FAIRNESS TO GIFTED GIRLS: ADMISSIONS TO NEW YORK CITY'S ELITE PUBLIC HIGH SCHOOLS

Jonathan Taylor

Hunter College Gender Equity Project, 695 Park Avenue, Room 1032E, New York, NY 10065, USA; Tel./Fax: 646-861-2910, E-mail: Jontaylor5819@gmail.com

The use of test scores in school admissions has been a contentious issue for decades. In New York City's elite public high schools, it has been particularly controversial because of disproportionate representation by ethnicity. Underrepresentation of girls has received less attention. This research compared the predictive validity and gender bias of the admissions criterion, the Specialized High School Admissions Test (SHSAT), with that of seventh grade GPA, a possible additional criterion. SHSAT ($r^2 = 0.20$) predicted high school grades less precisely than GPA7 ($r^2 = 0.44$) and underpredicted girls' grades in all academic domains and specific courses analyzed. Girls were overrepresented in the upper tail of STEM course grades. Simulated admissions using an index combining SHSAT and GPA7 suggest that different admissions criteria might improve the quality of the admitted cohort, increase diversity, and be gender-fair.

KEY WORDS: *test validity, gender bias, admissions, Specialized High School Admissions Test*

1. INTRODUCTION

The use of test scores in school admissions has been a contentious issue for decades. For eighth grade students in New York City, the two-and-a-half hour Specialized High School Admission Test (SHSAT) looms large because of the cascade of benefits that may result from admission to the top NYC schools. Children admitted may not only receive a superior education but are also likely to have better access to elite colleges, professional and graduate schools, and eventually to better employment. Stuyvesant High School and Bronx High School of Science are renowned for the number of finalists and winners they have produced in the Westinghouse/Intel Science Talent Search. The two schools combined have produced at least twelve Nobel prize winners and leaders in many fields, including business, politics, and the arts.

Admission to New York City's elite public high schools has been controversial because of the underrepresentation of Hispanic and African American students. Underrepresentation of girls has received less attention. These schools use the score on one test, the SHSAT, as the sole admissions criterion. The current procedure has resulted in the admission of classes that do not reflect the gender ratio of applicants. Although 51.2% of the applicants in 2014 were female, only 44.6% of those admitted were.

Proponents of the exam defend it as objective and meritocratic, while opponents contend that when used without consideration of school grades or other factors, it is not an appropriate criterion. The test is unquestionably objective. However, when merit is defined as achievement in school, the question of whether the test is meritocratic is, in part, an empirical question which can be answered with existing data. Because the goal of the test (Calandra and Hecht, 1971) is to

1072-8325/19/$35.00

identify academic accomplishment, academic criteria were used in this research.

1.1 Rationale for Research

Implicit in the use of the SHSAT to select students for the specialized high schools is the assumption that the test is a good predictor of who will succeed in these schools and that it predicts equally well for all subgroups. The city and test developer have been remiss in not having validated the exam long ago. According to the American Educational Research Association standards, "evidence of the validity of a given interpretation of test scores for a specified use is a necessary condition for the justifiable use of the test" (AERA, 2014, p.11).

If the SHSAT results in severe underrepresentation of African American and Hispanic students, and of girls, the only justification, legally and ethically, the city can have for its continued use is if the overall predictive validity is high. In the only previous study of SHSAT validity, Taylor (2015) found relatively low predictive validity and underprediction of girls' grades in the cohort that took the exam in 2008 and attended the three largest schools: Stuyvesant, Bronx Science, and Brooklyn Tech. This investigation represents an attempt to replicate that research and to address shortcomings of that study, which only included three schools and did not have data on the criterion variable, freshman grade point average (FGPA), for students who did not attend a specialized high school. Estimates were therefore artificially low due to range restriction. This study was also designed to test the hypothesis that middle school grades would be superior to SHSAT scores as a predictor of high school success.

Because of the underrepresentation of females in STEM areas, an additional focus is the representation and performance of girls in science and math courses. The possibility that any underprediction of girls' grades is due to enrolling in less challenging courses is examined, as is the representation of girls in the upper tail of the grade distribution in STEM subjects.

1.2 Admissions to Selective Public High Schools Nationwide

New York is not the only city that has had to address the tension between selectivity and equal representation. However, the admissions process and the demographic results in New York are in stark contrast to those in many other cities. Nationwide there are 165 selective public high schools (Finn and Hockett, 2012). Admissions policies vary among these schools: almost 80% give consideration to prior academic performance; state or district achievement tests are a factor in admissions to 60%; 55% give weight to student essays; 52% to teacher recommendations; and 40% to a proprietary exam developed for the school. New York is unique in its reliance on one exam to select students.

1.3 Research Questions and Hypotheses

In 2015, Taylor (2015) found that the overall predictive validity of the SHSAT was low for the cohort admitted in 2009 to Stuyvesant High School, Bronx High School of Science, and Brooklyn Technical High School. It was expected that these findings would be replicated for the 2014 cohort and would be extended to all NYC public high schools. Furthermore, by investigating the predictive validity for all NYC students who took the SHSAT, the problem of range restriction in Taylor's 2015 analysis was avoided.

1.4 Gender Predictions

AERA standard 3.7 states that the test user is "responsible for evaluating the possibility of differential prediction for relevant subgroups for which there is prior evidence or theory suggesting differential prediction" (AERA, 2014, p. 66). Prior evidence with respect to the validity of the SAT suggests that possibility. Mattern et al. (2008) found that the SAT-Math test (SATM) underpredicted women's FGPA, with an effect size of 0.17 standardized residuals. A literature review by Stricker et al. (1993) of earlier research also showed that SAT scores underpredict women's grades.

In a review of the literature on the relationship between SATM scores and math grades, Wainer and Steinberg (1992) found that in general, although women had lower SATM scores than men, they earned higher grades in math courses. To test the hypothesis that this difference was attributable to men enrolling in more difficult courses rather than to bias, grades and SAT scores were obtained for 47,000 men and women who attended 51 different colleges. Women who earned the same math course grade as men had lower SATM scores in general. When matched for course subject and grade earned, women's SATM scores ranged from 21 to 55 points lower than men. Similarly, in a study by Subotnik and Strauss (1998), despite lower SATM scores, women achieved grades equal to men on the advanced placement (AP) calculus exam.

Research has repeatedly shown that women do poorly relative to men on multiple choice questions, which may explain, in part, the underprediction of grades by the SAT. Tannenbaum (2012) attributed the gender gap in SAT scores to girls on average being more risk-averse, and therefore not guessing as often as boys, although guessing on the SAT was a useful strategy. He estimated that 40% of the gender gap in scores could be accounted for by risk aversion. Other researchers have reported similar patterns (Gallagher et al., 2000; Baldiga, 2014).

In an investigation of gender differences on AP exams, Mazzeo et al. (1993) found a male advantage on multiple choice questions. Women, however, outperformed men on constructed answer questions, leading the authors to hypothesize that multiple choice and constructed answer questions tap different competencies on which there are real gender differences. Bennett (1993) also hypothesized that constructed answer questions, particularly essays, require multiple abilities and may reflect a more complex understanding of material than do questions that require only the selection of a correct choice. Demars (1998, 2000) reported evidence for the item format effect, particularly in the top of 5% of scorers, which may be relevant for specialized high school admissions. Furthermore, she suggested "that the two formats are measuring something slightly different (and that 'something' is also related to gender)" (Demars, 2000, p. 69). Bonner (2013) did not find a gender-item format effect. However, she did find that the student approaches to multiple choice questions did not always involve a good understanding of the problem.

The SAT writing section is the only SAT subtest on which females outperform males (Mattern et al., 2008). Although females have a small advantage on the multiple choice section (d = 0.04), the advantage on the essay portion is more than six times as large (d = 0.25).

Traub and MacRury (1990, reported in Mazzeo et al., 2013) reviewed gender differences on AP exams, the California bar exam, and an English placement exam used at California state universities. On the multiple choice section of the English placement exam, there was an effect size favoring males of 0.05, while females had an advantage of 0.39 on the essay portion. On all 11 of the AP exams studied, males were superior on multiple choice sections. Females had the advantage on constructed response portions on 10 of the exams.

In 2015, Taylor (2015) found that the SHSAT underpredicted FGPA for girls, a finding consis-

tent with the SAT research discussed. It was expected that this research would replicate those results.

1.5 Middle School Grades

The availability of applicants' middle school grades makes it possible to add a research question that is important to policy-making: Does the SHSAT predict high school success as well as seventh grade GPA (GPA7) does? Does GPA7 underpredict girls' grades? GPA7 is compiled over a full year across the full range of academic domains and is based on many different methods of assessment. In contrast, the SHSAT was one 2½ hour test. With the exception of five paragraphs with scrambled sentences that students must place in the proper sequence, all SHSAT questions were multiple choice. Because GPA7 is based on a wider range of academic skills assessed in a greater variety of ways, it is hypothesized that it is a better predictor than the SHSAT and statistically less biased against girls.

2. METHODS

2.1 Participants

In the fall of the eighth grade, students who wish to apply to the specialized New York City high schools must take the SHSAT. Of the approximately 81,000 eighth graders in New York City, 27,818 students took the SHSAT in 2013 for entry to high school in 2014. Students taking the exam came from 560 public middle schools. In addition, 3411 students came from unidentified private schools. At the time of the test, prospective students listed schools in order of preference. Students were then ranked in order of their scores and admitted according to those preferences, resulting in different cutoffs for each school. In 2013, scores ranged from 40 to 701, with school cutoffs (Table 1) ranging from 479 to a high of 559.

TABLE 1: SHSAT cutoff scores and number of Discovery students

	Cutoff	Low	High
Stuyvesant	559	559	701
Bronx Science	517	517	680
Brooklyn Tech	486	465	680
SI Tech	506	474	667
Lehman	506	506	646
Queens	511	478	645
CCNY	504	477	649
Brooklyn Latin	479	475	657

2.2 Data

SHSAT scores were provided by the NYC Department of Education (DoE) for all eighth grade students who took the exam in 2013. Additionally, demographic data were reported, including ethnicity, gender, and type of middle school attended. Middle and high school grades, as well as seventh

grade achievement test scores, were provided for all students who attended NYC public schools.

Gender was unknown for seven private school applicants. Ethnicity was unknown for 2736 students, all but two of whom were private school students. Ethnic and gender identification were not missing for any students who actually attended a specialized high school.

2.3 Constructed Variables

GPA: Grades received from the DoE were used to compute GPAs for seventh grade (GPA7), eighth grade (GPA8), and ninth grade (FGPA) students. For this purpose, grades in nonacademic courses such as physical education and performance arts were excluded.

FGPA Categories: To determine the relationship between SHSAT and FGPA at different levels of achievement, students at each school were assigned to six FGPA categories:

(1) below 75, (2) 75–80, (3) 80–85, (4) 85–90, (5) 90–95, and (6) 95+.

Admission Index: A hypothetical admissions index was created weighting SHSAT and GPA7 by coefficients found in the regression of FGPA. Because GPA7 was not available for students from private middle schools, this index was based only on public school students.

3. RESULTS

Of the eighth graders who sat for the SHSAT in 2013, 1383 retook the exam in 2014. Ninth graders were given a somewhat more difficult form of the exam, requiring somewhat higher level math and vocabulary (Princeton Review, 2018). The correlation between eighth and ninth grade scores ($r = 0.758$) indicated that the exam is highly reliable.

GPA7 was correlated with GPA8 as a measure of the consistency of grading. The correlation for all public school students in the city ($n = 64{,}606$) was 0.841, an indication that the skills and assessments were consistent over time. GPA was in fact a more reliable measure than the SHSAT, despite being the product of a variety of subjective methods of assessment done by many different teachers, in contrast to the SHSAT which consists exclusively of objective questions.

4. OVERALL VALIDITY

4.1 Linearity and Homogeneity of Variance

The plot (Fig. 1) of FGPA × SHSAT scores suggests a nonlinear relationship with very little variance at the top of the SHSAT scale. A vertical line has been drawn at SHSAT = 479 to indicate the lowest admissions cutoff for any of the schools. Students with extremely high SHSAT scores generally also had high grades. In contrast, there was a great deal of variance in the portion of SHSAT scores around the cutoffs for admission (479–559), with FGPAs ranging from around 50 to 100, indicating that the SHSAT was a very imprecise predictor in the crucial decision range.

4.2 Regression of FGPA on SHSAT

Statistical analyses described below are based on freshman year GPA (FGPA). This parallels common validity studies in which SATs are related to college freshman GPAs. The logic behind this is that freshmen are more likely to enroll in similar courses. Additionally, not all students who entered remained in these schools through graduation. Taylor (2015) found that the mean

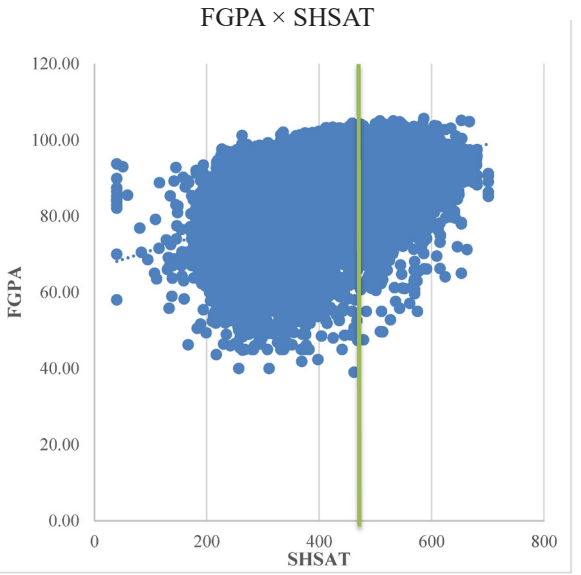

FIG. 1: FGPA × SHSAT - All applicants admissions cutoffs range from 479 to 559. Vertical line is at lowest cutoff.

ninth grade GPA of students in the three large specialized high schools who did not remain for four years was very low, 13.4 points lower than for those who remained. Using cumulative GPA rather than FGPA would remove the lowest achieving students from the cohort.

Table 2 displays the results of the regression of FGPA on SHSAT scores for all students who took the exam, and separately within each of the specialized schools. In addition, the number of FGPA points by which girls' grades are underpredicted is shown, along with the standard error of estimate (SEE).

TABLE 2: Percent of variance in FGPA predicted by SHSAT, SEE, and gender underprediction

	N	**SHSAT**	**SEE**	Underprediction
Applicants	22,576	20.0%	8.49%	4.20%
Stuyvesant	823	5.7%	5.97%	3.55%
Bronx Science	744	3.2%	5.37%	5.06%
Brooklyn Tech	1345	3.0%	7.23%	4.87%
SI Tech	312	9.8%	5.56%	3.93%
Lehman	94	2.4%	7.19%	4.17%
Queens	105	4.8%	7.23%	7.58%
CCNY	114	14.0%	7.42%	4.16%
Brooklyn Latin	94	6.3%	6.61%	4.77%

Although all regressions of FGPA × SHSAT were highly significant ($p < 0.001$), the variance in FGPA predicted was very small within the largest schools, (e.g., 3.2% at Bronx Science, 5.7% at Stuyvesant, and 3.0% at Brooklyn Tech) and ranged from 2.4% to 14% at the smaller specialized schools. To avoid the problem of range restriction, FGPA was also regressed against SHSAT for the 22,576 students for whom both were available. In this diverse group, SHSAT scores predicted 20% of FGPA variance (Table 2).

Because admissions decisions at the specialized high schools are based on total SHSAT scores, the above regressions were done on the total score. When separate verbal and math SHSAT scores were entered together into the regression, prediction improved by a trivial amount, from 20.0% to 20.1%, and underprediction of girls' scores increased from 4.20 to 4.27.

Dividing the sample in each school into FGPA categories was only done at the three largest schools, where the number of students in each category was meaningful. Mean SHSAT scores hardly differ among students with FGPAs ranging from 75 to 90 (Fig. 2). For example, at Stuyvesant, those with FGPAs below 75 had SHSAT scores only two points lower on a 701 point scale than those with FGPAs of 85–90. At Bronx Science, SHSAT scores of students with FPGAs below 75 were in fact higher than all but the 95+ group. In general, larger differences in SHSAT

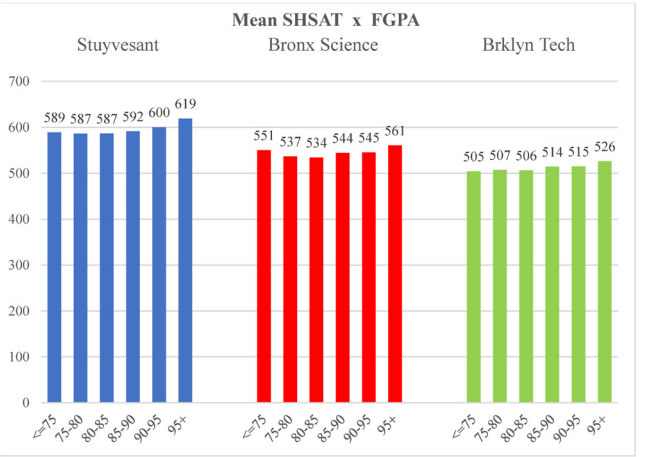

FIG. 2: Mean SHSAT × FGPA

scores emerged for students with FGPAs above 90, especially above 95.

The R^2 estimates are not the only evidence of the insufficient predictive value of the SHSAT. The SEE gives a more easily interpretable understanding of the precision of FGPA prediction based on SHSAT scores. This may be especially important in the range around the cutoff scores that determine admission. The high SEE (Table 2), which ranged from 5.49 to 8.37, suggest a very large 95% confidence interval of anywhere from 22 to 34 points. With such large confidence intervals, it seems clear that the exam is not a precise predictor. Even using a narrower 68% confidence interval at the school with the smallest SEE, Bronx Science, the margin of error in predicting FGPA is 5.4 points in either direction. A student predicted to have a FGPA of 85 could easily have a FGPA of 80 or 90. Furthermore, the plot discussed above shows that the greatest imprecision occurs in the lower parts of the distribution in each school, which is the decision range for admissions. When broken down into categories of FGPA, the lesser accuracy of prediction confirms the graphic displays in the plot.

4.3 Gender Predictions

Girls scored an average of 12.5 points lower than boys on the SHSAT, with most of the difference (10.1) resulting from lower scores on the math portion of the test. Standard deviations were also somewhat lower for girls, who were underrepresented in the top 3% of test scores (59.5%–40.5%).

When gender was included in the regression equations of FGPA on SHSAT, gender coefficients all had positive signs, indicating that course grades of girls are underestimated by the SHSAT (Table 2). On a 100-point scale, the underestimation ranged from 3.55 points to 7.58 points ($p < 0.0001$ for all).

Reverse regressions of SHSAT on FGPA found that girls achieved grades equal to boys who had higher SHSAT scores. At Stuyvesant, the difference was 6.6 SHSAT points; Bronx Science 5.8; Brooklyn Technical 9.0. For the entire sample of 22,576 the difference was much larger, 29.0 points.

Dempster (1988) has suggested that for a regression to be unbiased, *e,* often referred to as an error term in regression, but which Dempster describes as representing "unobserved characteristics," must have equal means for both genders, assumed to be zero. However, when the generic regression equation was used to predict FGPA, the gender difference in residuals was highly significant ($p < .001$). The mean residual for males was -2.28, whereas for females it was 1.91, another indication of underprediction of girls' scores.

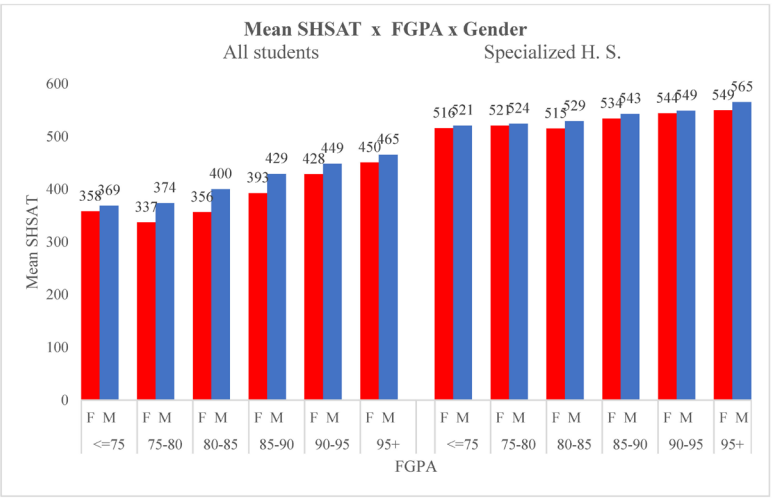

FIG. 3: Mean SHSAT × FGPA × gender

Further confirmation of these results can be found in Fig. 3, which displays results using the FGPA categories described above. All twelve gender comparisons indicate that girls earned the same grades as boys who had higher SHSAT scores. A two-way ANOVA test found significant gender effects (Table 3, all applicants: $F = 979.3$; $p < 0.001$; Table 4, specialized: $F = 389.2$; $p < 0.001$). Across the eight elite schools, girls outnumbered boys in the highest achieving group (95+) by 350:239, despite their overall underrepresentation in the schools (58% to 42%) and the upper tail of SHSAT scores. In the lowest FGPA category, below 75, boys outnumbered girls by a factor of more than six to one.

TABLE 3: FGPA × mean SHSAT × gender of all applicants

	Gender	SHSAT	*N*	Std. Dev.
≤ 75	F	358	3981	82.1
	M	369	4809	86.1
75–80	F	337	1066	71.4
	M	374	1425	86.0
80–85	F	356	1552	73.0
	M	400	1906	88.7
85–90	F	393	2408	86.5
	M	429	2347	89.2
90–95	F	428	3299	84.7
	M	449	2145	87.2
95+	F	450	1924	74.7
	M	465	923	83.0
Total	F	391	14230	89.4
	M	403	13555	93.7

Gender effects ($F = 979.3$; $p < 0.001$)

TABLE 4: FGPA × mean SHSAT × gender specialized high schools

	Gender	SHSAT	*N*	Std. Dev.
≤ 75	F	516	26	30.8
	M	521	163	38.2
75–80	F	521	44	35.1
	M	524	201	37.2
80–85	F	515	108	35.8
	M	529	364	38.2
85–90	F	534	343	41.4
	M	543	602	42.0
90–95	F	544	683	41.8
	M	549	609	49.1
95+	F	549	350	42.0
	M	565	239	48.4
Total	F	540	1554	42.1
	Total	541	3732	44.1

Gender Effects ($F = 389.2$; $p < 0.001$)

4.4 Gender Differences across Academic Domains

A variety of analyses were done to rule out the possibility that the underprediction in girls' grades resulted from differences in course selection. Because the underrepresentation of females in STEM majors and occupations has been a source of concern, it was of special importance to examine the relationship between SHSAT scores and grades in STEM areas and to ascertain whether girls in New York City's specialized high schools enroll as frequently as boys in challenging STEM courses.

Gender comparisons were done across different academic domains for all students who took the SHSAT and attended NYC high schools. Similar comparisons were done for those who actually attended specialized high schools. Finally, analyses were done at Stuyvesant High School, which, because it is the most selective of the schools, might possibly have the most difficult courses. Overall, girls (Table 5) earned significantly higher grades (84.4) in STEM courses than boys (81.7). The proportion of students enrolled in STEM courses who were female (50.7%) almost exactly matched the proportion of students who took the SHSAT (51.2%), which underpredicted their STEM grades by an average of 2.99 points. The results across specialized high schools were similar. Girls had higher mean STEM grades (89.6) than boys (86.7) and were represented in proportion (41.8%) to their representation of students in these elite schools (41.6%). Results across all math courses did not differ from results in science courses (Table 5).

TABLE 5: Gender comparisons and underprediction of girls' grades in STEM

STEM	Gender	% of Sample	% of Grades	Mean	Underprediction
All SHSAT	F	51.2%	50.7%	84.4	3.25
	M	48.8%	49.3%	81.7	
Specialized	F	41.6%	41.8%	89.6	2.99
	M	58.4%	58.2%	86.7	
Science					
All SHSAT	F	51.2%	50.9%	84.8	3.14
	M	48.8%	49.1%	82.2	
Specialized	F	41.6%	41.1%	89.1	2.87
	M	58.4%	58.9%	86.3	
Math					
All SHSAT	F	51.2%	50.9%	83.8	3.47
	M	48.8%	49.1%	81.1	
Specialized	F	41.6%	41.0%	89.2	3.29
	M	58.4%	59.0%	86.0	

Grades in specific courses, rather than across an entire domain, offer the best test of the hypotheses that girls are less capable of succeeding in STEM subjects and that the underprediction of their grades by standardized tests such as the SHSAT is due to enrollment in easier courses. At Stuyvesant High School (Table 6), girls, on average, earned better grades than boys in each

of the ten specific STEM courses analyzed, including geometry, integrated algebra, biology, and physics. In four of the ten courses, girls were represented in greater numbers than in the overall Stuyvesant ninth grade cohort. Furthermore, their grades were underpredicted by the SHSAT. All grade comparisons were highly significant ($p < 0.001$) except for those in integrated algebra 3 ($p = 0.067$), integrated algebra 4 ($p = 0.062$), honors analytic geometry (n.s.), enhanced Euclidean geometry ($p = 0.072$), and physics (n.s.).

TABLE 6: Stuyvesant High School: Gender Comparisons and Underprediction × STEM Course

	Gender	*N*	% of Grades	Mean	*p*	Underprediction
Analytic geometry	F	299	44.8%	90.4	< 0.001	2.93
	M	369	55.2%	87.5		
Honors analytic geometry	F	36	37.1%	94.5	0.296	0.89
	M	61	62.9%	93.5		
Euclidean geometry	F	303	45.0%	88.9	0.024	1.36
	M	371	55.0%	87.5		
Euclid geometry enhanced	F	34	34.7%	94.4	0.072	2.17
	M	64	65.3%	92.3		
Integrated algebra 3	F	37	39.8%	86.7	0.067	3.99
	M	56	60.2%	82.8		
Integrated algebra 4	F	38	39.2%	84.1	0.062	5.08
	M	59	60.8%	79.2		
Modern biology 3	F	256	42.0%	88.0	< 0.001	2.76
	M	353	58.0%	85.1		
Modern biology 4	F	242	41.4%	90.3	< 0.001	4.03
	M	342	58.6%	86.2		
Physics 1	F	24	48.0%	89.6	0.345	2.80
	M	26	52.0%	87.4		
Physics 2	F	24	48.0%	91.4	0.304	2.89
	M	26	52.0%	89.0		

In 2005, Lawrence Summers, then president of Harvard, suggested that the reason there were fewer women in STEM fields was due to their underrepresentation in the upper tail of the distribution of ability in STEM subjects. Because one of the goals of the specialized high school admissions process is to identify the truly exceptional, it is important to determine if the higher mean grades earned by girls in STEM classes were achieved without those extreme high achievers. It would be possible for girls to have higher mean grades but fewer exceptional grades. However, this is not the case. Girls represented only 41.6% of the cohort and only 40.5% of the top 3% of SHSAT scores, but in STEM courses they earned 50% of the grades of 95 or better and

29.7% of grades 80 or lower.

In non-STEM subjects, girls demonstrated somewhat larger superiority in grades, with greater underprediction by the SHSAT (Table 7). The mean grade in non-STEM courses for all female SHSAT-takers in NYC public high schools was 87.1, compared to 82.8 for male students. The SHSAT predicted non-STEM grades for girls that were 4.77 points lower than actually achieved. In the specialized high schools, girls (91.1) outscored boys by a similar margin (86.5), with underprediction of 4.68 points. These differences were found across the range of non-STEM subjects (humanities, languages, and social studies). However, the greatest differences and underprediction existed in language courses and the smallest in social studies.

TABLE 7: Gender comparisons and underprediction of non-STEM grades

Non-STEM	Gender	% of Sample	% of Grades	Mean	Underprediction
All SHSAT	F	51.2%	51.4%	87.1	4.77
	M	48.8%	49.0%	82.8	
Specialized	F	41.6%	41.2%	91.1	4.68
	M	58.4%	58.8%	86.5	
Humanities					
All SHSAT	F	51.2%	50.7%	86.6	5.00
	M	48.8%	49.3%	82.1	
Specialized	F	41.6%	41.3%	91.6	4.97
	M	58.4%	58.7%	86.6	
Languages					
All SHSAT	F	51.2%	52.7%	89.3	5.48
	M	48.8%	47.3%	84.5	
Specialized	F	41.6%	41.2%	91.8	5.45
	M	58.4%	58.8%	86.3	
Social Studies					
All SHSAT	F	51.2%	51.3%	86.0	3.89
	M	48.8%	48.7%	82.7	
Specialized	F	41.6%	41.2%	89.9	3.48
	M	58.4%	58.8%	86.5	

4.5 Seventh Grade GPA as Predictor

The proportion of variance in FGPA predicted by GPA7 (Table 8) far exceeded the proportion predicted by SHSAT scores. For example, at the three largest of the schools, Stuyvesant, Bronx

Science, and Brooklyn Tech, SHSAT scores were associated with small percentages of FGPA variance (5.7%, 3.2%, and 3.0%, respectively). In contrast, the associations with GPA7 were 6–12 times as large (35.4%, 35.2%, and 37.0%). Comparisons for all students who took the SHSAT also revealed large differences, with SHSAT predicting 20.0% of the FGPA variance for the entire group and GPA7 predicting 43.8%, while reducing underprediction of girls' grades from 4.2 FGPA points to 1.3 points. The SEE was also smaller when GPA7 was used to predict FGPA. The combination of SHSAT and GPA7 produced a very small increment over GPA7 alone, raising the predictive validity to 44.1% (Table 8).

TABLE 8: Percent of variance in FGPA predicted by GPA7: combined, SEE, and gender underprediction

	N	**GPA7**	**SEE**	**Underprediction**	**Combined**	**SEE**	**Underpredicted**
Applicants	20,018	43.8%	7.27	1.33	44.1%	7.20	1.59
Stuyvesant	726	35.4%	4.95	1.64	37.3%	4.88	1.84
Bronx Science	662	35.1%	4.29	1.46	35.6%	4.28	1.61
Brooklyn Tech	1207	37.0%	5.88	2.30	38.8%	5.80	2.51
SI Tech	280	26.2%	4.90	2.42	32.5%	4.69	2.89
Lehman	73	44.7%	5.09	0.44	46.6%	5.04	0.56
Queens	97	55.9%	4.88	3.52	56.2%	4.89	3.90
CCNY	91	40.1%	5.99	3.69	47.1%	5.65	3.68
Brooklyn Latin	145	40.7%	5.21	1.69	43.1%	5.13	1.64

In contrast with Fig. 2, which displays SHSAT means for the FGPA categories, Fig. 4 shows a clear linear relationship between grades in seventh grade and grades in ninth grade, further confirmation that middle school grades are a better predictor than SHSAT scores.

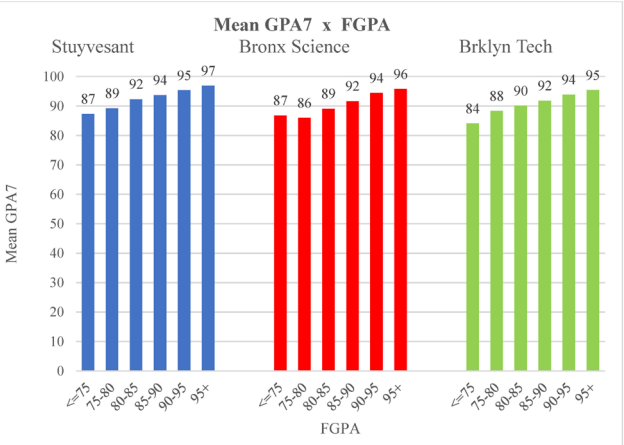

FIG. 4: Mean GPA7 × FGPA

4.6 Simulated Admissions

A hypothetical admissions index was constructed weighting SHSAT and GPA7 by coefficients from the regression of FGPA (20.243 + 0.008 *SHSAT + 0.717 *GPA7). The index was then used to simulate admissions to the specialized high schools. Because middle school grades were not available for the students from private school feeders, simulations only included public school students. For this purpose, public school applicants were ranked in order of their scores on the index and then "admitted" to the schools based on the choices indicated at the time of the exam. The number of students assigned to each school matched the number of public school students actually admitted and resulted in substantially different gender ratios (Table 9), with the proportion of girls admitted rising from 45% to 62%, a difference of 716 girls. At Stuyvesant, using the index, the representation of girls increased from 44% to 65%, a difference of 177 girls. If GPA7 were the sole admissions criterion, the proportion of girls would have been 68%, a difference of 202 girls.

TABLE 9: Female simulated and actual admissions

	Simulated	**Actual**	***N***
Lehman	70%	61%	122
Bronx Science	65%	47%	824
Brooklyn Latin	66%	52%	383
CCNY	62%	32%	146
Queens	67%	43%	140
SI Tech	58%	45%	297
Stuyvesant	65%	44%	825
Brooklyn Tech	57%	44%	1618
Total	62%	45%	4355

Use of this index would also have resulted in substantially different ethnic proportions. At Stuyvesant, for example, an additional 10 African American, 25 Hispanic, and 61 white students would have been admitted. Asian students, though reduced in numbers from 78%, would still comprise 66% of those admitted. Similar shifts would occur in the entire cohort of students admitted to specialized high schools, with increases of 40 African American, 209 Hispanic, and 205 white students. With 49% of the hypothetically admitted class, Asian students would still constitute the largest segment and would be admitted in numbers far exceeding their proportion of applicants (32%).

Because the hypothetical criterion predicts far more of the variance in FGPA than the actual criterion, with a smaller standard error of estimate, its use should not dilute the quality of the entering class. In fact, it might even result in a stronger cohort, while simultaneously increasing diversity and gender equity.

5. DISCUSSION

Analyses of the data make clear that the SHSAT measures an ability which is stable across time, and the ability measured contributes to success in high school. However, as a sole criterion for admission it is deficient and is arbitrary around the cutoff scores. Course grades earned in the sev-

enth grade are a far better predictor than the SHSAT. It is ironic that a standardized test designed to be a uniform metric does not predict as well as past school performance. The SHSAT represents a 2½ hour sample of a limited range of skills and knowledge. In contrast, GPA7 reflects a full year of student performance across the full range of academic subjects. An exam which relies almost exclusively on one method of assessment may fail to measure abilities that are revealed by the variety of assessment methods that go into course grades. Additionally, middle school grades may capture something important that the SHSAT fails to capture: motivation.

According to Cleary (1968), a "test is biased if the criterion score predicted… is consistently too high or too low for members of the subgroup" (p. 115). In this view, the SHSAT is biased against girls. Regression equations, analysis of residuals, reverse regression equations, and groupings by GPA categories all reveal the same phenomenon: girls earn higher grades than boys with equal SHSAT scores.

6. LIMITATIONS

Because the results presented in this paper are based on an analysis of data from one cohort, more general inferences may not be justified. However, the results with respect to the SHSAT validity and gender are very close to those found previously (Taylor, 2015).

Very few student IDs were missing for the students who actually enrolled in the three specialized high schools. However, larger numbers were missing in the full sample of students who took the test. Furthermore, because those students were private school students, they were not missing at random. The inability to link those students to grades may limit the inferences that can be made for the full sample. The simulated admissions results were similarly limited to students from public school feeders. Nevertheless, because public school student applicants represented 88% of the total applicant pool and 92% of the students attending the specialized high schools, the missing data probably do not meaningfully compromise the results of this study.

It is important to note that the approach outlined is limited to examining GPA as the sole metric of success at a specialized high school. There are certainly other criteria for success, such as artistic achievement or citizenship, and those may have implications for the admissions process. However, they are beyond the scope of this research. In any case, students with talents not measured by GPA would probably not be identified by the SHSAT. In order to identify these students, a more holistic admissions process employing multiple criteria is probably required.

7. POLICY IMPLICATIONS

The purpose of this research was to provide guidance for evidence-based policy decisions with respect to the admissions process to the specialized high schools. In light of the underrepresentation of female, African American, and Hispanic students, the imprecision in prediction found for the SHSAT may make it difficult for the city to justify its continued use as the sole gatekeeper to New York's elite high schools. The fact that seventh grade GPA is a far better predictor of high school success and is relatively gender-fair suggests that it should be an important part of the admissions process. While the data support that use, they also do not point to a unique alternative. Before choosing a policy, it may be important for the DoE to better define what it means to be a successful student in a selective high school, which in turn may require careful consideration of the schools' mission. With a definition in place, it may then be possible to develop a screening process that will more successfully select students who can best enable the schools to fulfill

their missions. This may also require a redesign of the SHSAT, reducing its reliance on multiple choice questions.

All other selective high schools in the country employ multiple criteria for admissions. Elite universities could admit classes based solely on SAT scores, but most elect to consider additional factors in selecting students. They do so because they recognize that such tests provide limited information, and because they believe that the classes admitted based on multiple criteria are overall more likely to fulfill institutional missions. Yet, New York City has clung to the belief that the high quality of students admitted to the specialized high schools can only be maintained by the continued exclusive reliance on an admissions exam.

Various alternatives to the current specialized high school admission procedures have been suggested as correctives to what is seen as unfairly disproportionate representation of minorities and girls at these schools. Given the evidence that the exam more accurately predicts achievement at the very top of the SHSAT scale, the DoE might consider a policy which admitted all students above a certain high cutoff, 650 for example, and filled the remaining seats by consideration of multiple criteria. Although this research does not directly address these alternatives, it may help inform policy makers in considering those options.

With changes to the SHSAT and consideration of additional criteria, it may be possible to select a group of students who will be more representative of the community the school system serves, and the pool of students who apply, without sacrificing the quality for which New York City's specialized high schools are so justifiably famous.

REFERENCES

AERA. (2014). *Standards for education and psychological tests.* Washington, DC: American Educational Research Association.

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science,* 60(2), 434–448.

Bennett, R. E. (1993). On the meanings of constructed response. In R.E. Bennett & C. Ward (Eds.). *Construction versus choice in cognitive measurement.* (pp. 1–27). Hillsdale, NJ: Lawrence Erlbaum.

Bonner, S. M. (2013). Mathematical strategy use in solving test items in varied formats. *The Journal of Experimental Education.* 81(3), 409–428.

Calandra, J., & Hecht, B. (1971). Text of Calandra-Hecht bill amending sec. 2590g, Subdivision 12 of the education law. Retrieved from https://learning-curve.blogspot.com/2012/10/text-of-calandra-hecht-bill-amending.html.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement,* 5(2), 115–124.

Demars, C. E. (1998) Gender differences in math and science on a high school proficiency exam: the role of response format. *Applied Measurement in Education,* 11(3), 279–299.

Demars, C. E. (2000) Test stakes and item format interactions. *Applied Measurement in Education,* 13(1), 55–77.

Dempster, A. P. (1988) Employment discrimination and statistical science. *Statistical Science,* 3(2), 149–161.

Finn, C. E., Jr., & Hockett, J. A. (2012). *Exam schools: Inside America's most selective schools*. Princeton, NJ: Princeton University Press.

Gallagher, A. M., DeLisi, R., Holst, P. C., McGullicuddy-DeLisi, A. V., & Morely, M. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology,* 75, 165–190.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008), Differential validity and prediction of the SAT. College Board. Retrieved from http://professionals. collegeboard.com/profdownload/Differential_Validity_and_Prediction_of_the_SAT.pdf.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). Sex related performance differences on constructed-response and multiple choice sections of advanced place examinations. ETS RR No. 93-5.

Princeton Review. (2018), *Cracking the NYC SHSAT* (3 Ed.) New York, NY: Penguin Random House.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology,* 84(4),710–718.

Subotnik, R. F., & Strauss, S. M. (1995). Gender differences in classroom participation and achievement: An experiment involving advanced placement calculus classes. *Journal of Secondary Gifted Education,* 6(2), 77–85.

Tannenbaum, D.I. (2012). Do gender differences in risk aversion explain the gender gap in SAT Scores? Uncovering risk attitudes and the test score gap. Unpublished paper, University of Chicago, Chicago, IL.

Taylor, J. (2015). *Policy implications of a predictive validity study of the Specialized High School Admissions Test at three elite New York City high schools*. Unpublished Dissertation, Graduate Center, City University of New York, New York, NY.

Traub, R. E., & MacRury, K. A., Ontario Institute for Studies in Education. (1990). *Multiple-choice vs. free-response in the testing of scholastic achievement*. Toronto, Ontario, Canada: Ontario Institute for Studies in Education.

Wainer, H., & Steinberg, L.S. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review,* 62(3), 323–337.