High Stakes, but Low Validity? A Case Study of Standardized Tests and Admissions into New York City Specialized High Schools

Joshua Feinman, Ph.D.

October 2008

EPRU | EDUCATION POLICY RESEARCH UNIT

Education Policy Research Unit
Division of Educational Leadership and Policy Studies
College of Education, Arizona State University
P.O. Box 872411, Tempe, AZ 85287-2411
Telephone: (480) 965-1886
Fax: (480) 965-0303
E-mail: epsl@asu.edu
http://edpolicylab.org

Education and the Public Interest Center
School of Education,
University of Colorado
Boulder, CO 80309-0249
Telephone: (303) 447-EPIC
Fax: (303) 492-7090
Email: epic@colorado.edu
http://epicpolicy.org

• Suggested Citation:
Feinman, J. (2008). *High Stakes, but Low Validity? A Case Study of Standardized Tests and Admissions into New York City Specialized High Schools.* Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved [date] from http://epicpolicy.org/publication/high-stakes-but-low-validity

EPIC/EPRU policy briefs are peer reviewed by members of the Editorial Review Board. For information on the board and its members, visit: http://epicpolicy.org/editorial-board

High Stakes, but Low Validity?
A Case Study of Standardized Tests and Admissions into New York City Specialized High Schools[1]

Joshua Feinman

Executive Summary

This is a study of the admissions process at a select group of New York City public high schools. It offers the first detailed look at the admissions practices of this highly regarded and competitive group of schools, and also provides a window into the broader national debate about the use of standardized tests in school admissions. According to New York State law, admission to these schools must be based solely on an exam. The exam used is called the Specialized High Schools Admissions Test (SHSAT). This study makes use of the individual test results from 2005 and 2006.

Several key findings emerge:

- The SHSAT has an unusual scoring feature that is not widely known, and may give an edge to those who have access to expensive test-prep tutors. Other reasonable scoring systems could be constructed that would yield different results for many students, and there is no evidence offered to support the validity of the current system.
- Thousands of students who are not being accepted have scores that are statistically indistinguishable from thousands who are granted admission. And these estimates are derived using the less precise, classical-test-theory-based measures of statistical uncertainty, which may understate the problem. The New York City Department of Education (NYCDOE) fails to provide the more accurate, item-response-theory-based estimates of the SHSAT's standard error of measurement (SEM) near the admission cutoff scores, which would offer a clearer picture of how well the test is able to differentiate among students who score close to the admission/rejection line. This omission violates generally-accepted testing standards and practices.
- Students who receive certain versions of the test may be more likely to gain admission than students who receive other versions. No evidence is offered on how accurate the statistical equating of different test versions is. The mean scaled scores vary across versions much more than would be expected given the chance distribution of ability across large random samples of students, suggesting that the scoring system may not be completely eliminating differences among test versions.

- No studies have ever been done to see if the SHSAT is subject to prediction bias across gender and ethnic groups (i.e., if SHSAT scores predict things for different groups).

Of course, no test is "perfect." All face difficulties distinguishing among close candidates. The same is true of other potential admissions criteria, such as grades, which is a key reason why it is contrary to professional testing standards and practice to use any single metric as the sole criterion for admission. Since uncertainty and imprecision are inherent in all potential admissions criteria, standard psychometric practice is to choose the criteria that minimize this uncertainty. This is generally done by conducting predictive validity studies—studies designed to measure how well potential admissions criteria correlate with specific, quantifiable objectives (like future student performance). Predictive validity studies are regularly carried out for tests like the SAT and for high school grades, to help test-makers refine the test, and to help colleges decide how much weight to put on SAT scores, grades, and other factors in the admissions process. Overwhelmingly, these studies have found that multiple imperfect criteria, used in tandem, provide better insight into future student performance than a single imperfect criterion. Indeed, it's partly because of results from these validity studies that virtually all educational institutions use multiple admissions criteria.

The admissions procedures at the New York City specialized high schools violate this standard and run counter to these practices. Worse, in all the years the SHSAT has been the lone determinant of admission to these schools, the NYCDOE has never conducted a predictive validity study to see how the test was performing. In addition, it has never been made clear what the objectives of the SHSAT are. Absent predictive validity studies, there's no way to know if any test is providing useful information; and without well-specified objectives, it's not even clear what the test is supposed to do or predict. The whole process flies in the face of accepted psychometric standards and practice, and reminds us why those standards and practices were established and should be maintained. The thousands of students who apply to these select high schools deserve a properly tested system of determining who gets access to these prestigious and potentially life-changing educational experiences.

The foregoing findings give rise to the following recommendations:

- Formal predictive validity studies of the SHSAT need to be carried out. At a minimum, these studies should look at the ability of SHSAT scores (separate verbal and math) and middle school grades to predict high school performance. They should also test for prediction bias across gender and ethnic groups. The NYCDOE should release details on how the scaled scores are derived from item response theory—

particularly IRT-based estimates of the uncertainty surrounding scores near the admission cutoffs—and on the accuracy of the equating of different test versions. Any inadequacies in equating across test versions need to be corrected.

- Based on the results of these studies and in keeping with generally accepted psychometric standards and practices, a determination should be made as to what admissions process—including such areas as the scoring system, other criteria considered, and weights of these criteria—is most likely to achieve a specific, quantifiable admissions goal in a transparent, equitable way.
- If this study concludes that it is best to use additional admissions criteria besides a standardized test, the New York State law—which says that admissions to these schools must be based solely on a test—would need to be changed.
- Findings such as those presented in this study, and the particular choices of admissions procedures for these schools, should be discussed and deliberated in New York, and an informed decision should be made about future practices. Whatever admissions procedures are established, all applicants should know their implications.
- These findings should also be disseminated so that they can contribute to the broader national debate on standardized tests, school admissions, and high-stakes testing such as exit exams.

High Stakes, but Low Validity?
A Case Study of Standardized Tests and Admissions into New York City Specialized High Schools

Joshua Feinman

Introduction

Every year, the New York City Department of Education (NYCDOE) offers a competitive exam to determine admission to a select group of "specialized" public high schools. According to a state law passed in 1971, admission to these schools must be based solely on the results of an exam.[2] The test is called the SHSAT (Specialized High Schools Admissions Test), and is constructed and scored by a private firm, American Guidance Service. It is given each fall to students seeking admission for the school year beginning the following September.

The pool of applicants is large, and the competition keen. In 2005 and 2006, between 25,000 and 27,000 eighth graders (including 4,500 to 5,000 private school students), took the SHSAT for admission to the ninth grade at the specialized public high schools. Only 18% to 20% of all test takers were offered a seat at one of these schools; fewer than half of those were admitted to their first choice school.[3]

Although this process has been going on for decades, there has never been a published study of the admissions procedure, of how it compares with generally accepted psychometric standards and practice, or of the test itself: how it is scaled, the statistical properties of the distribution of scores, measures of test reliability, confidence intervals around scores, and so on. This paper seeks to remedy these deficiencies, using the individual test results from the 2005 and 2006 SHSAT test administrations.

The national debate on the use of standardized tests for student evaluation and school admissions has a rich history in the education literature. Proponents stress that standardized tests provide a common yardstick for comparing students, reduce the influence of personal biases in admissions decisions, and have the effect of promoting meritocracy. Critics contend that these tests are imperfect and narrow measures of students' abilities or knowledge, have difficulty distinguishing among candidates with similar abilities or knowledge, and are biased along racial, gender, and class lines.[4]

There is some common ground among proponents and critics though. Most agree that a necessary condition for standardized tests to be considered valid guides for student evaluation and school admissions is that they be shown to improve predictions of how students will perform in

the future. To that end, batteries of predictive validity studies have been conducted over the years to see whether standardized tests really do help predict future student performance.[5] Although results vary and disagreements persist, a rough consensus has emerged that forms a core tenet of educational standards and practice. Standardized tests are generally viewed as imperfect but valid supplementary aids for evaluating students and making admissions decisions provided that (1) the tests are properly constructed for their intended use as predictors, (2) students are familiar with the content and format of the tests, (3) evaluators understand the limitations of the tests, (4) the tests are used in conjunction with other factors, and (5) their efficacy is supported by predictive validity studies.[6]

These widely accepted psychometric standards and practices provide a benchmark for this study of the admissions process at the New York City specialized high schools. This study's findings will remind us why these standards were established, and why most other selective educational institutions—including other prominent test-admission high schools like Thomas Jefferson in Virginia and Boston Latin—adhere to these standards by using multiple admissions criteria and by relying on predictive validity studies to inform those criteria. Even a well-designed test like the SHSAT is subject to a lack of precision and uncertainties. For example:

- The SHSAT exhibits an unusual scoring feature that is not widely known, and may give an edge to those who have access to expensive test-prep tutors. Someone with a very high score in one section of the test and a relatively poor one in the other will have a better chance of admission than someone with relatively strong performances in both. Reasonable alternative scoring systems would yield different results for many students, and there is no evidence offered to support the validity of the current system.
- Thousands of students who are not being accepted have scores that are statistically indistinguishable from thousands who are granted admission. And these estimates are derived using the less precise, classical-test-theory-based measures of statistical uncertainty, which may understate the problem. The NYCDOE fails to provide the more accurate, item-response-theory-based estimates of the SHSAT's standard error of measurement (SEM) near the admission cutoff scores, which would offer a clearer picture of how well the test is able to differentiate among students who score close to the admission/rejection line. This omission violates generally-accepted testing standards and practices.
- Different test versions are used. Details about how these versions are statistically equated and how accurate that equating is are not provided. The mean scaled scores vary across versions more than the chance distribution of ability levels across large random samples of students would suggest is plausible, suggesting that the scaling system may not

completely eliminate differences among test versions when estimating student ability. Thus, students who receive certain versions of the test may be more likely to gain admission than students who receive other versions.

- No studies have ever been done to see if SHSAT scores predict different things for different genders and ethnic groups.
- No predictive validity studies have ever been done linking SHSAT scores to any outcomes. In fact, the NYCDOE has never published what specific, measurable objectives the SHSAT is supposed to predict (high school performance, SAT scores, etc.). Without well-specified objectives and carefully constructed validity studies, there's no way to know if admissions criteria are serving their purpose, what that purpose is, or if an alternative admissions system would be more appropriate.

By failing to provide detailed information about many aspects of the SHSAT, by not making all the implications of the scoring system known to all test takers, and especially by relying on a single imperfect criterion whose predictive validity has never been established, the admissions practices at the New York City specialized high schools run counter to educational standards and practices advocated by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.[7] By pointing out some of the flaws of the New York policy, this case study illustrates why relying on multiple imperfect criteria guided by predictive validity studies is the preferred standard and practice.

The Process

There are eight specialized high schools in New York City that use the SHSAT. They are Stuyvesant, Bronx Science, Staten Island Tech, Queens College High School, Lehman College High School, Brooklyn Tech, City College High School, and Brooklyn Latin (a school that just began using the SHSAT for ninth graders entering in September 2007). These schools have excellent reputations, as suggested by recent results of a national ranking system that placed four of them—Stuyvesant, Bronx Science, Staten Island Tech, and Brooklyn Tech—among the top 40 public high schools in the country.[8] Students taking the SHSAT must list which of the specialized schools they would like to attend, in order of their preferences (their first-choice school, their second choice, etc.). They can rank as few as one or as many as eight, but they can only be offered a seat at one of these schools.

After the students make their choices, the testing company ranks the students, based solely on their scores on the SHSAT. The highest-scoring students are offered seats to their first-choice school, until all the seats at one school have been offered. That school will have the highest "cutoff" score (the score obtained by the students who are offered the

school's last seats). Students who just miss the cutoff for that school will be offered a seat at their second-choice school (or at their first choice, if it is a different school than the one with the highest cutoff), until all the seats at a second school have been offered. [9] This process continues until all seats at all the schools have been offered.

Since not all students offered seats will enroll, the number of seats offered exceeds the capacity of these schools. A school's capacity and its expected yield (how many of those offered seats are likely to enroll), determine how many seats the school can offer. How many test takers will qualify for those seats depends on how many want them—i.e., on how the students rank the schools. If many of the best scorers on the SHSAT select the same school as a top preference, that school will not have to go very far down the list before its limit of seat offerings will be reached. That school will have the highest cutoff score; equivalently, a smaller fraction of test takers will qualify for a seat at that school than at the other schools.

Though preferences vary somewhat from year to year, Stuyvesant and Bronx Science have historically had the highest and second-highest cutoffs, respectively, because more students who do well on the SHSAT tend to select Stuyvesant and then Bronx Science as their top choices. For example, in 2005 and 2006, only the top 4.5% to 5% of scorers qualified for a seat at Stuyvesant, the top 11% to 12% qualified for a seat at Bronx Science, and the top 18% to 20% qualified for a seat at the school with the lowest cutoff.

The Test

The SHSAT consists of 95 multiple-choice questions, divided into two sections: verbal and math. The math segment contains 50 questions on topics including elementary number theory, algebra, and geometry. The verbal is subdivided into three parts: 30 reading comprehension questions, 10 logical reasoning questions, and five scrambled paragraphs (each of which counts as the equivalent of two questions). So the maximum number of correct answers ("raw score") is 50 on math and 50 on verbal. Four main versions of the test—A/B, C/D, E/F, and G/H—are given, in part to reduce the potential for cheating. The versions are designed to be similar, and students are randomly assigned a test version.[10]

Summary Statistics for the Raw Scores

Verbal raw scores had a mean ranging from 25 to 29, while the mean on the math was 20 to 22 (Table 1). The verbal/math gap was a bit wider in 2005 than 2006, though in both years there were some statistically significant differences in mean raw scores across test versions.[11] Other aspects of the distributions varied somewhat as well. The distribution of verbal raw scores was flatter, without as clear of a peak as the math raw scores (especially in 2005). More students scored above the mode—the

most frequent score—on math than below it, while the verbal raw scores were somewhat more symmetric on all versions (Figures 1 and 2).

Table 1: Summary Statistics for Raw Scores

2006 Test	**Verbal**				**Math**			
	A/B	**C/D**	**E/F**	**G/H**	**A/B**	**C/D**	**E/F**	**G/H**
Mean	25.1	26.5	26.5	24.8	22.3	20.8	22.1	20.4
Stand. Dev.	10.5	11.4	11.0	10.7	10.7	10.3	11.9	10.7
Skewness	0.21	0.07	0.14	0.24	0.57	0.72	0.59	0.69
Kurtosis	-0.91	-1.01	-0.94	-0.88	-0.61	-0.28	-0.77	-0.45
Number of test takers	6704	8130	6029	3597	6704	8130	6029	3597
2005 Test	**Verbal**				**Math**			
	A/B	**C/D**	**E/F**	**G/H**	**A/B**	**C/D**	**E/F**	**G/H**
Mean	28.0	27.3	29.4	25.7	21.9	19.6	20.6	21.3
Stand. Dev.	10.8	11.0	11.3	10.6	10.7	10.1	10.4	11.1
Skewness	-0.09	0.05	-0.20	0.22	0.67	0.85	0.75	0.69
Kurtosis	-0.92	-0.95	-0.95	-0.83	-0.44	0.00	-0.25	-0.47
Number of test takers	8552	5803	5929	5780	8552	5803	5929	5780

Figure 1: Frequency Distribution, Verbal Raw Scores (2006)

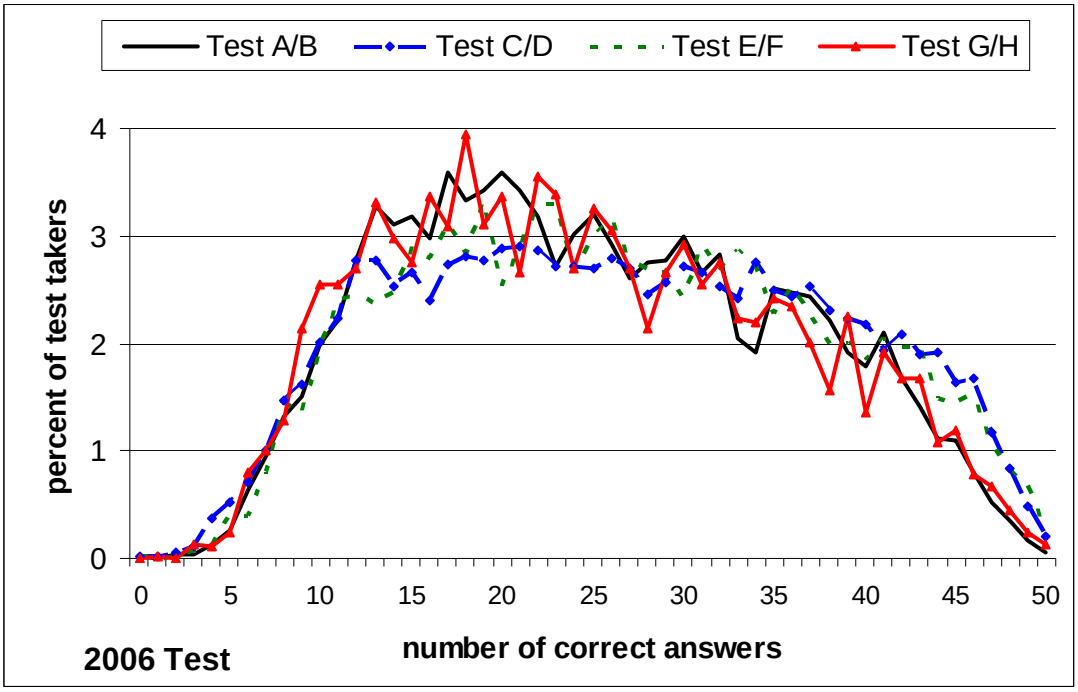
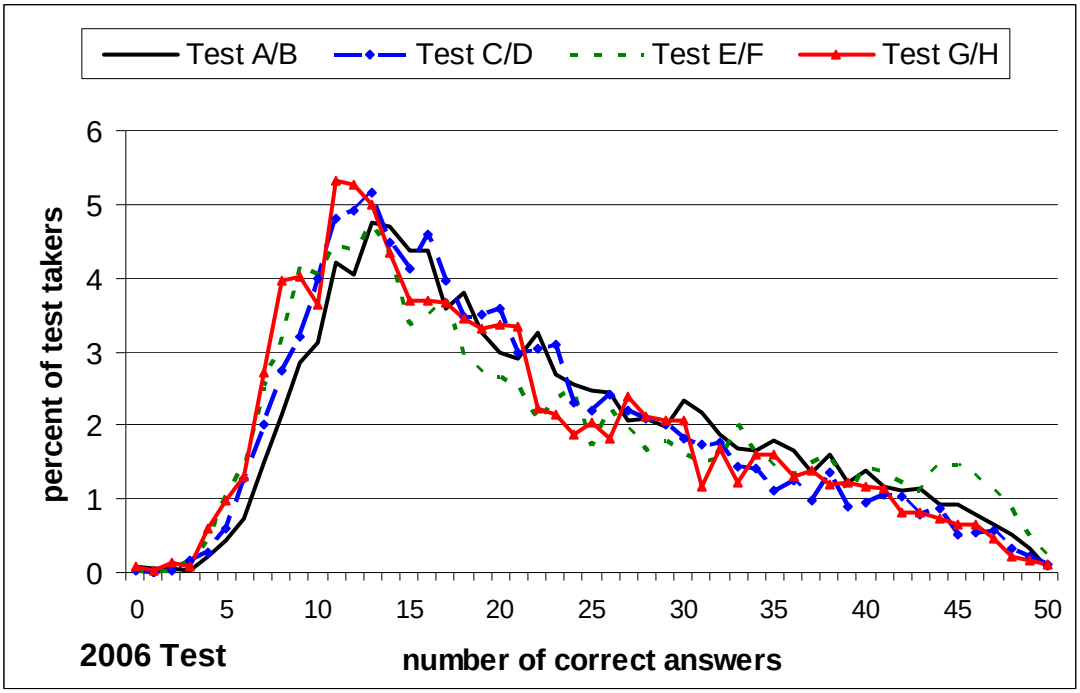


Figure 2: Frequency Distribution, Math Raw Scores (2006)



Grading the Test: Converting Raw Scores to Scaled Scores

On each test version, students are scaled relative to others who took that version. The scaling is done separately for the verbal and the math because the two sections aim to measure somewhat independent skills, and because the distributions of math and verbal raw scores have different statistical properties. According to the NYCDOE, the scaled scores in each section are derived from item-response-theory calibrations, and are estimates of a student's "ability" in the skill that each section aims to measure. In item response theory (IRT), a relationship is estimated between the level of ability (or knowledge, or whatever a given test is intended to measure) and the likelihood of a correct response to each question on the test. These estimates are summed across all questions to produce an estimate of the relationship between the total number of correct responses (raw score) and the ability level (the scaled score). IRT also generates estimates of the overall goodness-of-fit of the relationship between raw scores and scaled scores, as well estimates of the variance of each scaled score (i.e., how precisely each ability level is measured).[12]

Despite several requests, the NYCDOE did not make the details of the item-response-theory estimates available for this study.[13] For example, no information was provided on the overall goodness-of-fit of the relationships between raw scores and ability levels nor on how precise the estimates of individual ability levels are and how that precision varies across ability levels. Without those estimates, it's hard to know how good of a job the scaling system is doing in linking raw scores to the underlying

ability that is intended to be measured or how confident one can be that the test is able to distinguish among students who score close to the cutoffs. That's why failing to provide IRT-based estimates is contrary to Standard 3.9 of the *Standards for Educational and Psychological Testing*:

> When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.[14]

For this study, the NYCDOE provided only the raw scores and the corresponding scaled scores—but not the details of how the latter were derived. In 2005 and 2006, the scaled scores ranged from 20 to about 370 on both math and verbal, varying slightly by test version. The relationship between raw scores and scaled scores is nonlinear; so too is the relationship between the percentile rank based on the raw scores and scaled scores. That is, a change in raw score (and percentile) alters the scaled score by more near the upper and lower ends of the range of scores than near the middle. For example, on the verbal section of test G/H in 2006, an increase in raw score from 25 to 30 (an increase in percentile from 55.0 to 68.5) boosted the scaled score 19 points. In contrast, a rise in raw score from 40 to 45 (a rise in percentile from 90.2 to 97.7) added 28 points to the scaled score, and an increase in raw score from 45 to 50 (percentile increase of 97.7 to 100.0) caused the scaled score to leap 80 points (Figures 3 and 4, following). Similar relationships hold for the other test versions, for the math section, and for the 2005 test as well.

If the scaled score is taken to be the "true" barometer of the latent trait being measured—in this case, something akin to math or verbal "ability"—this scaling system implies that an increase in raw score (or percentile) reflects a bigger increase in ability if it occurs near one of the tails of the distribution, rather than near the middle. For example, on the math section of test G/H in 2006, the three-point difference in raw score between 45 and 48 (percentile 98.4 vs. 99.7) and the seven-point difference in raw score between 21 and 28 (percentile 62.0 vs. 76.5) correspond to the same difference in scaled score (25 points), and hence are interpreted by the SHSAT as reflecting a similar differential in underlying ability.

Of course, no scale that is designed to measure a latent trait such as verbal or math ability can ever be a perfect interval scale, with each increment representing exactly the same difference in the underlying trait throughout the scale. This is an issue confronting all such scales. In fact, scaled scores are generally considered to be only approximations to interval scales, though closer approximations than are raw scores or percentiles, whose increments are assumed to overstate changes in ability near the middle of the distribution, where many students are clustered, and

understate it near the tails, where there are few students. The scaling system used for the SHSAT is not unlike that used on many standardized tests.

Figure 3: Raw Scores & Scaled Scores

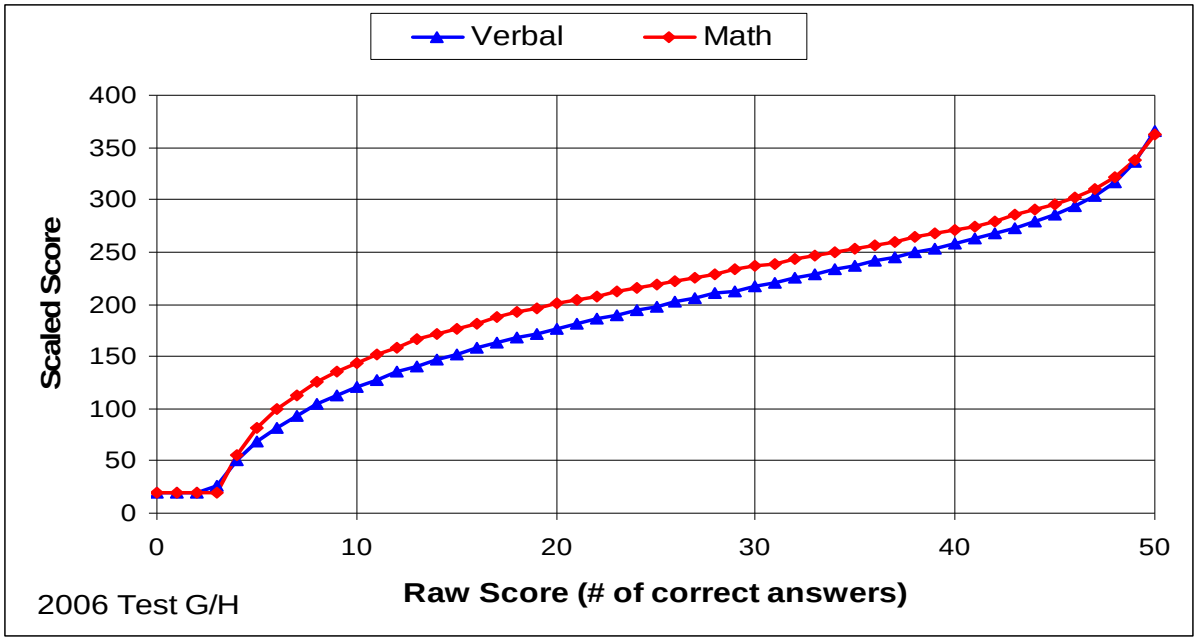
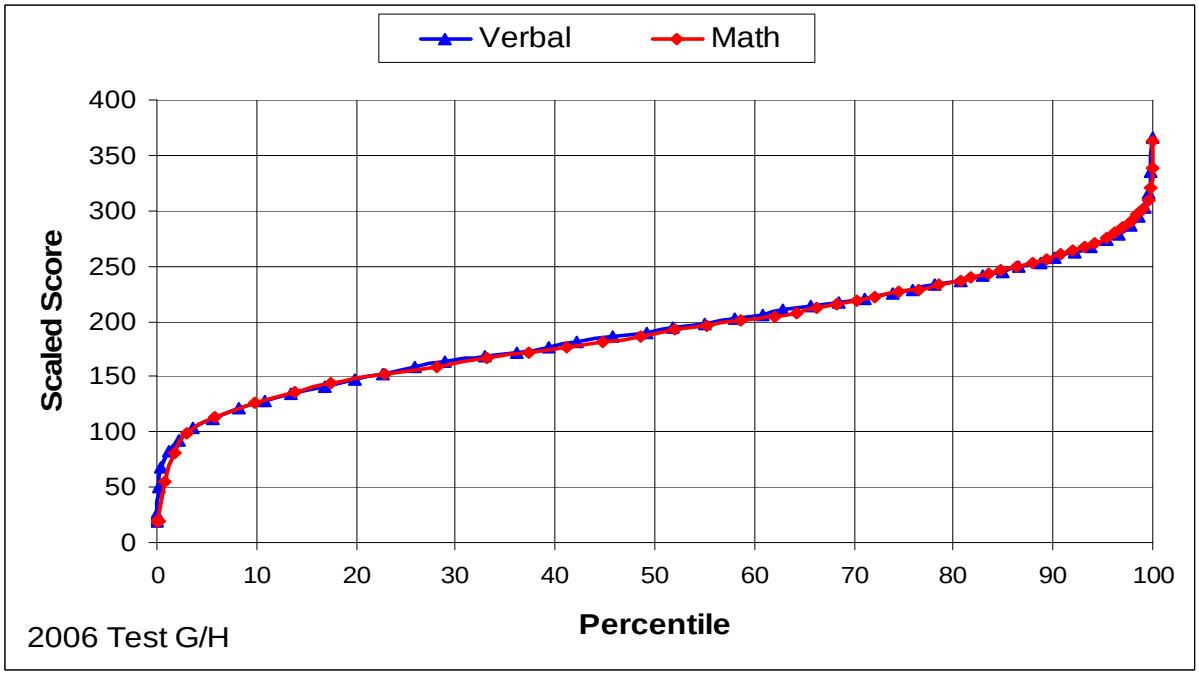


Figure 4: Percentiles & Scaled Scores



But where the SHSAT differs greatly from other standardized tests is that the verbal and math scaled scores on the SHSAT are added together to form a single, composite scaled score for each test taker, and this is the only number used to rank the students and to create cutoffs for the

specialized high schools. This approach of adding the math and verbal scaled scores together and putting equal weight on each, creating a single scaled score, is apparently intended to measure some combined trait. That collective trait is then used as the sole determinant of admission. Doing so implicitly assumes a perfectly compensating, one-for-one relationship between the two scores; it assumes that a one-point increase in math scaled score exactly compensates for a one-point decline in verbal scaled score, leaving the combined ability construct unchanged. Thus, a scaled score of 300 verbal, 200 math is treated as an equivalent measure of some putative combined ability as a 250 verbal, 250 math because the 50-point rise in verbal scaled score is assumed to compensate exactly for the 50-point decline in math scaled score.

But the math and verbal sections are designed to measure different skills, are scaled separately, and each scale is only an approximation to an interval scale. The NYCDOE has no reason to assume the existence of a perfectly compensating, one-for-one relationship between the two different scales. Yet that is exactly what the SHSAT scoring system implies. In fact, given that the NYCDOE provides no support for this assumption, their system runs counter to the *Standards for Educational and Psychological Testing*, which states:

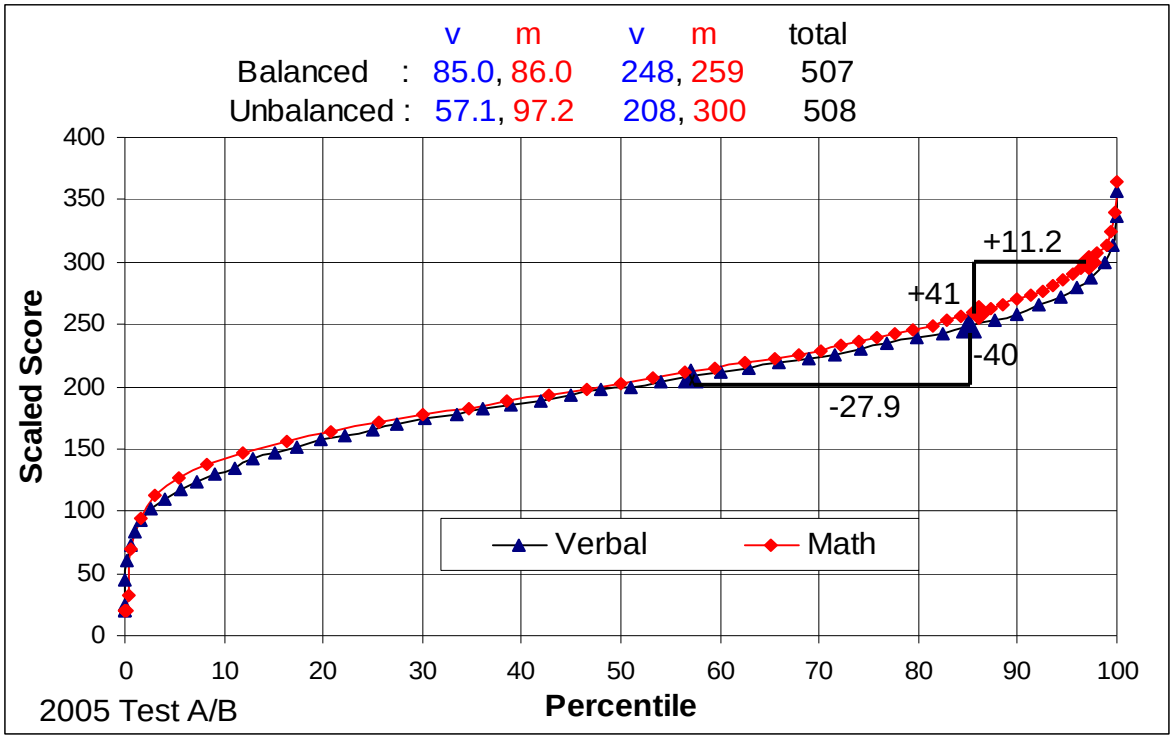
> Where composite scores are developed, the basis and rationale for weighting subscores should be given.[15]

Other standardized tests, such as the SAT and ACT for college admissions, and the Independent Schools Entrance Exam (ISEE) and the Secondary Schools Admissions Test (SSAT) used for admission to private secondary schools, report separate scaled scores for each section, which avoids the assumption that NYCDOE makes. This allows schools to look at the scores individually and to place different weights on them if they choose, using the results of validity studies to inform their decisions.[16] The NYCDOE has conducted no validity studies to support their approach of ranking students for admission solely on the equally weighted sum of their math and verbal scaled scores on the SHSAT.

The unorthodox system used to derive these combined SHSAT scaled scores results in an advantage for a subset of students. Those students who score toward the upper end of the distribution of raw scores in one section and much lower in the other will be deemed to have more combined ability (will get higher total scaled scores) than those who score moderately high in both sections. For example, as shown in Figure 5 (following), a student scoring in the 97.2 percentile in math (scaled score of 300) on test A/B in 2005 needed to score only in the 57.1 percentile in verbal (scaled score of 208) to be considered by this scoring system to have greater combined ability (higher total scaled score) than a student scoring in the 85.0 percentile in verbal (scaled score of 248), and 86.0 in

math (scaled score of 259). Similar results hold for other test versions in both years and for those whose stronger area was the verbal.

Figure 5: Balanced vs. Unbalanced Scores

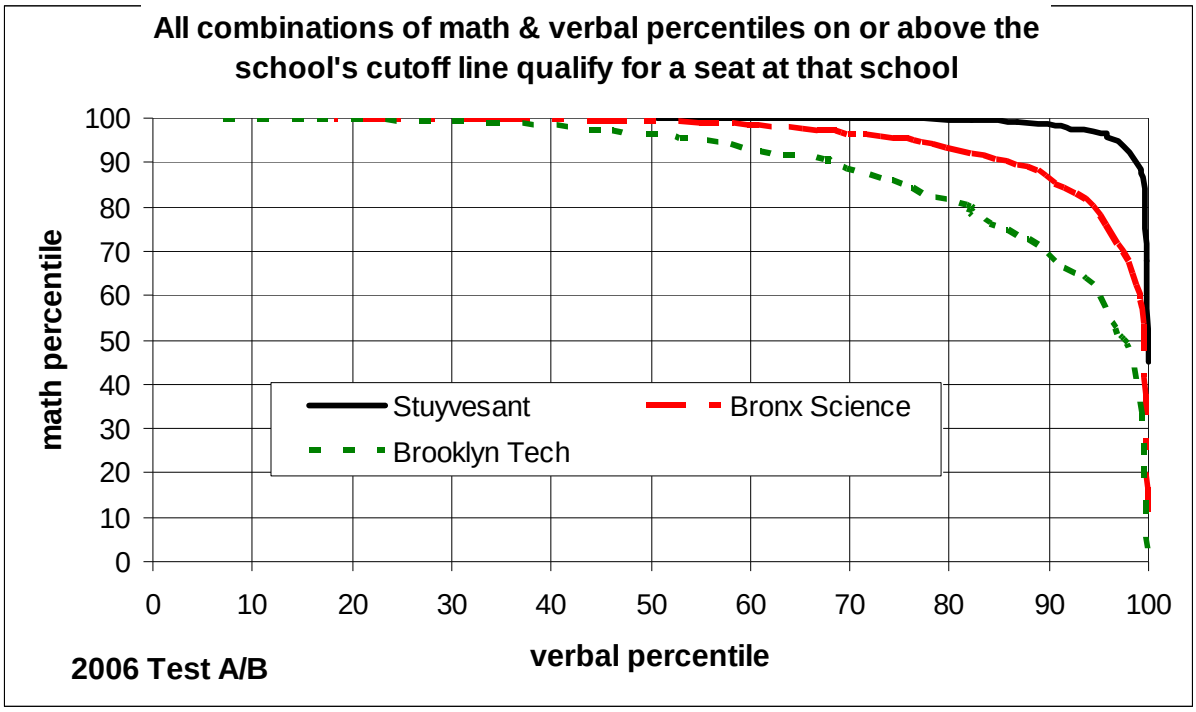


This can matter a lot for test takers near the cutoff for admission to a specialized school. A student with a perfect score in one dimension of, say, test A/B of the 2006 SHSAT, needed to score only in the 45th to 50th percentile in the other dimension to push his or her total scaled score high enough to meet the cutoff for Stuyvesant (Figure 6, following). A near-perfect score in one dimension (99th percentile), needed to be accompanied by only a mid- to high-80s percentile score in the other to meet the Stuyvesant cutoff, while someone scoring in the mid-90s in both would have been considered to have less combined ability and would have fallen just short.

The effects of this scoring system become greater at schools with lower cutoffs. To meet the cutoff for Bronx Science in 2006, a student with a perfect score in one section had to score only in the 11th to 18th percentile in the other—about 10 to 14 correct answers out of 50, or not much better than pure guessing. Some might argue there is merit in giving students who obtain perfect scores in one area an edge in admissions because they may be the future "geniuses" who will go on to great things. No predictive validity studies have ever substantiated this claim for the SHSAT, but even if we accept the argument, the benefits of the nonlinear scoring system are not confined solely to students with perfect or near-perfect scores in one dimension. Scoring in the 96th to 97th percentile in one area—very strong, but not nearly the perfect "genius"—was enough to compensate for about a 70th

in the other to meet the Bronx Science cutoff, while a student scoring in the 87th to 89th in both would have been rejected.

Figure 6: Cutoff Lines for Admission



To meet the Brooklyn Tech cutoff (second-lowest cutoff in 2006), a perfect score in one section required only the 3rd to 5th percentile in the other, or 7 to 9 correct answers (a below-average guesser), while scoring in the 90th to 94th percentile in one dimension—far from a perfect score—required just the 50s or 60s in the other section. By contrast, a student scoring at the 80th percentile in both would have just missed. Similar results hold for the other schools, the other test versions, and for the 2005 test.

Is it "better" to admit students with scores in the upper end of the distribution in one area and low in the other rather than students with moderately high scores in both? Do the former really have more "ability"? The NYCDOE implicitly assumes that it is, and that they do. But ultimately this is an empirical question. To answer it, the NYCDOE first needs to define what "better" means by setting clear objectives for the admissions process. Specifically, what performance criterion variables are SHSAT scores supposed to predict? Then, validity studies need to be carried out to see how accurately the test scores predict the chosen criterion performance. This is recommended procedure according to the *Standards for Educational and Psychological Testing*.[17]

For the SHSAT, a key aim of predictive validity studies should be to examine what type of scoring system maximizes the probability of meeting the chosen criterion performance. Is it the current system, or one that puts different weights on the math and verbal scaled scores, or sets a minimum cutoff for each section? Absent clearly specified objectives and

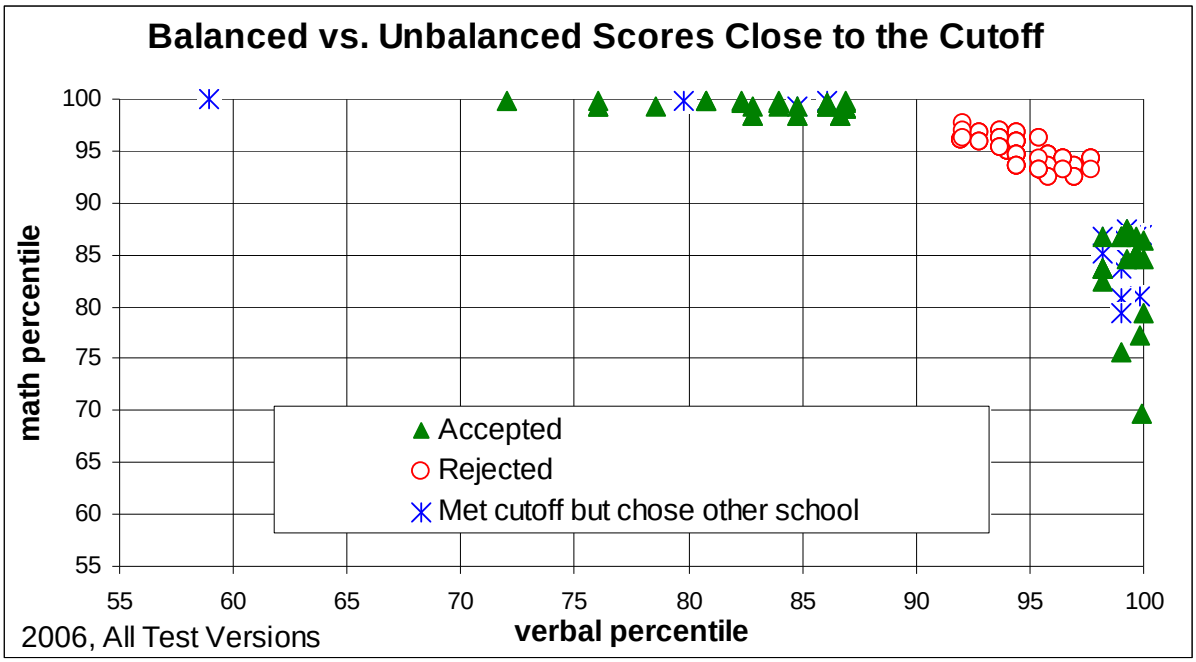
carefully constructed validity studies, the NYCDOE has little basis on which to determine which scoring system is optimal.

This is not just a hypothetical issue. There were many instances in 2005 and 2006 where the scoring system made the difference between admission and rejection. For example, of all the students who met the cutoff for Stuyvesant in 2006, 72 (nearly 6%) did so with "unbalanced scores" (Table 2, Figure 7).[18]

Table 2: Balanced vs. Unbalanced Scorers Near the Cutoffs

	Number of Students					
School	**Stuyvesant**		**Bronx Science**		**B'klyn Tech**	
Year	2006	2005	2006	2005	2006	2005
Met cut w/ unbalanced scores	72	74	107	111	102	82
Offered seats	53	49	54	53	97	65
Just missed w/ balanced scores, wanted to go	90	80	63	93	89	171

Figure 7: Stuyvesant



One made the cut with a 59th percentile verbal because of a perfect math score. That student happened to choose another school, but would have been offered a seat at Stuyvesant had he or she wanted one. All told, 53 of these unbalanced scorers did choose Stuyvesant first, and hence were offered seats, taking up more than 5% of the seats at this school. Of those,

nine scored in the 70th to 79th percentile in their lower dimension, but made the cut because of 99th percentile scores in the other section. By contrast, there were 90 students in 2006 who just missed the cut and wanted to go to Stuyvesant, and who had balanced percentile scores of mid-90s in both

At Bronx Science, 107 students met the cut in 2006 with unbalanced scores (Table 2, Figure 8). Of those, 54 were offered seats (about 5% of all those offered seats), and more than half had scores below the 70th percentile in one section, including one with a 49th percentile verbal. Nearly all the unbalanced who met the cut with scores in the low- to mid-70s in one section had only low- to mid-90s in their stronger area, belying the notion that the scoring system benefits only those at the very top in one dimension. On the other side of the ledger, there were 63 students in 2006 (93 in 2005) who just missed the cut and had balanced percentile scores of mid- to upper-80s in both sections. At Brooklyn Tech, 102 students met the cut in 2006 with unbalanced scores (Table 2, Figure 9, following). Of those, 97 were offered seats (more than 5% of all seats), and more than half scored below the 60th percentile in one dimension, including one with a 25th percentile verbal. And the unbalanced that got in with scores of low- to mid-60s in one section had only high-80s to low-90s in their stronger area, while 89 students were rejected with scores near the 80th percentile in both.

Figure 8: Bronx Science

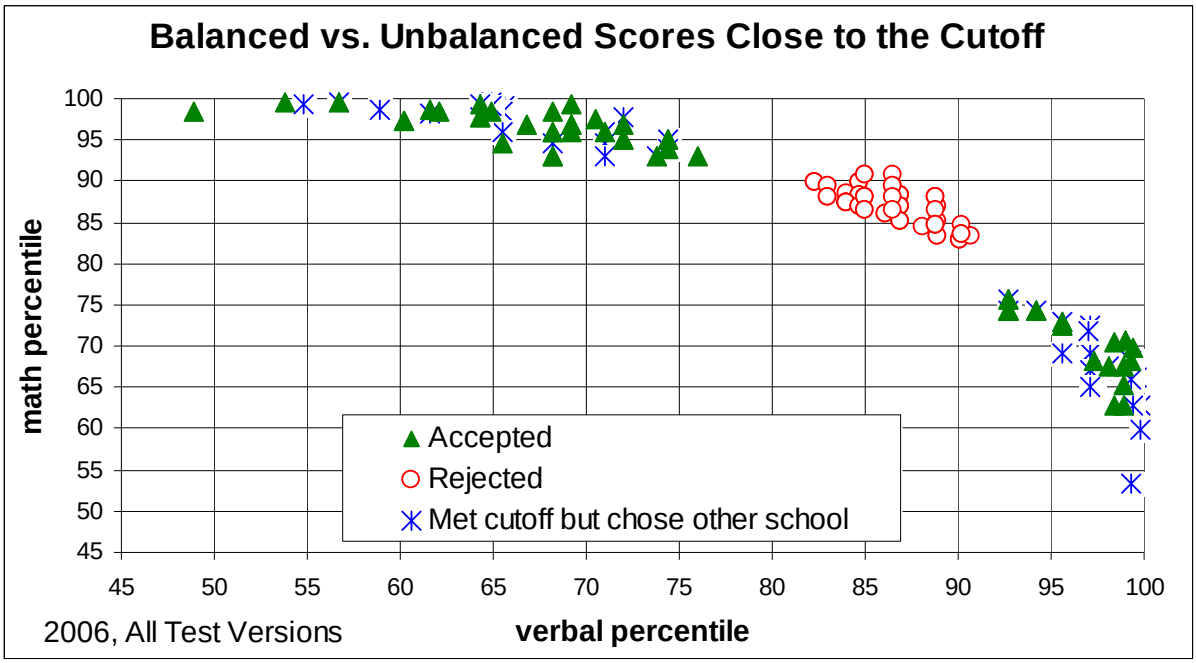
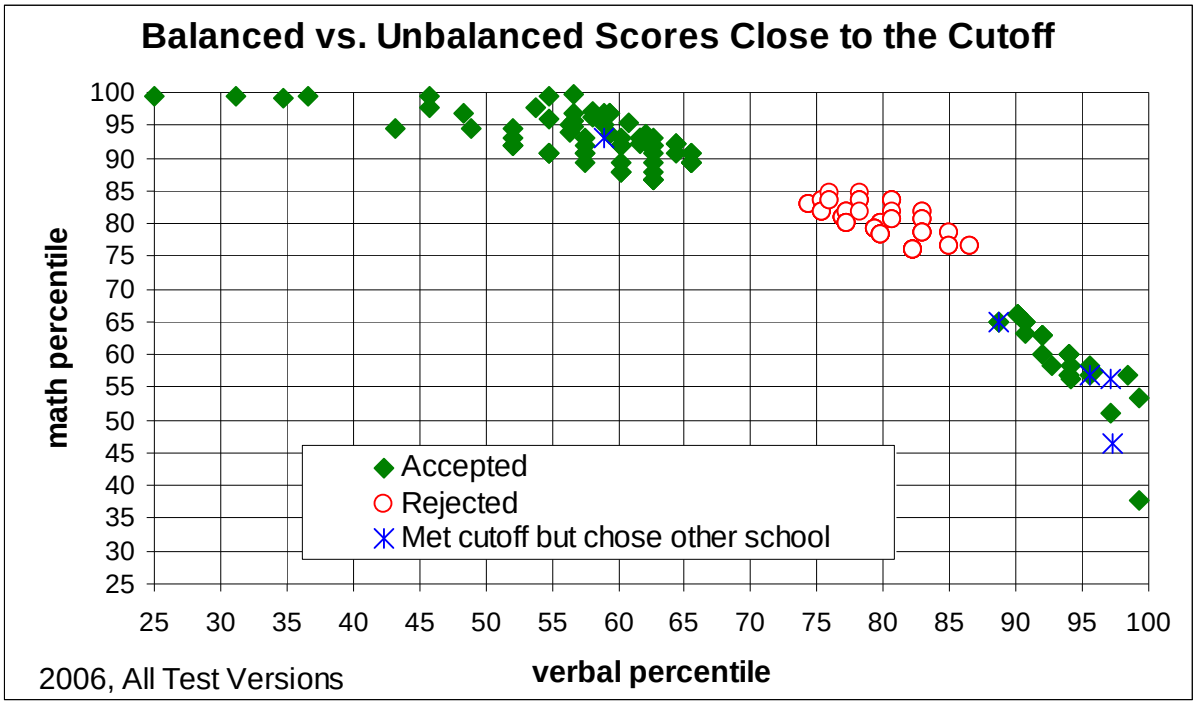


Figure 9: Brooklyn Tech



Reasonable alternative scoring systems—such as ones that used a more linear relationship between raw scores and scaled scores, or put different weights on math and verbal scaled scores, or reported separate math and verbal scores and allowed each school to decide, or set a minimum cutoff for each section—would yield different results for many students. On the SAT, and on the ISEE and SSAT used for admissions to private secondary schools, schools see the separate math and verbal scores (together with grades and other information), and can weight them differently, based on the results of predictive validity studies. Admission to the specialized high schools is based only on the composite score on the SHSAT, and the NYCDOE has never conducted a predictive validity study to see whether the current system is superior to some alternative.

The NYCDOE also fails to make students fully aware of the implications of the current scoring system. The Specialized High Schools Student Handbook published by the NYCDOE does mention that the relationship between raw scores and scaled scores is not proportional, and that the sum of the scaled scores in math and verbal is the sole determinant of admission.[19] But it does not point out that since a student will get a higher total score by concentrating more of his or her correct answers in one section, he or she might consider spending more time (both in preparation and on the test) on his or her stronger area. Many people might find this advice counterintuitive. Yet that is exactly what some expensive test-prep tutors advise their pupils—those fortunate enough to be able to afford test-prep services.[20] They also emphasize to their pupils that catching an error on the test in their stronger area is worth more to their

total score than catching an error in their weaker area. That should be pointed out by the NYCDOE, too. Whatever scoring system is used, everyone should know all of its implications. Indeed, if some test-taking strategies are shown to affect test performance, the *Standards for Educational and Psychological Testing* states that, "…these strategies and their implications should be explained to all test takers before the test is administered."[21]

Uncertainty of Test Scores

No test, no matter how it is scored or how well it is designed, is a perfect measure of a student's ability in whatever dimension being tested. All scores are merely estimates of the underlying trait that is being measured. There is uncertainty or imprecision around those estimates. In classical test theory, the degree of uncertainty, reflected in the standard error of measurement (SEM) of the test, is derived from two elements: the standard deviation of the scores across all the students who took the test, and a measure of the "reliability" of the test—how much the variation of scores across different students reflects true differences in their skills in whatever the test is supposed to measure and not just random errors (due, for example, to the questions being imperfect samples of the skill being tested).

Reliability metrics generally fall into two broad categories: test-retest reliability measures, and internal-consistency measures. The former are obtained by administering alternate forms of the same test to the same examinees twice. The correlation between the scores on the two tests provides a good measure of the reliability of the test. In practice, however, this method is difficult to implement because it is hard to give the same test twice. Instead, most tests that are given just once (like the SHSAT) use internal-consistency measures to gauge reliability. These focus on the correlation between the answers to individual questions on the test. The higher that correlation—i.e., the more likely it is that a student who gets one question correct will get other questions correct as well—the more reliable the test is as a measure of a particular uniform skill.

The advantage of internal-consistency measures is that they don't require multiple administrations of the test. The disadvantages are several: since they estimate how reliable a test is at measuring a single skill, they are best used for tests that are designed to measure a single skill. For the SHSAT, that means generating separate math and verbal reliability measures is more appropriate using this method than generating a combined reliability measure. Second, internal-consistency estimates of test reliability tend to be higher than test-retest reliability estimates. Most research has found that the correlation of scores across repeat tests is lower than the correlation of items on the same test.[22] So the measure of test reliability calculated for the SHSAT—.91 for the verbal, .92 for the math, and .95 for the total in both 2005 and 2006, on a scale of 0 to 1—

should probably be thought of as upper-bound estimates of the test's actual reliability.[23] And the estimates of the SEM derived from these reliability measures—15.0, 14.0, and 20.4 points, for the verbal scaled score, math scaled score, and total scaled score, respectively, in both 2005 and 2006—should be considered lower bounds.[24] The final problem with an SEM calculated under classical test theory is that it is a "one-size-fits-all" measure; all ability levels are assumed to be measured with the same degree of precision, reflected in the single SEM. In fact, average ability levels might be measured with more precision than very high or very low ability levels.

With item response theory (IRT), estimates of a test's precision can vary across ability levels. For each estimated ability level, IRT generates a corresponding variance of that estimate.[25] A key advantage of such finely calibrated estimates of uncertainty is that they would better enable us to hone in on the precision with which the SHSAT measures ability levels near the cutoff scores for admission to the specialized high schools. IRT estimates offer a potentially clearer window than classical test theory into how confident we can be that the SHSAT is able to differentiate between students whose ability levels are close to the cutoffs. That's why Standard 2.14 of the *Standards for Educational and Psychological Testing* states:

> Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.[26]

This is especially important for a test whose cutoff scores are the sole arbiter of the admissions decision. But the NYCDOE has not released any data on the IRT estimates for the SHSAT's standard error of measurement near the cutoff scores. Despite several requests, the department provided nothing about the test's information function—the relationship between estimated ability and the precision of those estimates.[27] That's a crucial omission, because such estimates would offer a more precise measure of the uncertainty of scaled scores near the admission cutoffs.

Instead, we have to rely on the SEM estimates derived from classical test theory—the only ones the NYCDOE provided. These suggest that thousands of students may have fallen well within the range of statistical uncertainty of the cutoff scores for the specialized schools in both 2005 and 2006. For example, about 2,200 students in 2005 and 2,400 in 2006—about half of all those offered seats at the specialized schools each year—exceeded the cutoff for a school they wanted to attend by 20 points or less (Tables 3 and 4). At no school was the figure less than 35%; at several it was more than 60%. And 2,600 to 2,700 students each year fell short of the cut for a school they wanted to attend by 20 points or less.

That means that there were 4,800 to 5,100 students in 2005 and in 2006 (18% to 20% of all test takers), who fell well within the bounds of statistical uncertainty (as defined by classical test theory).We simply can't be that confident statistically that the "correct" decision was made for many of these students. This may understate the problem, because these figures are derived using what is probably a lower-bound estimate of the SEM under classical test theory. The IRT-based measures of uncertainty near the admissions cutoffs might be even larger because it's often harder to measure high ability levels as precisely as average ability levels. But we can't know that for sure in this case because the NYCDOE does not make the IRT estimates available, at odds with the *Standards for Educational and Psychological Testing*.

Table 3: Students Scoring Within One SEM of the Cutoff

School	**Exceeded Cut by <= 1 SEM (& wanted to attend)**				**Missed by <= 1 SEM (& wanted to attend)**	
	2006		**2005**		**2006**	**2005**
	Number	**% of those offered seats**	**Number**	**% of those offered seats**	**Number**	**Number**
Stuyvesant	408	42%	361	36%	468	499
Bronx Sci	467	43%	459	46%	442	505
Staten Isl	95	37%	125	35%	145	114
Queens	101	66%	87	52%	111	112
Lehman	115	63%	132	69%	140	144
City College	134	64%	873	49%	169	210
Bklyn Tech	934	53%	162	70%	1085	1028
Bklyn Latin	142	82%			142	
Total	2396	50%	2199	47%	2702	2612

Table 4: Breakdown of Test Takers

Category	**2006**		**2005**	
	Number	**% of Test Takers**	**Number**	**% of Test Takers**
Exceeded Cut by > 1 SEM	2423	9.7	2530	9.5
Exceeded Cut by <=1 SEM	2396	9.6	2199	8.2
Missed Cut by <= 1 SEM	2702	10.8	2612	9.8
Missed Cut by >1 SEM	17564	70.0	19371	72.5

All tests—no matter how well designed and reliable—are subject to statistical uncertainty. Of course, so are grades and other metrics. This is one reason why psychometric standards caution against using any single measure as the sole criterion for admission, and why colleges and universities, as well as high schools in other states that use the SHSAT or similar tests, look at other indicators too. A common approach would be to use test results to winnow the applicant pool, and then look more closely

at those within one or two SEMs of some threshold (with the estimates of uncertainty around the threshold derived from item response theory). These institutions would typically also use the results of predictive validity studies to inform their selection criteria. New York does none of these.

Different Test Versions

Students are randomly assigned one of four major SHSAT versions to take. The test versions are designed to be similar, but they are not identical in terms of difficulty, content, and so on. In principle, any differences between the test versions are expected to be corrected for by the scoring system. Specifically, the item-response-theory calibration that converts raw scores to scaled scores and statistically equates different test forms to a common scoring scale is structured so that a student at an estimated ability level is expected to receive the same scaled score regardless of which test version he or she is assigned.[28] But there is sampling variation around that expectation (the estimates of ability levels are approximate, the questions may not all reflect the skill the test seeks to measure, there can be errors in equating across forms, etc.). As a result, the expectation that a student's estimated ability will be invariant to the test version used to measure that ability may not always be realized in practice.

The current SHSAT scoring system implicitly assumes that any differences in average scaled scores across test versions reflect differences in average student ability across test versions—not differences in the versions themselves, which the equating process is expected to have rendered immaterial in estimating student ability. But no empirical evidence is provided to support this assumption. The NYCDOE offered no information on the degree of uncertainty that surrounds the equating of different test forms. This is not compliant with Standard 4.11 of the *Standards for Educational and Psychological Testing,* which states:

> When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of the equating functions.[29]

Without that information, it's hard to know how confident one can be in the SHSAT's assumption that a given scaled score corresponds to the same ability level on all test versions. If the average scaled score on, say, version A/B is lower than on C/D, the current scoring system does not assume that A/B was a tougher test, because even if it was, this was expected to have been corrected for by the equating across test forms. Instead, it assumes that students who took A/B had lower average ability.

How likely is it that groups of 4,000 to 8,000 randomly chosen students have statistically significant differences in average ability levels? Not very. There is no reason to expect the average ability of one large random sample of students to differ from that of another, much as there's no reason to expect average height to differ across large random samples of people taken from the general population. So if the equating system is really eliminating differences between test versions, we shouldn't expect to find many statistically significant differences in average scaled scores across test versions.

But we do. Table 5 shows the mean scaled scores for verbal, math, and composite for the four major versions of the 2006 and 2005 SHSAT. In 30 of the 36 comparison pairs (all but A/B vs. E/F math, C/D vs. E/F verbal and total in 2006, and A/B vs. G/H verbal, E/F vs. G/H math, and A/B vs. E/F total in 2005), we can reject the null hypothesis of no difference in mean scaled score across test versions at the 95% confidence level.[30] That is, there's less than a 5% probability that these differences in mean scaled scores owe to chance. If the samples were smaller—say, only 100 students assigned each test version—we couldn't reject the hypothesis that these differences were due to chance. But with such large samples of students, we can. So if the equating system really does render differences in test versions irrelevant for estimating student ability, why are there so many statistically significant differences in average scaled scores across versions—much more than the chance distribution of ability levels across such large random samples of students would suggest is plausible?

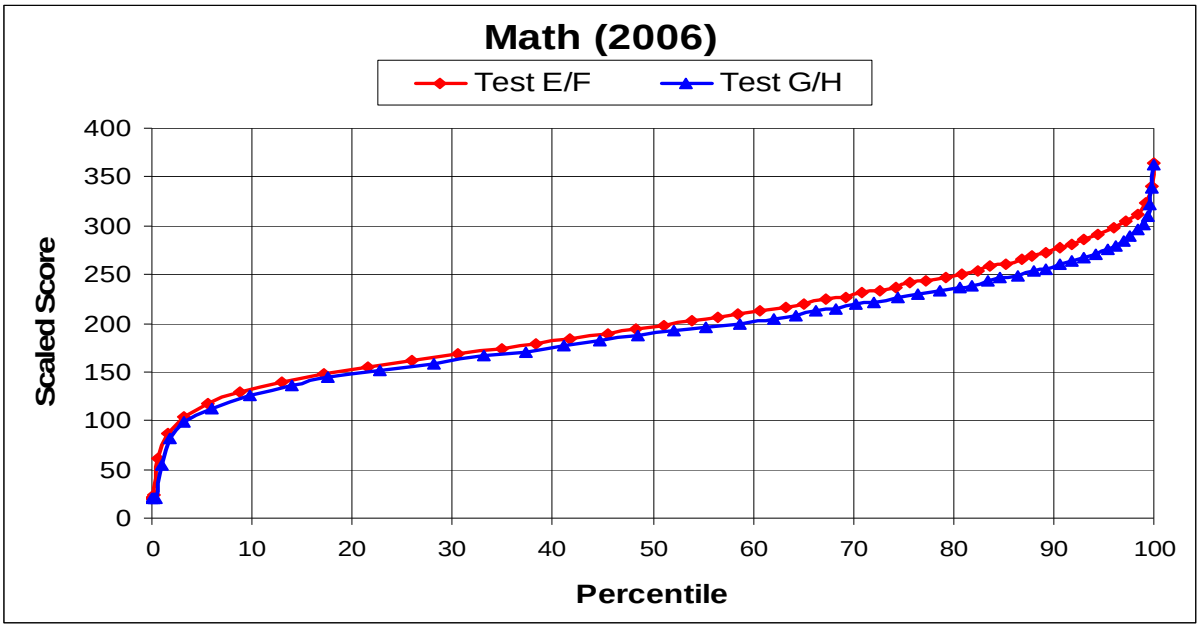
Table 5: Scaled Scores

2006 Test								
	Mean				**Standard Deviation**			
	A/B	**C/D**	**E/F**	**G/H**	**A/B**	**C/D**	**E/F**	**G/H**
Verbal	197.6	203.3	202.3	193.9	46.9	51.5	50.1	50.0
Math	202.3	200.5	202.3	193.0	47.4	45.6	54.2	50.3
Total	399.9	403.8	404.6	386.9	85.9	89.1	96.4	92.1
2005 Test								
	Mean				**Standard Deviation**			
	A/B	**C/D**	**E/F**	**G/H**	**A/B**	**C/D**	**E/F**	**G/H**
Verbal	200.0	195.3	203.9	201.0	48.7	51.4	52.7	46.7
Math	206.5	191.6	200.8	199.1	49.5	48.2	47.0	51.4
Total	406.5	386.8	404.8	400.2	90.3	91.5	91.4	90.1

The SHSAT implicitly drew the conclusion that students who took certain versions (G/H math and verbal in 2006, for example), had lower ability, on average. As a result, a given percentile rank translated into a lower scaled score if it was achieved on, say, test G/H rather than test E/F in 2006, because the former was presumed to have come against "weaker competition." For example, a 90.7 percentile on G/H math mapped into a scaled score of 260; on E/F math it translated into a scaled score of 277

(Figure 10). To achieve a comparable scaled score on G/H required scoring in the 95.4 percentile.

Figure 10: Differences between Test Versions



Although the magnitudes of the differences in mean scaled scores across test versions were not terribly large—even the biggest differences, when scaled by their standard deviations yield estimated effect sizes (Cohen's "d") of only around 0.2, which are considered small (Table 6, following)[31]—they affected the admissions decisions for many students who scored near the cutoffs for the specialized high schools. Indeed, acceptance rates varied considerably across test versions. For example, a smaller percentage of those who took version G/H in 2006 met the cutoffs (3.7% for Stuyvesant, 10.3% for Bronx Science, and 17.4% for any specialized school), vs. 7.3%, 15.2%, and 23.2%, respectively, for version E/F (Table 7, following). This would be appropriate if any differences between test versions were, in fact, being fully corrected for by the equating system, so that any differences in average scaled scores across test versions reflected differences in average ability across the versions (i.e., if those taking G/H in 2006 really were weaker, on average). But if the equating system did not fully adjust for differences between test versions, and G/H in 2006 was really a bit harder, then the students who were assigned that version were put at a disadvantage.

The issue boils down to this: differences in average scaled scores across test versions could be due to one, or both, of two sources—differences in the difficulty of the test versions that are not fully adjusted for by the equating system, or differences in the average ability of groups of 4,000 to 8,000 randomly selected students. The current system makes the strong assumption that it's all due to the latter. If even some is due to

the former—which seems likely given how many statistically significant differences there were in mean scaled scores across test versions—this would inject another element of arbitrariness into the admissions process, and provide another reason why the decision should not be based solely on one test score. Again, this is not a problem specific to the SHSAT; it potentially affects any test that uses multiple versions and is another reason to exercise caution in interpreting test results and not to use a single test as the sole criterion for admission.

Table 6: Effect of Different Test Versions on Total Scaled Score

2006 Test			
Test Versions	**Absolute Value of Difference in Means**	**Effect Size: Cohen's "d"**	**95% confidence band around "d"**
A/B vs. C/D	3.9	0.04	0.01 to 0.08
A/B vs. E/F	4.7	0.05	0.02 to 0.09
A/B vs. G/H	13.0	0.15	0.10 to 0.19
C/D vs. E/F	0.8	0.01	-0.03 to 0.04
C/D vs. G/H	16.9	0.19	0.15 to 0.23
E/F vs. G/H	17.7	0.19	0.15 to 0.23
2005 Test			
Test Versions	**Absolute Value of Difference in Means**	**Effect Size: Cohen's "d"**	**95% confidence band around "d"**
A/B vs. C/D	19.7	0.22	0.18 to 0.25
A/B vs. E/F	1.7	0.02	-0.01 to 0.05
A/B vs. G/H	6.3	0.07	0.04 to 0.10
C/D vs. E/F	17.9	0.20	0.16 to 0.23
C/D vs. G/H	13.4	0.15	0.11 to 0.18
E/F vs. G/H	4.6	0.05	0.01 to 0.09

Table 7: Percent of Students Meeting the Cutoffs, by Version

2006 Test			
Version	**Stuyvesant**	**Bronx Science**	**Lowest Cutoff School**
A/B	4.1	11.4	19.3
C/D	4.9	13.0	21.0
E/F	7.3	15.2	23.2
G/H	3.7	10.3	17.4
2005 Test			
Version	**Stuyvesant**	**Bronx Science**	**Lowest Cutoff School**
A/B	4.9	12.3	20.0
C/D	4.0	9.6	15.2
E/F	5.1	12.0	19.5
G/H	5.3	11.3	17.9

Gender & Ethnicity

Scores on standardized tests tend to vary systematically across gender and ethnic groups. For example, on average men score higher than women on the SAT, while African Americans and Hispanics tend to score lower than Whites and Asian Americans, for reasons that are vigorously debated.[32] On the SHSAT, females made up an average of 50.5% of all test takers in 2005 and 2006, but only 41.6% of those who met the cut for Stuyvesant, 44.0% of those who met the cut for Bronx Science, and 45.4% of those who met the cut for any specialized school (Figure 11). A complete breakdown of SHSAT scores by gender was not available for this study, so we don't know if males scored higher on average (their greater representation in the upper end of the distribution of scores could have been offset by greater representation in the lower end too, as is the case on the SAT), whether any difference in mean scores was statistically significant, and whether it was evident in the math section, the verbal section, or both.

SHSAT scores were not broken out by ethnicity either, but school enrollment data show that while African Americans and Hispanics together made up 72% of the NYC public school system in the 2005-2006 school year, they were only 5.5% of the student body at Stuyvesant, 11.2% at Bronx Science, and 16.5% at all specialized high schools—strongly suggesting that these students either did not take the SHSAT in great numbers, did not do well on the test, or both (Figure 12).

Figure 11: SHSAT Results by Gender

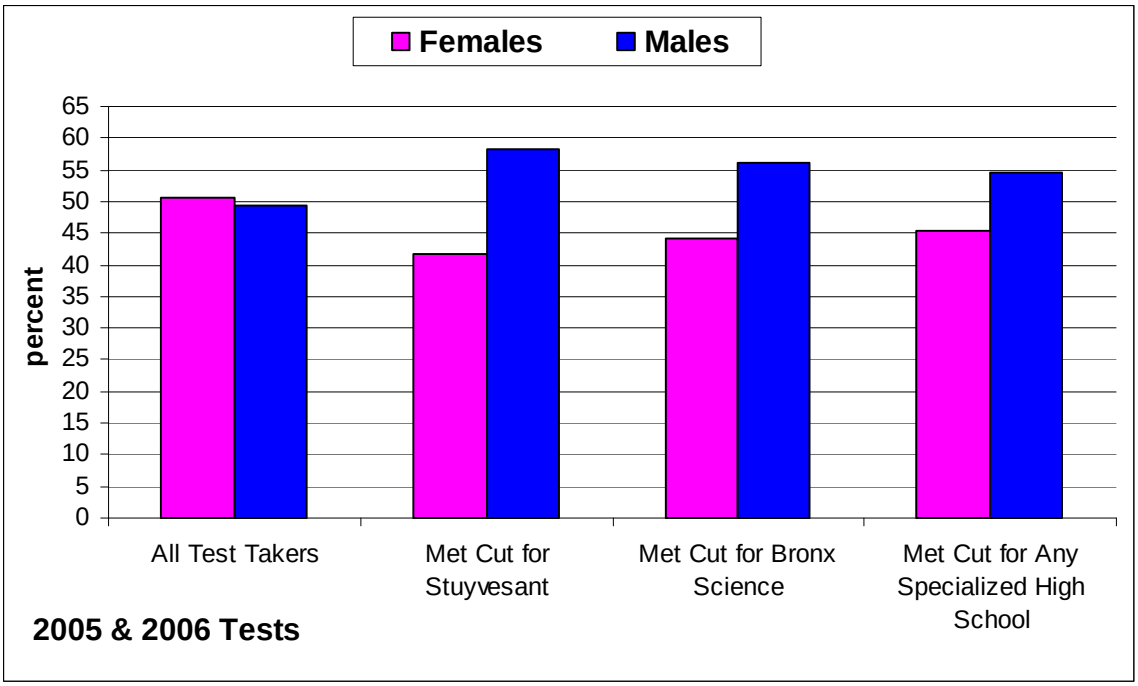
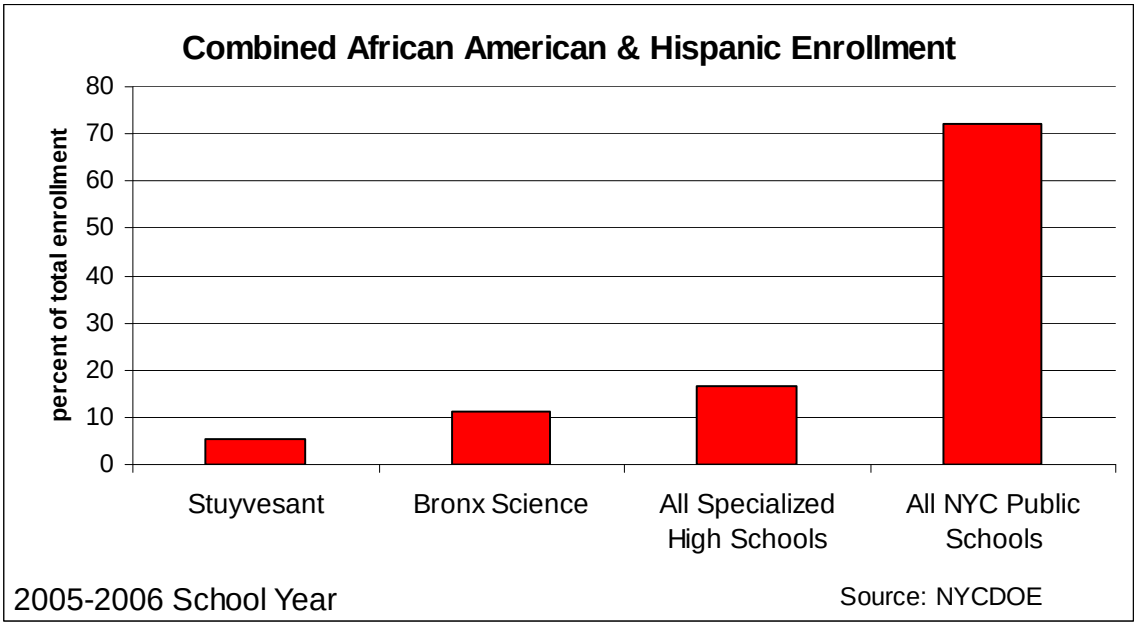


Figure 12: School Enrollment by Ethnicity



Just because test scores differ across gender and ethnic groups doesn't necessarily mean that a test is biased or of limited utility. From a psychometric perspective, the key is whether the predictive validity of a test varies across groups. If a given test score predicts different things for different groups, the test is said to exhibit "prediction bias," which diminishes its usefulness.[33] Studies have found several persistent cases of prediction bias in standardized tests like the SAT. For example, the correlation between SAT scores and college performance tends to be greater for women than for men and for Whites and Asians than for African Americans and Hispanics, while SAT scores tend to under-predict female performance in college and over-predict how well African Americans and Hispanics will do, for reasons that are also not completely understood.[34] These findings don't invalidate the SAT, but they do offer yet another reason to interpret SAT results with caution, and to use this test only in conjunction with other criteria.

Is the SHSAT also subject to prediction biases? No one knows, because no one has ever done a predictive validity study of the SHSAT. This offers another reason why this is a serious omission, and one that is not compliant with Standard 7.1 of the *Standards for Educational and Psychological Testing*, which states:

> When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup.[35]

Since there's considerable research suggesting that scores on other standardized tests do differ in meaning across gender and ethnic groups, this standard implies that the same kind of validity studies warranted for all SHSAT examinees should also be conducted separately for these groups. If these studies find that the SHSAT does exhibit prediction bias across groups—i.e., that the same SHSAT score produces systematically different inferences about future performance for members of different groups—this would be still another argument against using this test as the sole criterion for admission.

Conclusion

The SHSAT is widely assumed to produce clear-cut, valid, equitable results. But for many students, this may not be true. Thousands of rejected students have scores that are, for all practical purposes, indistinguishable from those of students who were accepted; the equating system may not fully adjust for differences in the test versions; and the peculiar scoring system benefits some at the expense of others, many of whom don't even know about the system or how to adjust for it because they don't have access to expensive test-prep tutors. All told, on a different day, many students might have flipped to the other side of the admission/rejection line by pure chance—if they'd been assigned a different test version, if the winds of random variation in test scores had blown a bit differently, if slightly different but equally logical scoring had been used, or if they'd been told how the actual scoring system works.

Sometimes, all of the test's idiosyncrasies combine to help one student and harm another. One student might benefit from the nonlinear scaling, a friendly test version, and a score that barely meets a school's cutoff, while another may be disadvantaged by the nonlinear scaling, a less-friendly test version, and a score that misses the cutoff by a hair—well within a standard error of measurement. There were many cases in 2005 and 2006 where this happened; a few examples are shown in Table 8 (following). The decisions here seem arbitrary, especially since there is no validity evidence to support them.

To be sure, no test is "perfect." All face difficulties distinguishing among close candidates. A line must be drawn, and the differences among candidates close to the line are usually tiny, beyond the ability of any test to differentiate. The same is true of other potential admissions criteria, such as grades. That's a big part of why it is contrary to professional testing standards and practice to use any single metric as the sole criterion for admission. According to Standard 13.7 of the *Standards for Educational and Psychological Testing*:

Table 8: Examples

2006 Test						
Version	**Percentile**		**Scaled Score**			
	Verbal	**Math**	**Verbal**	**Math**	**Total**	**Outcome**
E/F	99.0	75.7	317	241	558	Offered seat at Stuyvesant
G/H	97.7	94.2	286	271	557	Missed cut for Stuyvesant
E/F	48.9	98.4	198	312	510	Offered seat at Bronx Sci
G/H	86.5	90.7	249	260	509	Missed cut for Bronx Sci
E/F	25.0	99.3	164	323	487	Offered seat at B'klyn Tech
G/H	80.7	83.5	237	243	480	Missed Cut for B'klyn Tech
2005 Test						
Version	**Percentile**		**Scaled Score**			
	Verbal	**Math**	**Verbal**	**Math**	**Total**	**Outcome**
A/B	85.0	98.9	248	314	562	Offered seat at Stuyvesant
C/D	93.4	98.0	270	291	561	Missed cut for Stuyvesant
A/B	95.9	74.0	279	236	515	Offered seat at Bronx Sci
C/D	91.3	88.1	264	248	512	Missed cut for Bronx Sci
E/F	93.3	61.2	276	210	486	Offered seat at B'klyn Tech
C/D	85.2	82.4	248	235	483	Missed Cut for B'klyn Tech

> In educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision.[36]

Of course, this raises the question of what constitutes "other relevant information," and how to determine if it will "enhance the overall validity of the decision," or even how to define that term. Uncertainty and imprecision are inherent in all potential admissions criteria (test scores, grades, portfolios, etc.). Standard psychometric practice is to choose the criteria that minimize this uncertainty and that allow for the possibility that some students may demonstrate the skills needed to succeed in ways other than captured on a single standardized test. The only systematic, objective way to do this is by conducting predictive validity studies. Such

studies are regularly carried out for tests like the SAT and for high school grades, to help test-makers refine the test, and to help colleges decide how much weight to put on SAT scores, grades, and other factors in making the admission decision.[37] Overwhelmingly, studies like these have found that multiple imperfect criteria, used in tandem, are a better guide to future student performance than a single imperfect criterion.[38] Indeed, it's partly because of batteries of results from predictive validity studies like these that Standard 13.7 was adopted, and that virtually all educational institutions (including high schools in other parts of the country that use the SHSAT or a similar test) do not use a single test as the sole arbiter of the admissions decision.

The admissions procedures at the New York City specialized high schools violate this standard and run counter to these practices. The NYCDOE also ignores the standards by failing to provide detailed information about many aspects of the SHSAT. No evidence is offered to support the equal weighting of verbal and math scaled scores, no IRT-based estimates of the standard error of measurement near cutoff scores are provided, the accuracy of the equating of different test versions is not established, and test takers are not made aware of all the implications of the scoring system. Worse, in all the years the SHSAT has been the lone determinant of admission to these schools, the NYCDOE has never conducted a predictive validity study to see how the test was performing. In fact, the department has never published what specific, measurable objectives the SHSAT is supposed to predict (high school performance, SAT scores, etc.). Absent predictive validity studies, there's no way to know if any test is doing its job; and without well-specified objectives, it's not even clear what that job is—or whether it could be better accomplished by some alternative admissions system. The whole process flies in the face of accepted psychometric standards and practice and reminds us why those standards and practices were established and should be maintained.

Recommendations

- Formal predictive validity studies need to be carried out. At a minimum, these studies should look at the ability of SHSAT scores (separate verbal and math) and middle school grades to predict high school performance. They should also test for prediction bias across gender and ethnic groups. The NYCDOE should release details on how the scaled scores are derived from item response theory—particularly IRT-based estimates of the uncertainty surrounding scores near the admission cutoffs—and on the accuracy of the equating of different test versions. Any inadequacies in equating across test versions need to be corrected.
- Based on the results of these studies and in keeping with generally accepted psychometric standards and practices, a determination should

be made as to what admissions process—including such areas as scoring system, other criteria considered, and weights of these criteria—is most likely to achieve a specific, quantifiable admissions goal in a transparent, equitable way.

- If this study concludes that it is best to use additional admissions criteria besides a standardized test, the New York State law—which says that admissions to these schools must be based solely on a test—would need to be changed.
- Findings such as those presented in this study, and the particular choices of admissions procedures for these schools, should be discussed and deliberated in New York, and an informed decision should be made about future practices. Whatever admissions procedures are established, all applicants should know all of their implications.
- These findings should also contribute to the broader national debate on standardized tests, school admissions, and high-stakes testing such as exit exams.

Notes and References

[1] I would like to thank Keith Gayler, Walt Haney, Kevin Welner, and two anonymous reviewers for helpful comments, and officials at the New York City Department of Education, especially Jennifer Bell-Ellwanger and Susan Waddington, for helpful comments and for their aid in securing the data needed for this study.

[2] The "Hecht-Calandra Act" of 1971 requires that admissions to the specialized high schools be based solely on an exam.

[3] Limited seats for 10th grade (fewer than 200 in total) were offered to the roughly 2,000 ninth graders who took the 9th grade version of the SHSAT in 2005 and 2006.

[4] Zwick, R. (2007, February). *College Admission Testing*. Alexandria: National Association for College Admission Counseling.

[5] Many predictive validity studies have been conducted. A few examples are:

Bridgeman, B., Pollack, J., & Burton, N. (2004). *Understanding what SAT Reasoning Test scores add to high school grades: A straightforward approach*. College Board Research Report 2000-1). New York: College Entrance Exam Board.

Camara, W.J., & Echternacht, G. (2000, July). *The SAT and high school grades: Utility in predicting success in college*. College Board Research Notes RN-10. New York: College Entrance Exam Board.

Geiser, S. & Studley, R. (2001). *UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California.* University of California Office of the President.

Haney, W., Lee, T., & Center, L. (1984, September). *The Predictive Validity of the Secondary School Admissions Test at Selective Independent Secondary Schools.* Princeton: Secondary School Admissions Test Board.

Rothstein, J. M. (2004, July-Aug). College performance predictions and the SAT. *Journal of Econometrics*. Vol. 121, pp. 297-317.

[6] The "bible" of psychometric standards is:

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Other valuable references for the use of tests in college admissions include:

Commission on the Role of Standardized Testing in College Admission (1995, June). *Recommendations of the Commission.* Alexandria: National Association for College Admission Counseling.

Zwick, R. (2007, February). *College Admission Testing*. Alexandria: National Association for College Admission Counseling.

[7] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

[8] Morse, R. (2007). Gold Medal Schools. *US News & World Report*. *http://www.usnews.com/articles/education/high-schools/2007/11/29/gold-medal- schools.html*

[9] Technically, this is not a cutoff score in the traditional sense that it is not set in advance. The admissions process works more like a ranking system with replacement, in which students with the highest score may opt to attend the school, but if they decline, their option passes to students with the next highest score.

[10] About 2.5% of test takers who missed one of the official exam dates for medical or religious reasons took one of two other versions.

[11] Standard t-tests can be used to test for differences in means across test versions. See Mendenhall, W., & Scheaffer, R. L., & Wackerly, D. D. (1981), *Mathematical Statistics with Applications.* (2nd ed.). Boston: Duxbury press. pp. 393-395. Using these tests, we can reject the hypothesis of no difference in mean raw scores between test versions in 8 of 12 comparison pairs in 2006 (all but A/B vs. G/H verbal, C/D vs. E/F verbal, A/B vs. E/H math, and C/D vs. G/H math), and in all 12 comparison pairs in 2005 at the 99% confidence level.

[12] For a primer on item response theory, see Baker, F. (2001). *The Basics of Item Response Theory* (1st ed.) ERIC Clearinghouse on Assessment and Evaluation, College Park, MD: University of Maryland.

[13] On November 11, 2007 I sent an e-mail to officials at the NYCDOE requesting details on the IRT estimates (the test characteristic curves, estimates of their goodness- of-fit, and the test information function). I got no response. I made follow-up requests for that information on December 10 and on February 8, 2008 and have still not received a response.

[14] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, p. 45.

[15] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, pp. 14.

[16] See http://www.erbtest.org/parents/admissions/isee for information on how scores on the ISEE are reported.

[17] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, pp. 14-17.

[18] Unbalanced scores are defined for illustrative purposes in these examples as follows: for Stuyvesant, a difference of at least 11 percentiles between the verbal and math scores, and no higher than 87th percentile in the weaker area; for Bronx Science, a difference of at least 17 percentiles between the verbal and math scores and no higher than the 75th percentile in the weaker area; for Brooklyn Tech, a difference of at least 23 percentiles between the verbal and math scores and no higher than 66th percentile in the weaker area.

[19] New York City Department of Education (2007-2008). *Specialized High Schools Student Handbook.* p. 15. http://schools.nyc.gov/OurSchools/HSDirectory/SpecializedHighSchoolsStudent Handbook.htm

[20] Herszenhorn, D. M. (2005, November 12). Admission Test's Scoring Quirk Throws Balance into Question. *New York Times.*

[21] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, Standard 11.13, pp. 116.

[22] Haney, W. (2000, August). The Myth of the Texas Miracle in Education. *Education Policy Analysis Archives, Vol. 8* (Issue 4), pp. 10-11.

[23] The SHSAT uses a common type of internal-consistency measure called the coefficient alpha reliability, which is given by:

$$R = \frac{k}{k-1} \cdot \{1 - (\frac{S_1^2 + S_2^2 + \ldots S_k^2}{S^2})\}$$

where k is the number of questions, $S_1^2 + S_2^2 + \ldots S_k^2$, is the sum of the variances of the scores on each question, and S^2 is the variance of the total scores. The variance of the total scores will be greater than the sum of the variances of the scores on the individual questions to the extent that there is some positive covariance across the individual questions (i.e., to the extent that if a student gets the right answers on some questions, he or she is more likely to get the right answers on other questions too). The greater that covariance, the more likely it is that the questions are measuring the same thing – the "true" variability across students in a single dimension – thus the more reliable the test is as a gauge of that dimension (the greater is S^2 relative to $S_1^2 + S_2^2 + \ldots S_k^2$, and hence the higher is R). R can vary from 0 (there is no covariance across questions, so the questions are not measuring any common skill), to 1 (all the questions are measuring the same thing, so the test is perfectly internally consistent as a measure of the true variability across students in that single dimension). The internal consistency of the test also rises as the number of questions goes up; the larger the sample of questions, for a given correlation between those questions, the more likely it is that the questions are a representative sample of the uniform skill that the test is designed to measure. That's why longer tests are generally considered more reliable, albeit with diminishing returns. As the number of questions rises, for a given correlation between those questions, the internal consistency of the test increases, but at a decreasing rate.

[24] The SEM is calculated as follows:

$$SEM = S \bullet (1-R)^{.5}$$

where S is the standard deviation of the test scores, and R is the test's reliability. The SEM – which is an estimate of what the standard deviation of the same student's scores would be on repeated tests -- will be less than the standard deviation of scores across all the students on the actual test, S, to the extent the test is reliable (i.e., R is close to 1).

	2006			**2005**		
	Verbal	**Math**	**Total**	**Verbal**	**Math**	**Total**
Standard deviation	49.8	49.1	90.7	49.9	49.4	91.1
Reliability	0.91	0.92	0.95	0.91	0.92	0.95
SEM	15.0	14.0	20.4	15.0	14.0	20.4

[25] Baker, F. (2001). *The Basics of Item Response Theory* (1st ed.) ERIC Clearinghouse on Assessment and Evaluation, College Park, MD: University of Maryland.

[26] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, p.35.

[27] The author made several e-mail requests for IRT-based details and received no response: no data and no explanation for why the data were not forthcoming.

[28] Baker, F. (2001). *The Basics of Item Response Theory* (1st ed.) ERIC Clearinghouse on Assessment and Evaluation, College Park, MD: University of Maryland. pp. 89-91.

[29] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, p. 57.

[30] Standard t-tests can be used to test for differences in means across test versions. See Mendenhall, W., & Scheaffer, R. L., & Wackerly, D. D. (1981), *Mathematical Statistics with Applications.* (2nd ed.). Boston: Duxbury press.

[31] Cohen's "d" is the difference between the mean scaled scores across two versions, divided by the root mean square of the standard deviations of the scaled scores on those two versions. See Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Lawrence Earlbaum Associates. See Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press., for how to construct confidence intervals around effect sizes.

[32] Jencks, C., Phillips, M. (Eds.) *The black-white test score gap.* Washington, DC: Brookings Institution Press

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report 93-1). New York: College Entrance Examination Board.

Young, J.W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In Zwick, R. (Ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions.* New York: RoutledgeFalmer

Zwick, R. (2007, February). *College Admission Testing*. Alexandria: National Association for College Admission Counseling.

[33] Cole, N.S., & Moss, P.A. (1989). *Bias in Test Use*. In Linn, R.L. (ed.) *Educational Measurement* (3rd ed.). New York: American Council on Education/Macmillan. pp. 201-219.

[34] Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Prediction of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (College Board Report 2000-1). New York: College Entrance Examination Board.

Ramist, L., Lewis, C. & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report 93-1). New York: College Entrance Examination Board.

Young, J.W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In Zwick, R. (Ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions*. New York: RoutledgeFalmer

Zwick, R. (2007, February). *College Admission Testing*. Alexandria: National Association for College Admission Counseling.

[35] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, p. 80.

[36] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, p.146.

[37] Commission on the Role of Standardized Testing in College Admission (1995, June). *Recommendations of the Commission.* Alexandria: National Association for College Admission Counseling, recommends that colleges use a variety of admissions criteria, guided by predictive validity studies.

[38] Bridgeman, B., Pollack, J., & Burton, N. (2004). *Understanding what SAT Reasoning Test scores add to high school grades: A straightforward approach.* College Board Research Report 2000-1). New York: College Entrance Exam Board.

Camara, W.J., & Echternacht, G. (2000, July). *The SAT and high school grades: Utility in predicting success in college.* College Board Research Notes RN-10. New York: College Entrance Exam Board.

Geiser, S. & Studley, R. (2001). *UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California.* University of California Office of the President.

Haney, W., Lee, T. & Center, L. (1984, September). *The Predictive Validity of the Secondary School Admissions Test at Selective Independent Secondary Schools.* Princeton: Secondary School Admissions Test Board.

Rothstein, J. M. (2004, July-Aug). College performance predictions and the SAT. *Journal of Econometrics*. Vol. 121, pp. 297-317.