USF HORIZONTES

ENSAIO

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

New tests, same old problems: how high-stakes standardized tests have always reproduced inequity

Wayne Au[1]
https://orcid.org/0000-0002-5533-7394

Here in the United States, there has been a movement to make high-stakes, standardized tests like the Scholastic Assessment Test (SAT) or the American College Testing (ACT) optional for admission into many universities. Over 1900 U.S. universities, including major systems like the California State University system and some elite Ivy League universities, currently do not require student applicants to provide an SAT or ACT test score, let alone use a particular test score to determine admission. However, recently a few elite U.S. universities, like Dartmouth College, have changed course and have once again started requiring potential students to submit high-stakes exams scores like the SAT and ACT in order to be admitted (Nadworny; Ahmad, 2024). While it is not clear just how many other universities will reinstate such tests, the struggles around whether or not to use such testing for college admissions still revolve around the same old issues of what all of our high-stakes, standardized tests measure and whether or not they are a tool of equity or inequity.

Looked at historically, the use of standardized testing as a tool of mental measurement and learning has always been steeped in inequality. The origins of mass standardized testing, particularly in the U.S., but also for much of the West, begins with intelligence testing: the idea that we can measure human intelligence through a series of tests. In the context of the United States, psychologists in the early 1900s like Goddard, Terman, Brigham, and Yerkes perverted Binet's original testing for developmental issues in young children in France into much more crass and flawed tests that could supposedly measure human intelligence. In their turn, these psychologists developed intelligence tests that believed were objective, but that in reality were wildly biased products of their own flawed beliefs (Au, 2023b).

For instance, in 1917, using a massive pool of 1.75 million World War I army recruits who they tested for "mental fitness," Goddard, Terman, and Yerkes arrived at the conclusion that

[1] University of Washington. Bothel, WA: wayneau@uw.edu.

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

lighter-skinned immigrants from Northern Europe were more intelligent than darker-skinned immigrants from Eastern and Southern Europe, that the rich were more intelligent than the poor, and that Black Americans were the least intelligent of all peoples. These findings aligned with the racist, classist, and sexist eugenics movement of the time – a movement that not only advocated for the debunked "science" of attributing behaviors, morals, and intelligence to genetic makeup, but also believed in notions of racial purity, Western superiority, and, ultimately, white supremacy (Au, 2023b).

In turn, psychologists like Terman developed similar tests for school children, which, when combined with desires to efficiently sort students within the growing system of mass public schooling in the U.S. at the time, were then implemented in school districts in major cities across the country. As we might expect of racist and classist standardized educational measurements, these tests were then used to label Black and Brown children as less intelligent, fueling eugenicist arguments about the kind of schooling those children deserved (Au, 2023b). Indeed, Terman (1916) himself remarked:

> The fact that one meets this type with such frequency among Indians, Mexicans, and negroes suggest quite forcibly that the whole question of racial differences in mental traits will have to be taken up anew and by experimental methods…. Children of this group should be segregated in special classes and be given instruction which is concrete and practical. They cannot master, but they can often be made efficient workers, able to look out for themselves. There is no possibility at present of convincing society that they should not be allowed to reproduce, although from a eugenic standpoint of view they constitute a grave problem because of their unusually prolific breeding (Terman, 1916, p. 91-92).

Using these racist I.Q. tests developed by Terman and others, Mexican students in the U.S. West and Southwest were then placed in specific educational tracks with inferior offerings (Blanton, 2003). Through this history we can see that at their conceptual heart, our use of high-stakes, standardized tests to supposedly measure learning within systems of mass schooling have always produced inequitable educational outcomes, particularly relative to race and class (Au, 2023b).

The Irony of the SAT

The college entrance exam, the SAT, also came out of this political and historical milieu, and it highlights one of the ongoing disputes with the use of high-stakes, standardized testing

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

more generally. The Scholastic Aptitude Test was first administered in 1926. In the 1990s, in an attempt to move away from the concept of "aptitude," it was then renamed the Scholastic Assessment Test. Then by the 2000s, in another marketing shift, it was rebranded as just the SAT – with no meaning attached to the acronym. The SAT was originally developed by Brigham, who adapted them from Yerkes' original intelligence testing of World War I Army recruits. Brigham was also a eugenicist who believed in the biological basis of intelligence, as well as the superior intelligence of Europeans (Rosner, 2012; Viera, 2018).

Perhaps ironically, given the how much Brigham and others were mired in racist, classist, and sexist eugenics, the development of the SAT was driven by an egalitarian impulse. Up until the early 1900s, only members of white, elite, wealthy families could gain admission into the predominantly white universities, as such admissions were based on whether or not your father or grandfather attended university as well. These heritage admissions policies meant that only a highly select group were allowed to attend university. Brigham developed the SAT with the explicit idea of challenging such privileges driving university admissions. From his vantage at the time (a vantage that is shared by many test-proponents now), he saw the SAT as a chance to measure individuals fairly and based on individual merit, not on the status of their families. Ideally for Brigham, then, the SAT would produce more educational equality, not less (Au, 2023b).

However, the SAT has never escaped its racist, classist, eugenic origins. Here we sit, over 100 years removed from when the SAT was first administered as a college entrance exam, and the overwhelming evidence points to how the SAT has continuously reproduced the race, class, and educational inequalities of the students who take the test. Indeed, SAT scores correlate so strongly with the combination of the economic class of a student's family and the educational level of a student's parents or grandparents that we could take a room full of SAT-takers and accurately predict the overall distribution of scores without any of them taking the test at all (Au, 2023b). Research has pointed to multiple reasons for the SAT's deep seated inequality, including how the selection of test questions create a self-reinforcing racial bias (Kidder; Rosner, 2002) and racial bias in early test questions (Santelices; Wilson, 2010).

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

The Conceit of the Individual Superseding the Rule

All of this points to one of the central conceits of all high-stakes, standardize testing, SAT or otherwise. The overall trend for all of our high-stakes testing in the U.S. is remarkably consistent with the blatantly racist and classist I.Q. testing that birthed educational psychometrics 100 years ago: Working class Black and Brown children score lower on these tests than affluent white children. Ultimately, to make sense of this overwhelming trend, either you prescribe to white supremacist notion that these children and communities are intellectually inferior, or you prescribe to the idea that there are underlying, systemic issues (either in the tests or in society or both) that are producing these outcomes. Yes, there are *individuals* from working class backgrounds whose parents do not have college degrees and who are from Black and Brown communities who do quite well on standardized tests. However, by definition, these students are the exception, not the rule (Au, 2023b).

And this is the conceit I mentioned, above. Much like those 100 years ago or right now who romantically yearn for the SAT to challenge university elitism – despite all evidence to the contrary, advocates for the use of high-stakes testing typically point to the exceptions and suggest we can make them the rule. The problem with this conceit is at least two-fold. First, it suggests that individuals can overcome systemic outcomes to such a degree that individual outliers (those from working class backgrounds that score highly) can become the statistical norm. This, of course, is absolute nonsense. Data produced by standardized tests, whether norm-referenced or criterion-referenced, are typically interpreted along a statistical bell curve of "normal distribution." This means that no matter what, the tests are made – and derive their validity from – a central assumption that there is a "thing" called "intelligence" and that this thing is distributed unevenly across human populations. Here we see not only the presumptive legacy of eugenics sneaking into the foundations of psychometric assumptions, we also see that the tests themselves are designed to sort human populations into those who pass and those who fail. Under regimes of high-stakes, standardized testing, then, we can't all be "winners" and there have to be "losers" – otherwise everyone questions the validity of the tests themselves. Put more simply, if everyone passes, either the test is deemed too easy or we assume people cheated (Au, 2023b).

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

The second problem with this conceit of testing for individual merit is that it belies the reality of systemic oppression. Test data is test data, and we have decades upon decades of data showing that our standardized assessments correlate more strongly with income, race, and family education than anything else (Berliner, 2013). However, by pushing the idea that standardized tests are a true measure of individual hard work and merit (and, by extension, a measure of teaching and learning), testing advocates refuse to recognize that the systemic issues of stable housing, access to medical care, access to food, access to educational resources inside and outside of schools, and livable wages for student care providers are necessities for student success and learning. Instead, the ideology of meritocracy embedded in the tests suggest that the issue is not whether or not students' needs are being met by society, but rather that the issue is that students are just not working hard enough (Au, 2023b).

Testing (In)Validity

Critics, like myself, point to the systemic outcomes and challenge both the validity of measurement and the applied use of these tests (Amrein-Beardsley, 2014; Au, 2023b; Baker, B., 2013; Holloway-Libell; Amrein-Beardsley, 2015), and with good reason. For instance, there are a lot of statistical issues with relying on standardized tests to make high-stakes decisions about teachers and students. Studies have found huge error rates in using tests to evaluate teachers (Baker, B., 2010; Schochet; Chiang, 2010), instability in test scores year-to-year (Sass, 2008), instability in test scores due to random events students experience on the day of a test (Kane; Staiger, 2002), problems with non-random student assignment skewing test scores (Baker, E. *et al*., 2010), and manipulation of test scores within testing companies hired to provide the assessment (DiMaggio, 2010; Farley, 2009, 2010).

Not only are these tests problematic for making high-stakes decisions about students teachers due to their statistical imprecision, it is also clear that we don't really know that they are measuring learning and teaching. What many people don't understand about high-stakes, standardized testing as a tool of educational assessment, is that they are based purely on correlation. That is to say, our standardized tests can only assess a sample of student answers (answers which are already greatly influenced by factors such as emotional state, amount of

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

sleep, hunger, health, etc.), and then we assume a correlation between that sample and whether or not a student knows a larger body of knowledge. None of it is direct, causal knowledge of student learning. It is all correlation (Au, 2023b). Importantly, there are a whole host of non-knowledge-based correlations for high-stakes, standardized testing. For instance, there are studies that show a correlation between the amount of natural, plant greenness around schools and higher test scores (Kuo *et al*., 2018; Wu *et al*., 2014). There is also research pointing to correlations between test scores and student cardiorespiratory fitness (Garber *et al*., 2018), classroom temperature (Chang; Kajackaite, 2019; Goodman *et al*., 2018), stress and cortisol levels (Adam *et al*., 2017; Heissel *et al*., 2017, 2021), and overall cognitive fatigue (Sievertsen *et al*., 2016).

Based on these correlations, we could raise test scores simply by making sure kids are well rested, relaxed, in good respiratory shape, in a cool room, surrounded by forest, and testing at the beginning of the day before their minds get fatigued. Notice that none of these have anything to do with what student has learned or what a teacher has taught. All of which raises the important issue of what these tests are measuring if they are not really measuring learning. Elsewhere (Au, 2023a), I have argued that while high-stakes, standardized tests are terrible measures of learning and teaching, they are excellent measures of the amount of social labor and resources have been put into children's and their community's lives. In my view, the social construction of knowledge, language, and classroom discourse all combine transmit inequitable social, economic, and cultural relations into all levels of teaching and learning (Au, 2008), and that these relations are a function of inequitable distribution of resources (Au, 2023a).

New Tests, Same Old Problems

In the end, contemporary fights around the use of high-stakes, standardized tests to measure teaching and learning raise the same issues with educational assessment that existed over 100 years ago at their origin. We mistakenly presume our tests are objective, thereby obscuring deep-seated biases in their construction and outcomes. Similarly, we also mistakenly assume that our tests provide a pathway for equal, individual success in education, thereby obscuring the overwhelming role that systemic inequities play in educational outcomes. We also

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

build educational policy around the mistaken belief that our tests are providing valid measures of teaching and learning, thereby obscuring the reality that, perhaps they are measuring something else entirely. In the end, despite whatever technological advancements psychometricians have made over the last 100 years, all of our new tests somehow have the same old problems.

References

ADAM, E. K., HEISSEL, J. A., HITTNER, E. F., DOLEAC, J. L., MEER, J.; FIGLIO, D. Adolescent cortisol responses to high-stakes testing in school-based settings. *Psychoneuroendocrinology*, v. *83*, n. 85. 2017. DOI: https://doi.org/10.1016/j.psyneuen.2017.07.465

AMREIN-BEARDSLEY, A. *Rethinking value-added models in education: Critical Perspectives on tests and assessment-based accountability*. New York: Routledge, 2014.

AU, W. Devising inequality: A Bernsteinian analysis of high-stakes testing and social reproduction in education. *British Journal of Sociology of Education*, v. 29, n. 6, p. 639–651, 2008. DOI: https://doi.org/10.1080/01425690802423312

AU, W. Commodification, the violence of abstraction, and socially necessary labor time: A Marxist analysis of high-stakes testing and capitalist education in the United States, 2023a. *In*: HALL, R.; ACCIOLY, I.; SZADKOWSKI, K. (Eds.), *The Palgrave international handbook of Marxism and education*. p. 223–242. PalgraveMacmillan, 2023a. DOI: https://link.springer.com/chapter/10.1007/978-3-031-37252-0_12

AU, W. *Unequal by design: High-stakes testing and the standardization of inequality* (2. ed.). New York: Routledge, 2023b.

BAKER, B. D. *School finance 101: Rolling dice: If I roll a "6" you're fired!* 2010. Available in: https://schoolfinance101.com/2010/07/28/rolldice/

BAKER, B. D. The value added & growth score train wreck is here. *School Finance 101*. 2013. Available in: https://schoolfinance101.com/2013/10/16/the-value-added-growth-score-train-wreck-is-here/

BAKER, E. L., BARTON, P. E., DARLING-HAMMOND, L., HAERTEL, E., LADD, H. F., LINN, R. L., RAVITCH, D., ROTHSTEIN, R., SHAVELSON, R. J.; SHEPARD, L. A. *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute, 2010.

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

BERLINER, D. C. Effects of inequality and poverty vs. Teachers and schooling on America's youth. *Teachers College Record*, v. *115, n.* 12. 2013. DOI: https://doi.org/10.1177/016146811311501203

BLANTON, C. K. From intellectual deficiency to cultural deficiency: Mexican Americans, testing, and public school policy in the American Southwest, 1920-1940. *Pacific Historical Review*, v.*72*(1), 39–62, 2003. DOI: https://doi.org/10.1525/phr.2003.72.1.39

CHANG, T. Y.; KAJACKAITE, A. Battle for the thermostat: Gender and the effect of temperature on cognitive performance. *PLoS ONE*, *14*(5), 1–10. 2019. DOI: https://doi.org/10.1371/journal.Pone.0216362

DIMAGGIO, D. The loneliness of the long-distance test scorer. *Monthly Review*, v. 62, n. 7. 2010. Available in: http://monthlyreview.org/2010/12/01/the-loneliness-of-the-long-distance-test-scorer

FARLEY, T. *Making the grades: My misadventures in the standardized testing industry*. Berrett-Koehler Publishers, 2009.

FARLEY, T. A test scorer's lament. *Rethinking Schools*, v. 23, n. 2. 2010. Available in: http://www.rethinkingschools.org/archive/23_02/test232.shtml.

GARBER, M. D., STANHOPE, K. K., CHEUNG, P., & GAZMARARIAN, J. A. Effect of cardiorespiratory fitness on academic achievement is stronger in High-SES elementary schools compared to low. *Journal of School Health*, *88*(10), 707–716, 2018.

GOODMAN, J., HURWITZ, M., PARK, J.; SMITH, J. *Heat and learning* (Working Paper 24639). National Bureau of Economic Research. 2018. http://www.nber.org/papers/w24639.

HEISSEL, J. A., ADAM, E. K., DOLEAC, J. L., FIGLIO, D.; MEER, J. Testing, stress, and performance: How students respond physiologically to high-stakes testing. *Education Finance and Policy*, v. *16, n.* 2, p. 183–208. 2021. DOI: https://doi.org/10.1162/edfp_a_00306.

HEISSEL, J. A., LEVY, D. J.: ADAM, E. K. Stress, sleep, and performance on standardized tests: Understudied pathways to the achievement gap. *AERA Open*, v. 3, n. 3, p. 1–17. 2017. DOI: https://doi.org/10.1177/2332858417713488

HOLLOWAY-LIBELL, J.; AMREIN-BEARDSLEY, A. "Truths" devoid of empirical proof: Underlying assumptions surrounding value-added models in teacher evaluation. *Teachers College Record Commentary*. 2015. Available in: https://www.tcrecord.org/Content.asp?ContentId=18008.

KANE, T.J., & STAIGER, D. Volatility in School Test Scores: Implications for Test-Based Accountability Systems. *Brookings Papers on Education Policy* 2002, 235-283. DOI: http://dx.doi.org/10.1353/pep.2002.0010

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

KIDDER, W. C.; ROSNER, J. How the SAT creates "built-in headwinds": An educational and legal analysis of disparate impact. *Santa Clara Law Review*, v. 43, p. 131–212, 2002.

KUO, M., BROWNING, M. H. E. M., SACHDEVA, S., LEE, K.; WESTPHAL, L. Might school performance grown on trees? Examining the link between "greenness" and academic achievement in urban, high-poverty schools. *Frontiers in Psychology*, v. *9 n.* 1669, p. 1–14, 2018.

NADWORNY, E.; AHMAD, H. *Dartmouth will again require SAT, ACT scores. Other colleges won't necessarily follow*. National Public Radio: All Things Considered. 2024. Available in: https://www.npr.org/2024/02/05/1229223433/sat-act-diversity-dartmouth-college-admissions

ROSNER, J. The SAT: Quantifying the unfairness behind the bubbles. *In*: SOARES, J. A. (Ed.), *SAT wars*. New York: Teachers College Press, 2012.

SANTELICES, M. V.; WILSON, M. Unfair treatment?: The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review*, v. 80, n. 1, p. 106–133, 2010.

SASS, T. R. *The stability of value-added measures of teacher quality and implication for teacher compensation* [Policy Brief]. National Center for Analysis of Longitudinal Data in Educational Research, 2008.

SCHOCHET, P. Z.; CHIANG, H. S. *Error rates in measuring teacher and school performance based on test score gains* (NCEE 2010-4004; p. 59). U.S. Department of Education, Institute of Educational Sciences, National Center for Educational Evaluation and Regional Assistance. 2010. Available in: http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf.

SIEVERTSEN, H. H., GINO, F.; PIOVESAN, M. Cognitive fatigue influences students performance on standardized tests. *PNAS*, v. *113,* n. 10, 2621–2624. 2016. DOI: https://doi.org/10.1073/pnas.1516947113

TERMAN, L. The measurement of intelligence, 1916. In W. Dennis (Ed.), Readings in the history of psychology (pp. 485–496). Appleton-Century-Crofts. 1948. DOI: https://psycnet.apa.org/doi/10.1037/11304-053

VIEIRA, M. The history of the SAT is mired in racism and elitism. *Teen Vogue*. 2018. Available in: https://www.teenvogue.com/story/the-history-of-the-sat-is-mired-in-racism-and-elitism

WU, C. D.; MCNEELY, E.; CEDEÑO-LAURENT, J. G.; PAN, W. C.; ADAMKIEWICZ, G.; DOMINICI, F.; LUNG, C. C. S. L., SU, H. J.; SPENGLER, J. D. Linking student performance in Massachusetts elementary schools with the "greenness" of school surroundings using remote sensing. *PLoS ONE*, v. 9, n. 10, p. 1–9. 2014. DOI: https://doi.org/10.1371/journal.pone.0108548

DOI: https://doi.org/10.24933/horizontes.v42i1.1811

Submetido: 19.03.2024.
Aprovado: 25.03.2024.