

The Global Testing Business: Standardization, Inequality, and the Unfairness of One-Size-Fits-All Assessment

S. Tong

ABSTRACT

Standardized English proficiency tests such as the Test of English as a Foreign Language (TOEFL) serve as gatekeepers to higher education for millions of students each year, yet the assumption that a single timed test fairly measures a student's ability is seldom examined critically. This review consolidates research on the global testing industry across three connected questions raised by the field: how standardized tests fail diverse learners, how the commercial testing business profits from exclusion, and how assessment might be reimagined beyond a single score. Drawing on educational psychology, language assessment, and the political economy of education, and using the theory of test validity and fairness as a connecting thread, the review argues that a standardized test measures something real but narrow, and does so under conditions that introduce systematic disadvantage for students who think, process information, or approach knowledge differently. The aim is to bring dispersed evidence within reach of students, educators, and admissions practitioners who encounter these concerns one at a time rather than as parts of a single problem.

KEYWORDS

standardized testing; TOEFL; test fairness; construct validity; language assessment; educational inequality

HOW TO CITE THIS ARTICLE

S. Tong (2026). The Global Testing Business: Standardization, Inequality, and the Unfairness of One-Size-Fits-All Assessment. *The Journal of Young Innovators*, 2(2), pp. 8–15.

DOI <https://doi.org/10.67419/jyi.v2i2.13>

READ ONLINE <https://young-innovator.org/issues/articles/global-testing>

Introduction

The relationship between standardized testing and educational opportunity has drawn sustained attention across several fields over the past few decades. For nonnative speakers of English, taking a standardized proficiency test is unavoidable, because the resulting scores shape consequential outcomes such as admission to a school, the awarding of a scholarship, and access to employment (Puspawati, 2014). At the same time, measurement researchers have long argued that a test score is not a neutral fact but an inference, and that the appropriateness of that inference is what fairness is meant to protect (Messick, 1989). These two conversations bear directly on one another, yet they are often held apart, with the technical theory of fairness on one side and the lived experience of test takers on the other.

This disconnect matters because a test result is treated, in practice, as an objective measure of ability, when in fact it is the product of many interacting processes. A score reflects not only what a student knows but also how the student was feeling, how quickly the student processes written information under time pressure, and how familiar the student is with the cultural assumptions embedded in the test format (Cassady & Johnson, 2002; Puspawati, 2014). When an admissions committee reads a TOEFL score as an indicator of readiness, it inherits all of these hidden influences at once.

This review organizes the research around three questions that structure much of the critical literature on the global testing industry. The first concerns how standardized tests such as the TOEFL fail diverse learners, marginalizing students whose minds, learning styles, or educational traditions fall outside a narrow definition of ability. The second concerns how the commercial testing business profits from exclusion, and whether its incentives work against educational fairness. The third looks beyond the TOEFL to ask what it would mean to measure ability differently, in ways that treat students as whole people rather than as scores. Running through all three is a single idea from measurement theory, most influentially articulated by Messick (1989) and extended in language testing by Kane (2010): that the fairness of a test rests on whether the inferences drawn from its scores are appropriate for the uses to which they are put. That idea serves here as a connecting thread rather than a rigid template, giving the three questions a common analytic center.

Literature Review

The Standardization Trap: When Tests Fail Diverse Learners

The first question is whether standardized proficiency tests reward too narrow a slice of ability, and in doing so systematically disadvantage students who think or learn differently. The most influential challenge to a single measure of ability comes from Gardner (1983), who argued in *Frames of Mind* that intelligence is not one capacity captured by an IQ score or a short-answer test, but rather a set of relatively independent intelligences, including linguistic, logical-mathematical, spatial, musical, bodily-kinesthetic, interpersonal, intrapersonal, and naturalistic, a list he revisited and expanded in a later volume (Gardner, 1999). Sternberg (1985) made a related argument through his triarchic theory, contending that schools reward analytical ability while neglecting the creative and practical intelligences, leaving many student strengths unmeasured. These particular theories remain contested, and this review does not rest on their being correct; psychometric researchers have long argued that a general intelligence factor accounts for much of the variance across cognitive tasks (Visser et al., 2006). What matters for the present argument is the more modest point on which both sides largely agree: a single test loaded with verbal and analytic demands measures one region of a person's abilities well and leaves the rest out of view. Even critics of multiple intelligences do not claim that one timed test captures everything a student can do.

That narrowing turns into a fairness problem the moment the testing situation itself begins to move scores. Research on test anxiety shows that a timed, high-pressure format can depress performance in ways unrelated to knowledge. In a behavioral-genetics study of more than 400 twin pairs, test anxiety was negatively associated with performance on a high-stakes reading comprehension test, operating partly through shared environmental influences rather than reading ability itself (Wood et al., 2016). The proposed mechanism is cognitive: anxiety consumes the working-memory resources needed to process text, so comprehension suffers even among students who read capably under calmer conditions (Cassady & Johnson, 2002). The effect of the format falls unevenly, because the same time pressure that unsettles an anxious test taker also penalizes those who read or process information more slowly.

Studies of extended-time accommodations make this uneven burden visible. When high school students with reading disabilities were given a reading comprehension test under standard timing, they were unable to attempt as many questions as their nondisabled peers, and only extended time allowed them to reach the same number of items

(Lewandowski et al., 2008). A large analysis of the SAT found a similar pattern, with extended time raising scores for students with learning disabilities more than for other test takers, which indicates that the standard time limit had been suppressing their measured ability rather than reflecting it (Mandinach et al., 2005). If a change in timing alone can move a score in this way, then part of what the standard format measures is speed under pressure rather than the underlying ability it claims to assess.

A further question is whether a score reflects academic readiness or merely familiarity with a Western testing tradition. Puspawati (2014) conducted focus groups with 13 international student teachers who had taken the TOEFL and found that many perceived the test as unfair, frequently because its tasks did not fit their cultural backgrounds; she concluded that proficiency tests could be tuned more carefully to cultural difference so that they measure actual language ability rather than test familiarity. A broader literature on cultural bias in standardized assessment points to a similar pattern; a frequently cited example is a Scholastic Aptitude Test analogy built around the word *regatta*, a term familiar to many affluent test takers but unknown to many African American students, so that the item effectively measured cultural exposure rather than reasoning (Bazemore-James et al., 2017). The general pattern is that a test normed primarily on a majority group can produce score gaps reflecting who wrote the test as much as who is more capable. Kane (2010) captures the underlying issue by distinguishing procedural fairness, in which all test takers face the same conditions, from substantive fairness, in which those conditions are genuinely appropriate for everyone; a test can satisfy the first while failing the second, which is exactly what cultural bias reveals.

The Testing Business: Profiting from Exclusion

The second question moves from the individual test taker to the industry that produces the tests. The most useful lens is Messick's (1989) argument that the social consequences of test use are part of validity itself, a principle that reframes questions once dismissed as political. Smith (2016), in *The Global Testing Culture*, develops this system-level view, arguing that over the past three decades standardized testing linked to high-stakes outcomes has spread worldwide until it has come to seem like common sense, to the point where testing becomes almost synonymous with accountability and accountability with educational quality. Smith notes that the problem is not testing itself but the way tests are administered, used or misused, and tied to consequences.

The TOEFL fits this account closely. It is marketed as a global standard, accepted by more than 13,000 institutions worldwide according to its publisher, which gives one

organization's definition of English proficiency substantial influence over students' prospects across many countries. A commercial dimension follows from this position. The organizations that produce these tests are not disinterested referees; a student who intends to study abroad frequently must pay for the test, pay again for a retake when a score falls short, and often pay for preparation as well. That tension is one the literature returns to repeatedly: the commercial incentives of test providers can pull against the goal of educational fairness (McNamara & Ryan, 2011). When difficulty and exclusivity are part of what makes a credential valuable as a gatekeeper, there is not always a strong incentive to make the test easier or more inclusive.

Beyond the TOEFL: Rethinking How We Measure Ability

If a single timed test represents ability narrowly and admits influences unrelated to that ability, what might widen the measure while preserving fairness? The most discussed answer is authentic assessment, and portfolios in particular. A portfolio is a collection of a student's work assembled over time, which researchers describe as a more authentic method because it captures performance in a naturalistic setting across many tasks rather than inferring ability from a handful of items on a single high-pressure day (Birgin & Baki, 2007). Darling-Hammond (1994) argued that standardized tests generally do not reflect the tasks that educators and society expect students to perform, nor do they encourage instruction closely connected to the development of those abilities. Project-based and performance assessments appear in the same literature as ways of measuring capacities a multiple-choice format cannot reach, such as sustained problem solving and collaboration.

But portfolios are labor-intensive to compile and score, and they are harder to grade, which raises concerns about reliability and subjectivity. This is the central trade-off in the debate. Standardized tests are consistent and easily compared but narrow and sometimes biased, whereas authentic assessments are fuller and more responsive to the individual but slower and harder to standardize. Neither approach is without cost, and the honest way to frame the choice is as a design problem rather than a slogan: gains in an assessment's appropriateness for each individual student can come at the expense of the comparability that identical conditions provide (Kane, 2010).

There is also a cultural dimension to rethinking measurement that connects back to the global spread of one testing model (Smith, 2016). Different cultures define educational success differently, and when a single Western model becomes the default, one culture's definition is treated as universal. A more holistic approach to admissions, one that reads a student as a whole person with a context and a history rather than as a score, is an effort to

correct for that. This shift is already visible in practice: a growing number of U.S. colleges have adopted test-optional admissions policies that let applicants decide for themselves whether to submit a standardized score at all (Education Writers Association, 2024). It claims that relying on a single number, produced under artificial conditions, and treating it as objective truth, conceals much of the inequality it is meant to resolve.

Discussion

Read across the three questions, a picture emerges even though the researchers disagree on many particulars. Standardized tests such as the TOEFL measure something real, but they measure it narrowly, and that narrowness lands hardest on students who think differently, process information differently, or come from a non-Western educational tradition (Gardner, 1983; Puspawati, 2014). The testing industry that produces these instruments operates within commercial incentives that do not always align with fairness. The alternatives, from portfolios to holistic admissions, offer a richer view of students but surface a trade-off rather than a simple fix (Darling-Hammond, 1994).

What unifies the three sections is a single underlying question, drawn from the theory of fairness that runs beneath them: whether the inference a school draws from a score is appropriate for the decision it is being used to make (Messick, 1989). Seen this way, anxiety effects, cultural bias, narrow constructs, and market incentives are not separate grievances but different expressions of the same concern. The evidence does not resolve every subordinate debate, and it need not. Whether multiple intelligences are genuinely distinct, or exactly how test anxiety operates, remains disputed; but these open questions sit beneath a point on which the evidence is far more consistent, that a single timed test captures one region of a student's ability and, under pressure and cultural distance, measures it unevenly. The reasonable conclusion is not that testing should be abolished, but that the weight now placed on it is out of proportion to what it can fairly bear. A score of this kind is one piece of evidence about a student, and treating it as the whole story is precisely the inference the theory of fairness warns against.

Implications for Practice

The three questions point to several practical ideas for institutions that rely on standardized scores. Admissions practitioners can read a proficiency score as one piece of evidence rather than a verdict, interpreting it alongside other indicators of preparation. Test designers can attend to the influences that have nothing to do with language ability by

reconsidering rigid time limits and by reviewing items for cultural assumptions that advantage some test takers over others (Bazemore-James et al., 2017). Institutions can weigh the consequences of their testing policies as part of the policy's fairness rather than as a separate political matter (Messick, 1989). None of these ideas requires discarding standardized assessment altogether; they require treating a score as an inference to be examined rather than a fact to be trusted.

Limitations and Future Directions

This review is conceptual rather than experimental, so it can show that the concerns raised across the three questions are plausible and well supported, but not that any specific reform would improve outcomes. Much of the cited evidence derives from particular populations and settings, and more work is needed to understand how these findings generalize across languages, age groups, and testing regimes. The critical discourse analysis of TOEFL policy documents proposed in the wider project would complement this review by examining directly how test providers frame proficiency and fairness in their own materials. Future research should also test alternative assessment models through controlled studies that compare their fairness and predictive value against standardized tests across diverse groups of test takers.

Conclusion

This review has consolidated evidence from educational psychology, language assessment, and the study of education systems around three questions about the global testing business: how standardized tests fail diverse learners, how the testing industry profits from exclusion, and how assessment might be reimaged beyond a single score. Beneath all three runs a common concern with whether the inferences drawn from a test score are appropriate for their use. A standardized test such as the TOEFL measures something real, but it measures it narrowly and under conditions that introduce systematic disadvantage for students who think, process information, or approach knowledge differently. The most defensible position the evidence supports is not the abolition of testing but a more honest relationship with it, one that stops treating a single number, produced under artificial conditions, as the objective truth of a student's worth.

References

- Bazemore-James, C. M., Shinaprayoon, T., & Martin, J. (2017). Supporting students who experience cultural bias in standardized tests. *Trends and Issues in Academic Support, 2016–2017*, 4–11.
- Birgin, O., & Baki, A. (2007). The use of portfolio to assess student's performance. *Journal of Turkish Science Education, 4*(2), 75–90.
- Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270–295.
<https://doi.org/10.1006/ceps.2001.1094>
- Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5–30. <https://doi.org/10.17763/haer.64.1.j57n353226536276>
- Education Writers Association. (2024). *Test-optional vs. test-required policies at colleges*.
<https://ewa.org/issues/higher-ed/test-optional-vs-test-required-policies-at-colleges>
- Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. Basic Books.
- Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century*. Basic Books.
- Kane, M. T. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182.
<https://doi.org/10.1177/0265532209349467>
- Lewandowski, L. J., Lovett, B. J., & Rogers, C. L. (2008). Extended time as a testing accommodation for students with reading disabilities. *Journal of Psychoeducational Assessment, 26*(4), 315–324. <https://doi.org/10.1177/0734282908315757>
- Mandinach, E. B., Bridgeman, B., Cahalan-Laitusis, C., & Trapani, C. (2005). *The impact of extended time on SAT test performance* (College Board Research Report No. 2005-8; ETS Research Report No. RR-05-20). The College Board. <https://doi.org/10.1002/j.2333-8504.2005.tb01997.x>
- McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*(2), 161–178.
<https://doi.org/10.1080/15434303.2011.565438>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education & Macmillan.
- Puspawati, I. (2014). Fairness issues in a standardized English test for nonnative speakers of English. *TESOL Journal, 5*(3), 555–572. <https://doi.org/10.1002/tesj.157>

- Smith, W. C. (Ed.). (2016). *The global testing culture: Shaping education policy, perceptions, and practice*. Symposium Books.
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge University Press.
- Visser, B. A., Ashton, M. C., & Vernon, P. A. (2006). Beyond g: Putting multiple intelligences theory to the test. *Intelligence*, 34(5), 487–502. <https://doi.org/10.1016/j.intell.2006.02.004>
- Wood, S. G., Hart, S. A., Little, C. W., & Phillips, B. M. (2016). Test anxiety and a high-stakes standardized reading comprehension test: A behavioral genetics perspective. *Merrill-Palmer Quarterly*, 62(3), 233–251. <https://doi.org/10.13110/merrpalmquar1982.62.3.0233>