Testimony in support of **Res. 196 -2018** calling upon the New York State Legislature to pass and the Governor to change the admissions criteria for New York City's Specialized High Schools.

Submitted bv Akil Bello
Co-Founder Bell Curves and Board Member PASSNYC
May 3, 2019

This testimony is being submitted by Akil Bello, Co-Founder of Bell Curves and board member PASSNYC. Bello has researched, developed, and taught test preparation courses for almost 30 years. In 2003, Bello founded Bell Curves as a social responsible test preparation company that focused on working with low income and underrepresented students. Bello left his role as CEO of Bell Curves in 2014. Bello is on the board of two non-profit organizations, PASSNYC and ESPI, that work to help students gain entry to the City's top schools.

Thank you for allowing me to testify today on the question of the admission process to specialized high schools and the SHSAT. My testimony will provide four compelling research-backed reasons that the use of the SHSAT should be discontinued.

I. **The use of the SHSAT as a sole measure of entry violates the recommendations and principles of the use of psychological testing as laid out by the majority of test writers and educational testing associations.** The standards of good practice established by all leading associations of assessment test developers (the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education) hold that "in educational settings, a decision or characterization that will have major impact on a student should take into consideration not just scores from a single test but other relevant information.[1]" The College Board, the owner of the SAT, has stated in the list of "score uses to be avoided" that the use of scores as "the sole indicator of overall performance of students, teachers, educational institutions, districts, states, or other groups.[2]" Other major admissions test owners and creators, including ACT, GMAC[3] (Graduate Management Admissions Committee which owns the GMAT), ETS[4] (creator of the GRE), and LSAC[5] (Law School Admissions Committee, which owns the LSAT) have all issued similar guidance. Pearson, the author of the SHSAT, states:

"***Pearson understands that concerns around the role of assessments are varied and real. We believe that quality assessments are useful to the learning experience, but they are just one measure of the knowledge and skills that learners need. They do not, and will never, completely define the sum total of what a good education ought to provide.***[6]***"***

Among K-20 educational institutions in the United States, New York's specialized schools stand alone in their policy to determine admission entirely based on the results of an admission test. Not only is NY singular in its decision to use a test in this way, it is doing so in direct contravention of scientific guidance provided by the very authors of the exam.

II. **The variability of the content, structure, and logistics of the SHSAT since 1991, in and of itself, suggests that there is nothing inherently special about this exam that contributes to the identification of students with the ability to succeed in specialized schools.** An analysis of the the Specialized High Schools Handbooks released since 1991 shows at least one substantial change occurring, on average, every 3 years. The DOE has released only one validity study[7] in the past 30 years to substantiate the utility and predictive validity of the exam. However, even if other validity reports were available they would not support the use of the current form of the exam since they would have been conducted on an entirely different exam. The frequency of change to the SHSAT is also concerning because whenever a standardized test is changed, responsible psychometricians base those changes on sound theory and substantial research. Changes to large-scaled admissions tests are therefore typically infrequent and based on historic data used to evaluate how well the assessment accurately, fairly, objectively, and reliably assesses the standards it is designed to assess. While longitudinal comparisons aren't strictly necessary for the SHSAT to sort students year to year, it helps the research basis of the test to have a comparable test to allow better comparisons year over year. The frequency of the changes, and even the relative stability from 2007 to 2016, raise questions about what logic guides the changes or non-changes year to year. The fact that the test changes so frequently with no impact on the quality of graduates from the specialized high schools also argues against the utility of the exam as a necessary factor in that success.

Below is a history of some of the changes to the SHSAT[8]:

Specialized School Admission Test Partial History of Changes	
1969	Test was called the Science High Schools Admissions Test (SHAT) and prepared annually by Institute of Psychological Research, Columbia University.
1971	Stuy, BxSci, and Bklyn Tech all lobbied aggressively to get the Hecht-Calandra Law, which codified the entrance exam into state (not city) law.
	---- records currently unavailable ----
1991	Test is renamed Specialized Science High Schools Admission Test (SSHSAT) is a one-and-a-half-hour multiple-choice test, which includes verbal (sentence completions, synonyms, reading comprehension, logical reasoning) and mathematics (problem solving and quantitative comparison).
1993	Test is a two hour and ten minute multiple-choice test, which includes 55 verbal (sentence completions, synonyms, reading comprehension, logical reasoning) questions and 50 mathematics (problem solving and quantitative comparison) questions.
	---- records currently unavailable ----
1995	Synonyms renamed “word meanings.” Sentence completions removed, scrambled paragraphs added. Verbal now has 5 fewer questions but the same amount of time. Math is unchanged. Test time increased to two hours and thirty minutes. Instructions given that in verbal test-takers “should not spend more than 80 minutes” and in math “70 minutes allotted. You may go back to either section.
	---- records currently unavailable ----
1998	Scrambled Paragraphs reformatted. Math formulas no longer given. Quantitative comparison questions removed from Math. Word meanings are removed from Verbal. Reading Comprehension increased from 5 passages with 5 questions each to 5 passages with 6 questions each. Test timing remains unchanged.
1999	Verbal reduced to 45 questions. Scramble Paragraphs are reduced from 8 to 5 and awarded 2 points each rather than 1. Scrambled Paragraphs reformatted again. Logical reasoning reduced from 12 to 10 questions.
2002	Test administration moved from December to October to reduce test preparation. Test is renamed from Specialized Science High Schools Admission Test to Specialized High School Admissions Test (since the founding of non-science specialized schools.

2005	American Guidance Service (author of SHSAT) acquired by Pearson.
2007	Scrambled paragraphs reformatted.
2018	Verbal renamed ELA. Scrambled Paragraphs removed. Logical reasoning removed. 20 Revising editing added (6 stand alone, 14 passage based). Reading comprehension increased to 37 questions from 6 passages. Number of scored verbal questions changed from 45 to 48. Number of scored math questions changed to 48. 20 unscored questions added. Math standards revised. 5 Grid in questions added to math. Test time increased to 180 minutes.
2019	Fiction passages (including poems) added to ELA. Revising/editing questions reduced from 20 to 9 (3 stand alone, 6 passage based). Reading comprehension passages lengthened and reformatted. Charts included with some passages. Reading comprehension increased from 37 to 48 questions. Total time remains 180 minutes.

Not only do other high stakes tests take longer before making changes to items or specifications of the test, but the development period for items (questions) appears to be longer. For the SAT, it's about 2.5 to 3 years between when an item is first written to the point at which a test-taker first sees the item. According to the College Board, the writing revision process is so rigorous that on average only about 50% of written items are actually used (this can be for many reasons, among them an item may be determined to advantage or disadvantage a group of students). Prior to the most recent SAT redesign (the first in about a decade), the College Board released publicly a 250+ page specifications document[9], a full two years before the first redesigned SAT was administered to test-takers. The most recent changes to the SHSAT were announced in the fall of 2016, a vendor was contracted in December of that year, and the test was administered in October of 2017. ACT also reports a similar timetable to develop items[10].

One example of substantial changes to the SHSAT can be seen in the Flesch-Kincaid Reading Level analysis I compiled below using released sample passages from the released Handbook exams from the 2013 - 2014 Handbook as compared to those in the 17 - 18 and 18 - 19 version.

	Average Word Count	Flesch Kincaid Grade level
SHSAT 13-14	437.8	11.3
SHSAT 17 - 18	479.7	11.2
SHSAT 18 - 19 (w/ poems)	626.9	9.1
SHSAT 18 - 19 (w/o poems)	706.1	10.2
8th Grade ELA 2017	854.6	7.6
NYT Opinion Article	1301	11.3

This analysis demonstrates one way in which the content of the exam is changed periodically for unexplained reasons and with undefined impact on students. Further variability and unreliability of this exam can be seen in the annual adjustment of items in the handbooks, indicating that Pearson is at very least reconsidering and reworking questions included in the Handbooks after delivering them to the public for use. Without technical reports or validity reports for the SHSAT there is no way to know whether this practice of using poorly developed items is carried over to the actual test.

In the 2017-2018 Handbook this question appears:

2. Read this paragraph.

> (1) When coal was used to heat homes, it frequently left a soot stain on the walls. (2) Brothers Cleo and Noah McVicker, who owned a cleaning product company, created a doughy substance to help people remove this soot. (3) Over time, as natural gas becomes more common, people had little need for soot cleansers, and the McVickers' family company struggled to stay in business. (4) Then one day, Joe McVicker, Cleo's son, learned that his sister-in-law had been using the substance for art projects in her classroom, so he remarketed the product as the toy known today as Play-Doh.

Which sentence should be revised to correct an inappropriate shift in verb tense?

E. sentence 1
F. sentence 2
G. sentence 3
H. sentence 4

However in the 2018 - 2019 it appeared again this time more clearly written to prevent possible confusion in the interpretation of the question. Here is the revised question:

2. Read this paragraph.

> (1) When coal was used to heat homes, it frequently left soot stains on the walls. (2) Brothers Cleo and Noah McVicker, who owned a cleaning product company created a doughy substance to help people remove this soot. (3) Over time, as natural gas becomes more common, people had little need for soot cleansers, and the McVickers' family company struggled to stay in business. (4) Then one day Joe McVicker, Cleo's son, learned that his sister-in-law had been using the substance for art projects in her classroom, so he remarketed the product as the toy known today as Play-Doh.

Which pair of revisions need to be made in the paragraph?

E. Sentence 1: Delete the comma after ***homes***.
Sentence 3: Change ***becomes*** to **became**.

F. Sentence 1: Delete the comma after ***homes***.
Sentence 4: Change ***remarketed*** to **had remarketed**.

G. Sentence 2: Insert a comma after ***company***.
Sentence 3: Change ***becomes*** to **became**.

H. Sentence 2: Insert a comma after ***company***.
Sentence 4: Change ***remarketed*** to **had remarketed**.

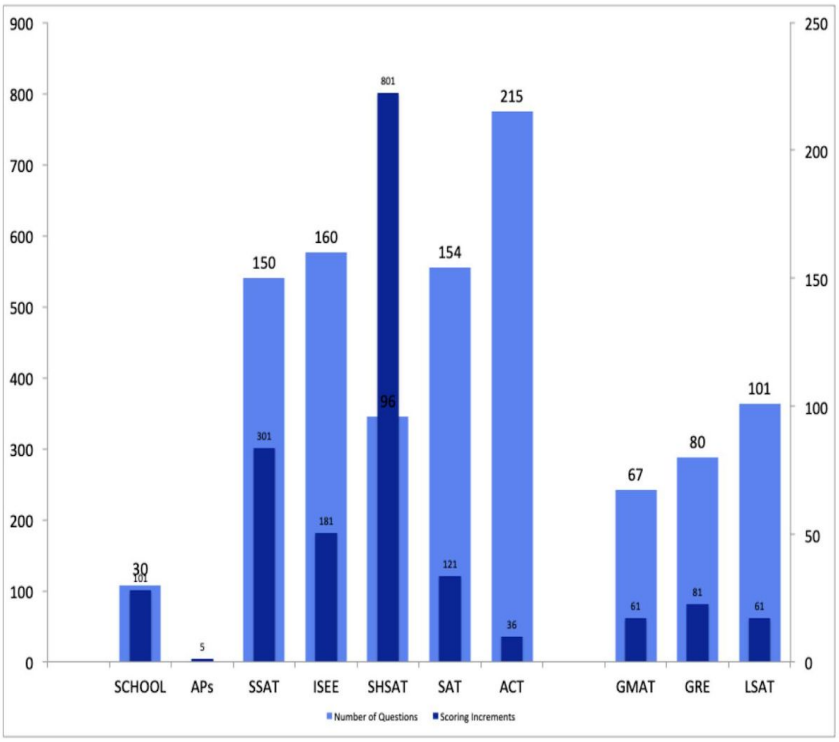
Multiple examples of this nature exist and can be provided.

The variability of the exam content and the constant editing of questions strongly indicates that the SHSAT is not primarily identifying those of the greatest academic ability and is hardly an objective measure of ability. It is instead, a sorting tool that assesses some academic skills but also a plethora of other skills that pertain performance under pressure, speededness, and the ability to overlook "noise" in the test. The SHSAT might be more than anything else identifying those individuals who are comfortable with this particular exam and all its particularities rather than identifying the most prepared students for high school.

III. **The number of confusing psychometric elements of SHSAT test design and scoring raise questions about the reliability and validity of the exam.** When the SHSAT was altered most recently the number of scored questions was reduced from 95 to 94 and the raw score value of those questions was reduced from 100 to 94, however the scaled scoring remained, from all available evidence, unchanged. This is not the first

time the test has altered its scoring algorithms without providing research, justification, or evidence of impact. Lacking evidence of impact of these odd changes it is important to consider the psychometrics of this exam in context of similar tests.

In addition, the SHSAT reports scores in 1 point increments on a scale of 0 to 400 in math and the same in ELA. This means that the SHSAT purports to show 801 levels of distinction between test-takers. No other major standardized test, which all have greater research pool of annual test-takers, claims that level of accuracy. This scoring suggests that either the SHSAT is a better, more precise measure of performance than other standardized tests and is able to achieve that level of precision with, in many cases, fewer questions or that the SHSAT is built on junk science.



The SHSAT is scored on just 94 questions, while the ACT uses 215 questions to provide 36 levels of difference and the SAT needs 154 questions to provide 121 levels of score distinction, the ISEE used for Boston Latin has 160 questions and 181 increments of distinction. It's generally accepted practice that the number of scaled scores possible to earn (as can be seen in the chart below there are a large number of scores that could not be earned by any test-taker) be fairly close to the number of theoretical increments in the scale. If there are less questions than increments in the scoring scale, score gaps will occur. The bigger this discrepancy the more score gaps appear. While some score

gaps are tolerable, a large number of them suggests a poorly designed scale or pattern scoring. Without further information it's impossible to conclude whether the SHSAT scale is appropriately designed.

What is clear is that by using the scale that it does, the SHSAT is making a claim to a greater level of precision than any other comparable assessment. This precision is questionable at best, and a score scale of this nature likely projects differences in ability between two test-takers that are not in actually representative of the test-taker's true ability.

Estimated Verbal Raw to Scaled Partial Conversion Chart	
Raw Score (number of questions correct)	**Scaled Score (0 - 400)**
35	240
36	245
37	248
38	253
39	257
40	262
41	267
42	271
43	277
44	284
45	291
46	299
47	310
48	324
49	345
50	365

A third conundrum in the scoring of the SHSAT is the quirk in the conversion from raw score (the total number of questions correct) to scaled score (0 to 800) which was brought under scrutiny by a 2005 New York Times article[15] (also captured in a 2008 research paper[15]) that revealed that the test strangely rewards test-takers with imbalanced scores over test-takers with the same number of correct answers but whose performance was balanced in both math and verbal (suggesting a type of pattern scoring). This scoring methodology, which by all reports is still part of the test, would award a higher scaled score to a test-taker who scored 40 on verbal and 47 on math (for a total of 87 raw

points) than a test-taker who scored a 44 on each section (earning a total of 88 raw points).

These scoring quirks suggest that the SHSAT uses some combination of "number correct scoring" theory and "pattern scoring," which would be an oddity among standardized tests. This once again reinforces that the SHSAT is an oddity of a standardized test with quirks and nuances that do nothing to advance good testing practice, identify the brightest students, or support the mission of New York City Department of Education.

IV. **The continued use of a test without strong scientific research to support that it is clearly a superior means of identifying academically prepared students than GPA or other measures increases cost to the city and inequality of educational outcomes by creating demand for test prep.** The use of the SHSAT and the accompanying need for test preparation should also be evaluated in terms of the economic impact on the city. It is at odds with the values of a public education to create a test that stands outside of normal school curriculum and the normal public school selection process, requires specialized preparation, and is a financial drain to the government and taxpayer. Given that the administration of the SHSAT comes with significant cost, the limited predictive validity as compared to the immense cost fails, as a taxpayer, fails any cost-benefit analysis. In 2016, the DOE renewed Pearson's contract to create the exam for an annual cost of $2.23 million[16]. The administration of the test (proctors, security, and administrative staff), the DREAM-SHSI program, the Discovery program, specialized high schools informational workshops, the salaries of at least two Office of Assessment staff responsible for liaising with Pearson as well as many other programs and services all exists in the DOE budget simply to serve the delivery of a quasi-scientific test of limited validity or reliability with highly inequitable participation and outcomes.

Test prep will never be able to level the gaps in opportunity and outcomes engendered by the SHSAT since the duration, effectiveness, and cost of test preparation greatly vary not only by type of preparation but also by community. While no studies have been conducted on the effectiveness of SHSAT preparation, there are studies on the effectiveness of SAT[19] and ACT[20] test preparation that can be used to draw comparisons. These studies, which do have some limitations that make seem to make

their findings underpredict the impact of test prep, have consistently shown that one-on-one tutoring is the most effective form of test preparation. One-on-one tutoring is also the most expensive form of commercial test preparation typically. For test preparation to be effective in leveling the playing field test preparation, all test preparation methods, would have to be equally available to all communities and it is not. Highly experienced test preparation experts can cost as much as $450 per hour[21] for one-on-one tutoring and allow wealthy families and communities the benefit of years of research and experience[22] while less affluent or connected communities do not benefit from the same high quality preparation.

Many recent immigrants from Asian countries, especially China[25] and South Korea[26] where a national single admission test is common and preparation for it lasts years, benefit from a familiarity with and expectation of test preparation that makes it more familiar and expected to spend months or years focused on one exam. This focus often helps make up for disparity in quality or experience of prep programs. The dearth of test preparation programs in Hispanic and African American communities combined with the costs of those programs that are available create a situation that dissuades all but the most stalwart families from pursuing test preparation. Further, the test preparation programs implemented are restricted to using DOE certified instructors which creates a significantly different pedagogy, expertise, and experience than that of private test prep vendors and this would likely be revealed in lower outcomes than those of private programs[17, 18], though that is hard to verify since the results from the city's programs are not made public.

The need for specialized test preparation to excel on the SHSAT enhances the point that this test is not testing for the skills that are taught in the public school curriculum but instead testing for skills particular to this exam and potentially unnecessary for success in a classroom. No research has shown that the type of study that the SHSAT engenders is of greater benefit than attending a school with a rigorous curriculum and effective teachers. Rather than investing additional millions in test preparation the state would be wise to eliminate the test and reinvest that money in schools.

There is a preponderance of evidence that establishes that the SHSAT not only does not meet the research standards to serve as sole means of admissions to NYC's specialized schools but also its

development may not be driven by evidence and ongoing research. Another way to put this is that there is no proof that the SHSAT isn't simply a bad test on which some bright students are doing better than others. Further, if the SHSAT were replaced with another test the results would be largely the same as the SAT[23], PSAT[23], and ACT[24] all demonstrate similar economic advantages for wealthy communities. Anyone interested in a fair educational system and good educational practice should question the use of this test at all. Anyone who believes in the integrity of the public school system should be against sending the message that the grades given by teachers, the results of statewide assessments, and the strength of schools curriculum are all less valuable than *any* standardized test that research shows adds almost nothing to the prediction[11, 12, 13] of success. Continuing the use of the SHSAT and encouraging test prep means that the DOE believes that its mandate is to encourage test prep rather than to encourage focus on learning and academics. The use of GPA and other measures encourages students to focus on improving their performance in the classroom to gain entry to the "best" high schools. This a laudable and educationally sound practice for the DOE to encourage through process and policy. The use of the SHSAT alone encourages students to ignore the classroom and focus on this measure that stands outside of the classroom. As a citizen of New York, I argue that we want our policies to reflect our morals and values and encouraging endless test prep is not a value that I would argue anyone but test writers have.

Lacking clear indications of the benefit of an exam *over* using other measures and the costs to produce and deliver the exam, I contend that the sole use of any admissions exam is a tool of political expediency (and vestige of historical racism) rather than a scientifically supported academic objective and fair measure of achievement. The single test entry system should be ended as bad educational, psychometric, and social policy.

References

1. 2014, American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.), Standards for Educational and Psychological Testing, Standard 12.4
2. 2018, College Board, Guidelines on the Uses of College Board Test Scores and Related Data, page 13
3. Retrieved May 2nd 2019, gmac.com, How to Use GMAT Scores to make business school decisions
4. Retrieved May 2nd 2019, ets.org, *GRE® Guidelines for the Use of Scores*
5. Retrieved May 2nd 2019, Law School Admissions Council, Cautionary Policies Concerning LSAT® Scores and Related Services
6. Retrieved May 2nd 2019, pearson.com, Our Position On Assessment
7. 2013, Metis Associates, The Specialized High School Admission Test and High School Academic Achievement
8. 2019, Bell Curves, A History of the SHSAT (unpublished internal company document), page 1
9. 2104, College Board, Test Specifications for the Redesigned SAT
10. 2018, ACT, ACT Technical Manual, chapter 2
11. 2018, Jonathan Taylor, http://www.gothamgazette.com/opinion/7871-new-research-shows-shsat-less-valuable-predictor-than-middle-school-grades
12. 2017, Jonathan Taylor, The Predictive Validity of the Specialized High School Admissions Test at Three New York City High Schools, page 6
13. 2015, Jonathan Taylor, Policy Implications of a Predictive Validity Study of the Specialized High School Admissions Test at Three Elite New York City High Schools,
14. 2005, David Herszenhorn NYT, https://www.nytimes.com/2005/11/12/nyregion/admission-tests-scoring-quirk-throws-balance-into-question.html?mtrref=www.google.com&gwh=9DC944D2E741414C6AF3173F5883102E&gwt=pay
15. 2008, Joshua Feinman, Ph.D., High Stakes, but Low Validity? A Case Study of Standardized Tests and Admissions into New York City Specialized High Schools, page 19
16. 2016, Christina Veiga Chalkbeat https://www.chalkbeat.org/posts/ny/2016/09/19/nyc-department-of-education-recommends-that-pearson-continue-to-provide-specialized-high-school-admissions-test/
17. Retrieved May 4 2019, Khan's Tutorial, https://www.khanstutorial.com/pressrelease/2019/5/4/khans-tutorial-annual-shsat-awards-ceremony-celebrates-385-shsat-students-who-received-offers-to-nycs-specialized-high-schools

18. https://www.wnyc.org/story/why-test-prep-may-be-key-improving-diversity-citys-specialized-high-schools/
19. 2009, Derek C. Briggs, Preparation for College Admission Exams. 2009 NACAC Discussion Paper, page 12, 15
20. 2018, Raeal Moore, Edgar Sanchez, Maria Ofelia San Pedro, Investigating Test PrepImpacton Score Gains UsingQuasi-Experimental Propensity Score Matching, page 3, 15
21. Retrieved May 4, 2019, https://www.noodlepros.com/tutor/profile/OTk2YmY3MmM3MzNmMjE
22. Retrieved May 4, 2019, Jonathan Arak, Cracking the New York City Specialized Science High School Admission Test
23. 2016, College Board, 2016 Total Group SAT Suite of Assessments Annual Report, page 4
24. Retrieved May 4 2019, ACT Enrollment Management Database
25. 2009, Sharon LaFraniere New York TImes, China's College Entry Test Is an Obsession
26. 2016, Anna Diamond the Atlantic, South Korea's Testing Fixation
27. https://hechingerreport.org/the-problem-with-high-stakes-testing-and-women-in-stem/