educational expenditures should increase or decline, and how much support to provide for alternative educational policies.

Such use of SAT scores would be beneficial if the scores reflected accurately the quality of schools and school systems. Unfortunately, the aggregate SAT scores, at best, convey no useful information about educational quality and, at worst, convey highly misleading information about educational quality. Research beginning in the 1980s has pointed out the flaws in the aggregate SAT scores stemming from the ever changing, self-selected sample of students taking the tests.¹ Yet aggregate SAT scores continue to be published and widely misinterpreted as indicators of educational quality.

The use and misuse of aggregate SAT scores are issues outside the purview of courts or legislators and largely lie in the hands of the College Entrance Examination Board, in relation to their accountability as a nonprofit institution and their professed commitment to educational excellence. In this article, I suggest that perhaps the only near-term remedy to such widespread misuse of aggregate SAT score data and its pernicious effect on educational decision making is a reexamination by the College Board of whether the benefits of publishing such data outweigh the possible cost from flawed public policymaking. The K–12 education community might be the avenue to raise this issue. In the longer term, the SAT scores may be discredited due to comparison with the spreading and statistically valid tests being undertaken in most states.

What Do the Aggregate SAT Scores Measure?

Differences in SAT scores among schools, school districts, and states primarily reflect the different characteristics of the changing, self-selected population taking the test and very little about the quality of the schools attended. The changes in average scores over time also primarily reflect the ever changing characteristics of the self-selected pool of test takers. Even worse, the national trends actually move in the opposite direction to school quality. The better job the K–12 education system does in preparing more students to meet college standards, the lower will be the average SAT score.

The decline in SAT scores over the past 25 years provides an example of these perverse trends. The National Assessment of Educational Progress (NAEP) tests given to representative samples of 9-, 13-, and 17-year-old students show gains for all age groups in both reading and math from the early 1970s to 1990 (Campbell, Voelkl, & Donahue, 1997). These tests reflect the most accurate assessment of how student achievement is changing in the nation. The data also show that all racial–ethnic groups score higher at each age group in both reading

¹For instance see Gohmann (1988), Powell and Steelman (1984), Wainer (1986), and Mennard (1988).

and math but that the largest gains have been made respectively by Black students, Hispanic students, and lower scoring White students (Hedges & Nowell, 1998). Higher scoring non-Hispanic White students have made the smallest gains in scores. The cause of the large Black gains is still not settled, but a small part is likely attributable to improved family characteristics and a much larger part to some combination of changing school characteristics or motivational and attitudinal changes among Black parents, students, and their teachers associated with the national movements for equality of opportunity and war on poverty (Grissmer, Flanagan, & Williamson, 1998; Grissmer, Kirby, Berends, & Williamson, 1994).

During this period of rising national test scores for all ages and racial-ethnic groups in both math and reading, the SAT scores show significant declines. From 1971 to 1990, the NAEP verbal scores show a gain of 0.30 standard deviation whereas the SAT scores show a decline of 0.10 standard deviation. For Black students, the SAT scores show increases of 0.18 standard deviation whereas the NAEP scores show gains of 0.60 standard deviation. The discrepancy between scores of representative samples of 17-year-old students and the self-selected population of SAT test takers is very large, approximately equal to 40 SAT points or 14 percentile points on a national scale. This decline in SAT scores combined with the perceived doubling of "real" per-pupil expenditures has partially formed the basis of widespread perception of K-12 education decline and inefficiency over the past 25 years.²

There are three primary reasons why SAT scores show different trends than NAEP scores, all relating to the nonrepresentative and changing self-selected sample of SAT test takers. First, the proportion of seniors who have taken SAT tests has expanded. In the late 1960s, less than one out of three 17-year-olds had taken an SAT test, whereas by 1990, over 40% had taken an SAT test. This broadening of the pool has changed the characteristics from a more to a less elite group and lowered average scores (Advisory Panel on the Scholastic Aptitude Test Score Decline, 1977; Rock, 1987). Second, the composition of the pool also has changed in ways that would be expected to lower average scores further (Rock, 1987). In 1971, 88% of test takers were White compared with 73% in 1990.³ Women have increased from 49% to 53% of test takers. Black students score almost a full standard deviation below White students, and Hispanic students score approximately two thirds of a standard deviation below White students. Women score almost one half a standard deviation below men in math and one tenth of a standard deviation below men in verbal scores. Asian American students score approximately 0.4 standard deviation below White students in verbal scores. The only racial-ethnic compositional change that would boost scores is that Asian American students score about 0.3 standard deviation above White students in math scores. The racial-ethnic compositional changes have slowed from the late 1970s, as have the changes in the average SAT scores.

²Recent research also has shown that the doubling of real expenditures in education is equally misleading. Increases in real expenditures to education that were directed toward improving academic achievement of regular students are closer to 33% rather than 100% (Ladd, 1996; Rothstein & Miles, 1995).

³The nonresponse rate on the racial-ethnic question has ranged from 20% to 40% of test takers, so there is even considerable uncertainty about the compositional shifts.

Besides the better known effects of an expanding pool of test takers and its changing composition, a third flaw is present in interpreting changing SAT scores as indicative of deteriorating quality of education. The largest score gains between the early 1970s and 1990s occurred among the non-college-bound senior population, those not included in the SAT pool. As mentioned above, NAEP data show that minorities and lower scoring White students made the largest gains and that higher scoring White students made little or no gains in scores. Thus, SAT scores, taken predominantly by higher scoring students, reflect that portion of the student population making the smallest gains from 1970 to 1990.

Research also has shown that differences in average SAT scores at the state level primarily reflect different participation rates among students in the state (Powell & Steelman, 1984, 1996). State participation rates vary from less than 5% to over 80% of seniors. These differences in participation do not reflect only differing abilities to succeed in college because some of the states with the highest NAEP scores have the lowest SAT participation rates.⁴ Comparisons of SAT scores below state levels, at school district or school levels, will suffer not only from different participation rates, but from smaller sample sizes (Fetler, 1991). Thus virtually no educational meaning can be attached to changes or differences in SAT scores at any level of aggregation.

Any aggregation of SAT test scores above the level of the individual student—by high school, school district, state, or the nation—is simply uninterpretable as a measure of student achievement trends or as a measure of quality among schools, school districts, or states. The question is why scores continue to be published, receive continuing publicity, and regularly are cited as evidence of school progress or deterioration in the research community, in political debate, and in the press, especially given the availability and statistical validity of the NAEP data. Social psychology may provide some answers to this phenomenon.

Do Aggregate SAT Scores Have Influence?

Do people use SAT scores to assess the quality of schools? Since 1974, an annual Gallup poll has asked adults to grade the schools in their communities. Figure 1 shows the movement of the percentage of adults giving schools in their community a grade of A or B versus movement of SAT scores.⁵ These data are certainly consistent with the hypothesis that SAT scores influence the belief of citizens about their schools.

If people use SAT scores as a basis for judging the quality of schools and SAT scores do not reflect actual school quality, then there should be a variance in people's beliefs about school quality depending on the availability and quality of sources of information about schools. The annual Gallup poll of adults concerning education also has asked respondents since 1985 to grade schools nationwide,

⁴The states of Iowa, North Dakota, South Dakota, Minnesota, Nebraska, Wisconsin, and Oklahoma score above state averages on NAEP tests but have among the lowest SAT participation rates of less than 10%.

⁵The graph normalizes both variables to a mean of zero. The regression fit for the equation, school grade = A + B (average SAT score) gives B = .79, $t(21) = 5.2$, $R^2 = .56$, $p < .10$.

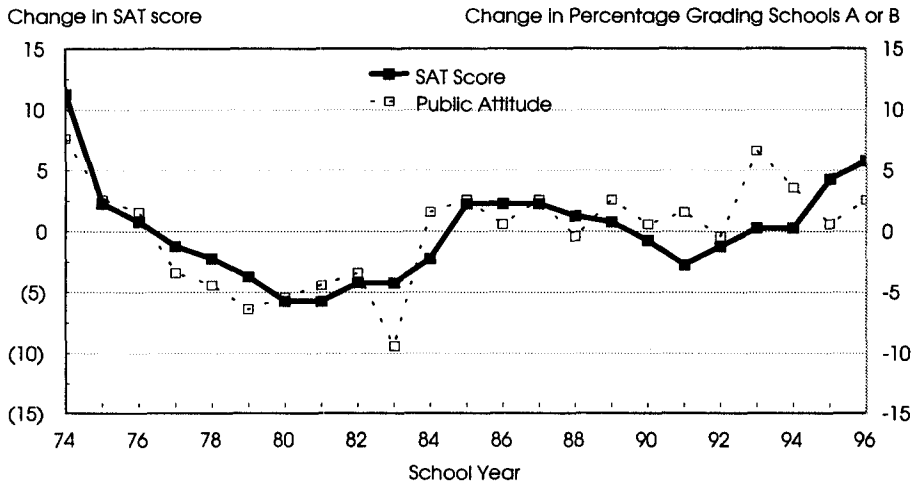


Figure 1. Comparison of trends in Scholastic Aptitude Test (SAT) scores with percentage of adults giving schools a grade of A or B.

schools in their community, and schools attended by their children (for parents of school-age children only). Table 1 contains the results for 1995.

The results, which have shown consistent gaps since all questions were asked in 1985, show that schools nationwide receive the lowest grade of both adults and parents of school-age children, whereas schools in the community and neighborhood score significantly higher grades and parents give the schools their children attend the highest grades.⁶ Alternate sources of information about local and national schools appear to be a likely source of this divergence. The more dependence there is on firsthand information about schools, the higher is the rating.

The primary problem with aggregate SAT scores is that consumers of such scores often assume that comparisons over time or among schools, school districts, or states have validity as an indicator of the quality of education. Theories regarding how people make inferences concerning statistical data suggest why SAT scores might exert undue influence despite their lack of validity. For example, Nisbett and Ross (1980) have reviewed evidence showing that people make inferential judgments from data that are more salient, vivid, emotionally interesting, and frequently reported than from data that are more statistically accurate but not as widely reported.

Because the SAT tests have been taken by 30% 45% of high school graduates annually for over 40 years and the results are quite critical to the college admissions process, these tests have much greater exposure and leave vivid impressions on students and parents alike. They are often reported several times a year in different forms: national results, state results, school district, and school results. In addition, local school scores are often routinely provided in real estate offerings and used as a basis for choosing neighborhoods and school districts.

⁶A regression of the parent rating for the schools their children attended shows a negative and insignificant coefficient for the regression in footnote 5.

Table 1
Percentage of Respondents Giving Schools an A or B Grade

Schools	All adults	Parents of school-age children
National	20	18
Community	41	49
Attended by children in your neighborhood	48	52
Attended by oldest child		65

Note. From "Of the Public's Attitudes Toward Public Schools," by S. M. Elam and L. C. Rose, September 1995, *Phi Delta Kappan*, 77(1), pp. 42–43. Copyright 1995 by the *Phi Delta Kappan*. Reprinted with permission.

In contrast, the NAEP tests, which provide a more statistically accurate picture of test score trends, have been taken approximately every 2 to 4 years by small samples of American students and have virtually no impact on the lives of individual students who take them or their parents. Thus, it is not surprising that people tend to rely on the more visible and more salient SAT results rather than the NAEP scores.

Research also indicates that mixed evidence, for example, evidence that NAEP scores are moving in an opposite direction from the SAT scores, often results in stronger, not weaker, trust in the originally held belief (Nisbett & Ross, 1980). This is partly because people tend to select and read information that agrees with prior expectations. During the 1980s especially, the perception of failing American education was widely reported in *A Nation at Risk* (National Commission on Excellence in Education, 1983). Press reports have also undoubtedly focused more on negative reporting on schools. As such, the more frequently reported SAT scores have tended to reinforce other sources of national information, whereas the less frequently reported NAEP scores might easily be dismissed partly because their source was the Department of Education, a much maligned agency during the 1980s.

The potential damage from public opinions based on SAT performance occurs if individuals believe that score differences among schools, school districts, or states reflect differences in educational quality or that changing scores over time reflect changing quality of education. Nisbett and Ross (1980) also have suggested that such naive inferences are consistent with evidence about how people form such inferences. In particular, people have strong tendencies toward single-cause explanations and tend to choose those that resemble the effect. Thus, the commonly held association between schools and test scores would lead to naive judgments such as declining test scores being the result of declining school quality.

Even in statistically valid samples of test scores, studies of achievement repeatedly show that family and demographic characteristics explain most of the variance in test scores having much stronger effects on scores than differences in schools or teachers.⁷ In the best of circumstances, it takes sophisticated statistical analysis to separate school and family effects. The SAT adds another significant

⁷A long line of research starting with the Coleman report shows family characteristics to be the

source of variance, a self-selected and changing sample, to try and separate out. The only possible way to do this would be for the SAT to collect extensive family- and schooling-related data. What data are collected with the SAT have high levels of refusal and miss most of the key variables that would be necessary to make such an analysis possible.

Why Are Aggregate SAT Scores Published?

The SAT is administered, analyzed, and published by the College Entrance Examination Board, a nonprofit organization established by colleges and universities to "help students succeed in the transition from school to college" (Press release, College Entrance Examination Board, February 14, 1997). The purpose of the SAT is to improve the college admissions process by providing scores that are comparable across individual students. As long as comparisons are restricted to individual students, the test may provide useful information about students applying to college and result in improving the college admission process.⁸

The College Board also publishes and seeks publicity for the aggregate test scores despite its own acknowledgment of the flaws in the scores. Starting in the 1980s, some of their publications began to have warnings attached. With respect to comparing aggregate scores, a warning states the following:

Aggregated data should not be used to compare or evaluate teachers, schools, districts or states. (Educational Testing Service [ETS], 1991, p. iii)

Their attitude toward comparing year-to-year scores is much more ambiguous:

Because the population of SAT takers is relatively stable from year to year, useful comparisons over time can be made among subgroups of the test taking population. . . . Year to year educational and demographic changes in this population along with changes in test performance could be of interest to the public, educators and educational policymakers at all levels. (ETS, 1991, p. iii)

Year-to-year changes in SAT scores contain no information about educational quality, nor do longer term trends, and releasing them to the public only invites misinterpretation. Not only are the scores published but publicity is sought around their release. So, on the one hand, the College Board acknowledges the scores cannot be used for the purpose the public uses them for, and on the other hand, it actively seeks widespread diffusion of the data.

The motivation for publishing such scores can only be speculated about. Although the aggregate data might have some value to admissions offices, access can be provided without seeking national publicity. One possible reason for publication is inertia and lack of knowledge about the impact and use of the aggregate scores. In this case, it might be an easy decision to quit publishing these scores. However, the motivation is probably more complicated. Like most all nonprofit organizations, The College Board seeks revenue to survive and revenue growth to expand. The SAT administration is the source of most revenue to the

dominant influence in explaining test score variance (see Coleman et al., 1966, Coleman & Hoffer, 1987; Jencks et al., 1972, and more recently, Gamoran, 1996; and Raudenbush, 2000).

⁸Some research challenges even this use of SAT scores (see Crouse & Trusheim, 1983).

College Board. So the publicity around SAT scores keeps the scores in the public eye. In this case, terminating the annual release of scores could be more difficult.

In the absence of voluntary action by the College Board, there are other avenues that might result in addressing the SAT problem. The organizations that have the largest stake in preventing misleading indicators of the quality of K–12 education are those that represent the major education constituencies. These include the National Governors Association, the National Council of the Chief of State School Officers, the National Association of Boards of Education, the Education Commission of the States, and the various teachers organizations. These organizations have been strangely silent on SAT scores despite their apparent damage to public opinion about schools. Concerted action on their part might be effective.

Another possible scenario is that the flaws in the SAT scores will become widespread, in which case the credibility of the the College Board and the SAT scores will suffer. Not only are the NAEP scores receiving more publicity but states are moving to widespread annual testing of students to track trends in scores and measure differences across schools and school districts. These state test scores have statistical validity and are widely publicized within states and may increasingly show differences with SAT scores. Particularly if the current wave of school reform is successful and state scores move upward while SAT scores move downward or stay stable, then the flaws in the SAT scores will become more widespread and ultimately damage the credibility of the SAT and its sponsor. It may then be in the College Board's self-interest to terminate publication of aggregate scores to prevent this.

Conclusion

An unfortunate fact is that the public perception of school quality is partly shaped by the ever available but flawed SAT scores. SAT scores can strongly influence public perceptions because they are more familiar, repeated frequently, have salience to people's lives, and often support existing opinions, even though the data have no statistical validity. Any reporting of aggregated unadjusted scores across schools, districts, states, or the nation appears not only to serve no useful public purpose but also to contribute to misleading impressions about K–12 schools and students and detract from an essential and well-informed public debate about our schools and students. In the absence of voluntary consideration by the College Board of whether publishing the aggregate scores is consonant with their commitment to "achieve Educational excellence for all Students" (Stewart, 1997–1998), concerted action by the K–12 education community could possibly be effective. However, the increasing prevalence and publicity of statistically valid tests at the national and state level may slowly result in widespread discrediting of the SAT scores.

References

- Advisory Panel on the Scholastic Aptitude Test Score Decline. (1977). *On further examination*. New York; College Entrance Examination Board.
- Campbell, J., Voelkl, K., and Donahue, P. (1997). *NAEP 1996: Trends in academic*

- progress* (NCES 97-985), Washington, DC: National Center for Education Statistics/Educational Testing Service.
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.
- Coleman, J. S., & Hoffer, T. (1987). *Public and private high schools: The impact of communities*. New York: Basic Books.
- College Entrance Examination Board. (1997, February 14). [Press Release]. Princeton, NJ: Author.
- Crouse, J., & Trusheim, D. (1983). *The case against the SAT*. Chicago: University of Chicago Press.
- Educational Testing Service. (1991). College bound seniors: 1991 profile of SAT and achievement test takers. Princeton, NJ: Author.
- Educational Testing Service. (1994). College bound seniors: 1994 profile of SAT and achievement test takers. Princeton, NJ: Author.
- Elam, S. M., & Rose, L. C. (1995, September). Of the public's attitudes toward public schools. *Phi Delta Kappan*, 77(1), 41–56.
- Fetler, M. E. (1991, Summer). Pitfalls of using SAT results to compare schools. *American Educational Research Journal*, 28, 481–491.
- Gamoran, A. (1996). Student achievement in public magnet, public comprehensize, and private city high schools. *Educational Evaluation and Policy Analysis*, 18, 1–18.
- Gohmann, S. F. (1988). Comparing state SAT scores: Problems, biases and corrections. *Journal of Educational Measurement*, 25, 137–148.
- Grissmer, D. W., Flanagan, A., & Williamson, S. (1998). Why did Black test scores rise in the 1970s and 1980s? In C. S. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 182–228). Washington, DC: Brookings Institution.
- Grissmer, D. W., Kirby, S. N., Berends, M., & Williamson, S. (1994). *Student achievement and the changing American family*. Santa Monica, CA: RAND Corporation.
- Hedges, L. V., & Nowell, A. (1998). Group differences in mental test scores: Mean differences, variability, and talent. In C. S. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 149–181). Washington, DC: Brookings Institution.
- Jacobs, W. R. (1995, Winter). Is the SAT Fair? *Journal of College Admissions*, 146, 22–31.
- Jencks, C. S., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Hayns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York: Basic Books.
- Ladd, H. F. (1996). Introduction. In H. F. Ladd (Ed.), *Holding schools accountable*. Washington, DC: Brookings Institution.
- Mennard, S. (1988, September). Going up, going down: Explaining the turnaround in SAT scores. *Youth and Society*, 20, 3–28.
- National Commission on Excellence in Education. *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.
- Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.
- Powell, B., & Steelman, L. C. (1984, November). Variations in state SAT performance: Meaningful or misleading. *Harvard Educational Review*, 54, 389–412.
- Powell, B., & Steelman, L. C. (1996, Spring). Bewitched, bothered and bewildering: The use and misuse of state SAT and ACT scores. *Harvard Educational Review*, 66, 27–59.
- Raudenbush, S. W. (2000). Synthesizing results from the trial state assessments. In D. Grissmer & M. Ross (Eds.), *Analytic issues in the assessment of student achievement*. (NCES 2000-050). Washington, DC: National Center for Education Statistics.

- Rock, D. A. (1987). "The score decline from 1972 to 1980: What went wrong? *Youth and Society*, 18, 239–254.
- Rothstein, R., & Miles, K. H. (1995). *Where's the money gone? Changes in the level and composition of education spending*. Washington DC. Economic Policy Institute.
- Stewart, D. M. (1997–1998). Report from the President. *1997–98 Annual Report*. New York: The College Board.
- Wainer, H. (1986, Spring). Five pitfalls encountered while trying to compare states on their SAT scores. *Journal of Educational Measurement*, 239, 69–81.

Received September 10, 1998

Accepted September 10, 1998 ■