FRASER INSTITUTE
2022

WHY EDUCATIONAL TESTING IS NECESSARY

Ze'ev Wurman

Contents

In school, you're taught a lesson and then given a test.
In life, you're given a test that teaches you a lesson.

—Tom Bodett, American author

Executive Summary

Nations invest enormous sums of money in their elementary and secondary education—on the order of two to four percent of their Gross Domestic Product. This is as much as, or even more, than their investment in national defense. Consequently, it's only natural that they are interested in finding out how effectively this money is spent. Furthermore, in recent decades the educational attainment of country's population was statistically linked to its economic success, a link that necessarily will only strengthen in the future as our society becomes even more dependent on technology.

Educational testing or, more properly, the *results* of educational testing, are an essential metric for evaluating the quality of the educational system and for informing the public and policymakers about what is and is not functioning as planned. Furthermore, individual students are necessarily interested in their own educational performance, as are parents in the performance of their child, so they can realistically evaluate the chances for the student's future educational progress as well as their future career path, whether in academia or the workplace.

To provide such metrics educational testing must be valid and objective. Valid in the sense that the results reflect the actual knowledge and skills that schools are supposed to impart to their students, and objective in the sense that they are accurate in their scoring and untainted by irrelevant considerations. Addressing the validity of a test is in the hands of content experts who create the test items. Addressing test objectivity is a matter of the testing process—the uniformity of its administration and of its scoring; in other words, its standardization. Hence the common expression "standardized test," which encompasses all those features and allows test results to be compared across schools and across time.

Other educational testing exists too. For example, teachers routinely assess the progress of their students by administering classroom tests that they generate themselves or lift from textbooks. These so-called formative tests are invaluable for the teacher, yet they may be neither objective nor even valid, and their results certainly cannot be compared across different schools and classrooms or over time.

While almost all teachers support formative in-class testing, many resist external objective testing. In that they are often joined by school administrators, and sometimes even by politicians. The reasons for such resistance are pretty obvious: objective test results often conflict with a teacher's own perception of his or her students' achievement, and no administrator or politician enjoys seeing disappointing results that they might be forced to explain to an unhappy public.

This report describes educational testing in some detail, argues for the importance of test standardization, and offers evidence of its objectivity particularly when contrasted with the evidence of grade inflation in public schools. Further, it documents its salutary effects on student achievement contrary to the cliché often whipped out by testing opponents that "weighing the cow doesn't make it fatter"—actually, it seems that it does, notably when there are significant stakes associated with the results of the test to schools and particularly to students. In the same vein the report also addresses other common, straw-man fallacies that test opponents often bring up, such as that testing necessarily narrows the curriculum.

The report concludes that well executed, valid, and objective educational testing seems irreplaceable as an oversight tool for politicians and the public and as a way to guide improvement for students and schools.

Introduction

Objective performance assessment is critical to any meritocracy, be it imperial China 2500 years ago or modern North America. However measured, a key aspect of successful societies lies in the relative objectivity of the assessment, rather than reliance on subjective evaluations often driven—or at least perceived as driven—by family ties, pecuniary benefits, or personal preferences.[1]

It is widely accepted that education is a necessary ingredient to the progress and welfare of a society, and hence it is important to make it effective. Further, educational effectiveness is accepted as a key element in the West's competitiveness. Given all this, it logically follows that if we want to improve our society, we should strive to make educational evaluation as valid and objective as possible so we can correctly identify the strengths and weaknesses of our educational systems and improve them as necessary.

This report describes many of the types of achievement tests that our educational systems use, their characteristics, and value, with a particular focus on their objectivity and utility and their importance for both systemic and individual student evaluation.

[1] While meritocracy is critical to society's success, it is not necessarily the sole criterion used to drive decisionmaking. Moral principles such as equity or social harmony necessarily play a role in a democracy, but we should be wary of conflating them with effectiveness-related meritocratic arguments.

Types of Tests

Objectivity in evaluation goes to the process of how the evaluation process is administered, and how its results are scored. Both should be standardized to minimize measurement errors, provide reliable comparability, and reduce bias. In itself this is insufficient—we can objectively score answers to meaningless set of questions, which will not be helpful. In other words, useful tests, in addition to being objective, must also validly sample the knowledge we attempt to assess. This does not imply, however, that identical scores must be interpreted in an identical way, or that they will have identical implications—these may also depend on the way that society collectively chooses to attach value to the scores, tempered by additional societal considerations like equity, fairness, and possibly others. It is important, however, to keep those additional considerations separate, explicitly balancing them with the meritocratic test score. Unless we do, we run the risk of corrupting the meritocratic results and tainting the assessment itself with overt biases, which may cause the loss of public trust in its inherent fairness and objectivity.

Educational tests can be classified by their purpose (diagnostic, formative, summative), by their type (achievement, aptitude), by their item types (selected response, constructed response), and their reporting (standard-based, norm-referenced). Diagnostic tests attempt to identify learning disabilities and will not be addressed here. Neither will formative tests, which are typically administered by classroom teachers and serve to assess students' progress and adjust instruction, nor aptitude tests, which are intended to gauge students' potential. This report will focus primarily on summative achievement tests, which in recent years tend to be standards-based, sometimes also called criterion-referenced, i.e., measuring achievement against a set of fixed educational grade-level expectations (standards). This is in contrast to older norm-referenced tests that assessed achievement compared to representative, typically national, sample reference groups.

Yet another important categorization comes up relating to the end-use of testing results. If the results should inform individual students and parents, the test must be administered in a census form—to each indi-

vidual student, in its fullness.[2] If, on the other hand, the test is intended to inform the general public or policymakers, it may be administered to a representative sample of students and each student may be exposed to just a subset of test items.[3]

Objective and standardized assessment is often associated in the public mind with the selected response ("multiple choice") test format. Such association, while understandable, is incorrect. The objectivity of assessment is predicated on the clear and unambiguous statement of valid test items, on the test's uniform administration, and on a uniform way of scoring and standardized scaling;[4] it is not dependent on the particular format of the test items. Remembering these requirements, it is obvious that it is easier to make a multiple choice item clear and unambiguous than it is to pose an open-ended question in that way; it is even more obvious that it is easier to assure objectivity in scoring a multiple choice item than it is to score an open-ended item. It is also much cheaper to objectively score a multiple-choice item than is to score an open-ended one.[5] Consequently, some people mistakenly associate objective assessment with the multiple-choice format, even though one can have objective assessments that include open-ended constructed-response questions.

To summarize, most people would agree that valid and objective tests are a vital component of our education systems. Without the information they provide, we would all be lost: parents would not know how

[2] In recent years Computer Adaptive Testing (CAT) came into fashion where each student may not necessarily be exposed to the same full set of questions as another, yet each must answer a sufficient number of questions for a reliable assessment of the complete breadth of the curriculum.

[3] In most countries census testing occurs only at specific transition points of their educational system, such as the transition between elementary, middle, and high school grades. Additionally, many educational systems have graduation and/or qualification testing at the end of high school, such as the French baccalauréat, the British A-Levels, or the American SAT. Census testing can both inform individual students, as well as form the basis of systemic evaluation. In contrast, sampled testing supports only systemic evaluation and may occur irregularly every so often.

[4] Scaling refers to adjusting the scores of periodically administered tests in such a way that they will be comparable over time. This is necessary since test item difficulty may vary slightly across multiple test administrations despite efforts to keep them similar in difficulty.

[5] Objective scoring of open-ended items will typically require intensive training of human scorers, assigning two scorers to evaluate each response, and having an adjudication process in case their scores differ. Clearly this makes objective scoring of open-ended items slower and much more expensive than scoring selected-response items.

well the school system was serving their children; policymakers would not know if their programs were working or how they could be improved; teachers would have no way of knowing how their students performed in previous grades or how they are currently performing in comparison to other classes and other schools; and even the students themselves would not know how well they truly grasped the material. To use an obvious metaphor, without objective tests, our education system would be adrift; like a ship with no navigation system, it would not know where it was or where it was going, and it would have no way to correct course. All it could do would be to rev the engines and hope for the best.

Having said that, we should recognize that many educators resent testing and believe that testing, and particularly high-stakes summative testing, is counterproductive. In some sense this is not surprising since testing, in effect, represents a potential critique of educators and the pedagogy they espouse. Yet such objections would be more acceptable had the outcomes of our education system, particularly in the West, been more robust. As it is, the general public recognizes the self-serving elements in educators' objections and overwhelmingly supports educational testing. For example, a survey conducted by Mark Holmes in Ontario in 1998 found that "while only 11 per cent of directors of education agreed with annual achievement testing of elementary students, the proposal was supported by 59 per cent of a similarly educated comparison group of non-educators" (Holmes, 1998: 142-43). In a more recent US poll some 68 percent of Americans supported annual student testing, while only 48 percent of teachers did. Moreover, 20 percent fewer teachers who were members of teacher unions supported testing than did non-members (Education Next, 2018). A 2022 poll of Canadian parents found that a large majority—84 percent—supported standardized testing, while noting the opposition of multiple teachers' unions to such testing (MacPherson, 2022).

Standardized Tests are a Vital Source of Information for Parents and Policymakers

In earlier times the grade level a student reached was deemed a sufficient indicator of how educated he or she was. For a host of historical reasons that is no longer the case, yet the need for more precise assessment of students' education has grown. For example, two Maryland graduates with similar GPAs, one from Baltimore and one from Bethesda, would likely not have attained the same level of literacy or numeracy. School Grade Point Averages (GPAs) and other teacher assessments are not reliable indicators of student achievement. Nor is the level of educational attainment as measured by the names of courses taken or grade promotions. For example, the recently released US 2019 National Assessment of Education Progress (NAEP) *High School Transcript Study* shows that while students are taking more rigorous classes and pass them with better grades, their actual knowledge of the subject matter has decreased (Hess, 2022). To be able to compare one school to another, there needs to be a reliable common measurement.

This is the most basic good that standardized test results provide—they give parents and policy makers objective and comparable information on student performance, enabling them to judge schools and districts in relation to one another, using a common unit of measurement.

Standardized tests are, in fact, the *only* real external check we have on the quality of our schools. As Richard Phelps, a former US Government Accounting Office (GAO) researcher laid it out: "If none of the curriculum is tested, we cannot know if any of it has been learned" (Phelps, 1999: 25). Without standardized tests, no one outside the classroom can reliably gauge student progress. No district or provincial education officials. No political leader. No taxpayer. No parent. No student. Each has to accept whatever the teacher says and, without standardized tests, no teacher has any point of comparison outside of their experience either.

Objective and comparable information about school performance is the basis for any clear and rational discussion about funding priorities or program formulation and implementation. Without such measurements, it

would be impossible to have a meaningful discussion about education on the state or national level.

Moreover, objective measurements help to frame and anchor debates about education policy, so that such debates do not become untethered from the reality of the problems at hand, as they can be if they are overtaken by demagogues who demand unhelpful reforms or interest groups who seek to protect the status quo.

Teacher assessments are inadequate in this regard. While teacher assessments may help to fill out our understanding of student performance, they can never replace standardized tests of students as a means of conveying information about school quality to parents and policymakers. There are several reasons for this.

One is that "individual teachers can also narrow the curriculum to that which they prefer. Grades are susceptible to inflation with ordinary teachers, as students get to know a teacher better and learn his idiosyncrasies. A teacher's grades and test scores are far more likely to be idiosyncratic and non-generalizable than any standardized test" (Phelps, 1999: 25). A summative objective standard-based test may be one of the more effective tools to ascertain that the intended curricular content is taught, as by design such a test broadly samples that content. In fact, the criticism that it encourages "teaching to the test" becomes a positive attribute with such standard-based testing as "teaching the curriculum" and "teaching to the test" become indistinguishable.

It is not uncommon for grades in schools and even in individual classrooms to drift toward an equilibrium, "so that most grades are As and Bs, without improvements in achievement" (Evers, 2001). This belies dramatic differences in the quality of instruction and the rate of learning between different schools and different teachers.

In-class assessment by teachers often does not provide an accurate picture of their students' subject matter knowledge. "It appears that nearly *everything* is considered when assigning a mark. There are probably two reasons for this. First, educators want to consider all relevant aspects of a student's classroom experience when assigning a mark. At the same time, there is apparently no clear consensus about which factors *are* relevant to assigning a grade" (Cizek, 1996 (emphasis in original)). For many teachers and professors in education schools, the mastery of subject matter is but one of the factors taken into account, and indeed, may not even be the most important one, as one elementary teacher said: "Getting the child through the level with a positive attitude and good memories is more important than a raw number grade... Shaping the kids' minds through group interaction, effort, and participation is more important than averaging tests and quiz scores" (Cizek, 1996; Farkas and Duffett, 2010).

Testing Makes Our Schools More Accurate and More Responsive

Far from discouraging students, formative and diagnostic testing enables teachers to identify problems early on, so they can respond promptly, before the student falls irretrievably behind and becomes truly discouraged.

An example of this is early reading intervention, which is absolutely necessary if struggling students are likely to avoid a lifetime of reading problems. Barbara Foorman and her colleagues write: "There is little evidence that [such students] catch up in reading skills, in spite of the widespread belief among educators in developmental delay—the late bloomer phenomenon" (Foorman, Fletcher, and Francis, 2019: 85).

Fortunately, research shows that early intervention can be effective and early identification via objective diagnostic testing is critical to that end. Foorman et al. cite a study that "identified children in kindergarten who had poor phonological awareness, that is, they had difficulty blending and segmenting sounds in speech. By second grade, one-on-one tutoring brought 75 percent of children to grade-level reading." Another study "identified middle-class children with very low word recognition skills at the beginning of grade one. After one semester of one-on-one tutoring, 70 percent were reading at grade level. After two semesters, more than 90 percent were at grade level" (Foorman, Fletcher, and Francis, 2019: 83).

- For intervention to succeed, it has to begin before the student is in third grade, making early detection vital. This in turn requires the systematic testing of all students.
- Foorman et al. believe that performance-based assessments, "in which the student constructs an original response (that is, displays procedural knowledge)" are inadequate to the task: "These assessments rarely present evidence of reliability or validity and generally do not measure transfer of knowledge independently of a specific curriculum" (Foorman, Fletcher, and Francis, 2019: 90).
- They conclude: "The judicious use of multiple-choice formats to assess declarative knowledge may be the most valid, reliable, and useful way to proceed" (Foorman, Fletcher, and Francis, 2019: 90).

We should not rely on teacher assessments of early reading. Because there can be a great deal of variance not just between schools but between individual classrooms, reliance on teacher assessments can obscure problems and hinder coordination between the educators themselves, as students move from grade to grade. Bill Evers recounts the experience of Judy Codding, a former principal of Pasadena High School in California, in her efforts to determine why her school's students were struggling in their math classes: "Wanting to know how much students coming into her school as freshman knew and what classes they should be placed in, she gave them objective diagnostic tests. The 'horrendous results' showed that incoming freshmen were much less prepared than their middle-school teachers had said they were" (Evers, 2001).

Phelps quotes a 1936 study that describes an experiment where two researchers, Starch and Elliott, made copies of two actual English examination papers and sent them to teachers to grade and return. The marks awarded ranged from 58 to 98 percent. One paper, graded by 142 teachers, received 14 marks below 80 percent and 14 above 94 percent.… Starch and Elliott repeated the procedure with duplicate geometry tests. Teachers' marks on the 116 returned papers ranged from 28 to 92 percent… According to the authors, "*This type of experiment has been repeated many time by investigators, and always with similar results*" (Lincoln and Workman, 1936, quoted in Phelps, 2007: 114 (emphasis added))**.**

Clearly, classroom grading by teachers is likely to be anything but reliable or standardized.

Standardized Tests Effectively Capture What Students Know

Even if teachers mean well, the fact remains that their assessments will likely not be as accurate an assessment of student knowledge as standardized tests are. The reason is that "even teachers who endeavor to grade their students on the basis of academic achievement are unlikely to have received more than cursory training in testing and measurement. Those who criticize standardized tests for their alleged imperfections of structure and content seldom mention that standardized tests are written, tested, and retested by large groups of Ph.D.s with highly technical training in testing and measurement" (Phelps, 2003: 226). Hence standardized tests are likely to be much more accurate and reliable than teachers' own assessments.

The effort put into developing a major standardized test is immense: "A typical large-scale test goes through multiple rounds of research, development, and pilot testing to make sure that it is reliable—in other words, that scores it yields are consistent and stable across multiple test administrations. A large-scale test is also expected to be valid, which means that it measures what it claims to measure and leads users to draw accurate and meaningful conclusions about what students know and can do" (Kober, 2002: 6).

The notion that standardized tests, and particularly multiple-choice ones, must be inherently simplistic—that they favour low-order thinking skills—is similarly misguided. As Richard Phelps explains: "Test items can be banal and simplistic or intricately complex and, either way, their response format can be multiple-choice or open-ended. There is no necessary correlation between the difficulty of a problem and the response format. Even huge, integrative tasks that require fifty minutes to classify, assemble, organize, calculate, and analyze can, in the end, present the test-taker with a multiple-choice response format. Just because the answer to the question is among those provided, it is not necessarily easy or obvious how to get from the question to the answer" (Phelps, 1999: 15).

Some critics argue that this does not matter, because even if a standardized test is valid and reliable, it may not measure the knowledge and

skills that are truly important in life later on, such as creativity, the ability to work in groups, and so on. This is put to the lie by a multitude of studies that link student performance on validated standardized tests to economic prosperity, both for individuals and nations. If the skills being measured by such standardized tests were irrelevant, then there would be no such correlation.

As an example, Stanford's Eric Hanushek cites three recent studies, all of which "employ different nationally representative data sets that follow students after they leave schooling and enter the labor force." These studies came to the same conclusion, which is that "when scores are standardized, they suggest that one standard deviation increase in mathematics performance at the end of high school translates into 12 percent higher annual earnings." Put another way, if the median earnings in 2001 were about $30,000, "a one standard deviation increase would boost these by $3,600 for each year of work life" (Hanushek, 2006: 450). And Hanushek suspects that "these estimates provide a lower bound on the impact of high achievement."[6]

Gregory Cizek makes an important point that is often lost on critics who say that tests do not measure everything a student needs to know:

> The relationship of standardized achievement tests to the essential goals of education--say, becoming a responsible, productive citizen--relies to a great extent on the principle of 'necessary but not sufficient.' It is easy to see that acquiring proficiency in reading, mathematics, writing, and so on, is *necessary* to accomplishing essential educational that goal. Admittedly, other student characteristics—such as personal responsibility and creative thinking or problem solving—are also necessary for the goal to be achieved. However, an extra measure of personal responsibility or creative thinking cannot compensate for deficits in a student's knowledge of language or mathematics, or in the student's ability to communicate his or her ideas. Students *may* become productive and responsible contributors to society if they master fundamental academic skills; they will almost certainly be unable to do without them. (Cizek, 1998: 2-3)

[6] One should note that the specific content that such tests assess is less important than it might seem, as long as the domains of the tested knowledge are similar and the questions are well posed. For example, PISA (Programme for International Student Assessment) is a curriculum-independent "literacy" test, while TIMSS (Trends in International Mathematics and Science Study) is a curriculum-based content test, yet achievement on both correlates similarly with economic impact (Eric Hanushek, personal communications, circa 2015).

In other words, while valid and reliable standardized tests may be unable to analyze everything necessary for a successful education, "they can yield highly accurate, dependable information about a finite but vital constellation of knowledge and skills" (Cizek, 1998: 3).

Standardized Tests Are Cost-Efficient

Standardized, largely or solely multiple choice tests are cost-efficient, in that they can draw out extensive information about large numbers of students in a short period of time and for a much lower cost than other, less standardized measures such as essays or large and complex open-ended "performance tasks."

Harvard economist Caroline Hoxby reviewed the 2000-01 costs of 25 US states' assessment systems, and found that they ranged from a low of $1.79 per pupil for South Carolina to a high of $34.02 per pupil for Delaware. Yet even Delaware's expenses represent only 0.4 percent of the average per pupil expenditure in America, which was $8,157 in the 2000-01 school-year. Hoxby also compares the cost of developing and implementing assessments to two other popular reforms: class-size reductions and pay raises for teachers.

Begin with class-size reduction: "Given that teacher compensation represents 54 percent of the average American school's costs and that items proportional to the size of school buildings (building, heating, etc.) represent another 22 percent of the average American school's cost, a 10 percent reduction in class size costs about $615 per student in the United States," or 12,399 percent more than the current average cost of assessment (Hoxby, 2002: 11).

Pay raises for teachers fairs similarly: "It would cost the average American school $437 per student to raise teachers' compensation by 10 percent," or 5,011 percent more than the current average cost of assessment (Hoxby, 2002: 11-12).

A 1993 US Government Accounting Office [GAO] study found the average cost of state testing to be around $15 per student including staff time for multiple choice tests and about $20 per student when it included some open-ended items. At the time (1990-91), the per-student expenditures were $5,885. In other words, the cost of testing was about 0.3 percent of per-student spending—clearly a miniscule cost to spend on quality assurance [US GAO, 1993).

What Does the Research Show About Testing and Student Achievement?

When coupled with an accountability system, standardized tests can have a powerful effect on student achievement. Studies (Phelps, 2019, 2012; Bishop, 2004, 2005) appear to show that accountability systems aimed at students are the most effective, but if they are aimed at the schools, as the No Child Left Behind (NCLB) assessments were, they can still have a significant positive effect.

Tom Loveless of the Brookings Institute cites "a 2003 study examining the effects of accountability on state NAEP scores, [in which] Martin Carnoy and Susanna Loeb rate the strength of each state's system on a five-level scale" (Loveless, 2005: 9; Carnoy and Loeb, 2002). Both student and school accountability were factored into the rating. Controlling for differences in spending and student demographics, Carnoy and Loeb found that "between 1996 and 2000, the stronger the accountability system,[7] the

[7] Carnoy & Loeb describe the accountability rating as follows:

> "The 0–5 scale captures degrees of state external pressure on schools to improve student achievement according to state-defined performance criteria. States receiving a zero do not test students statewide or do not set any statewide standards for schools or districts. States that require state testing in the elementary and middle grades and the reporting of test results to the state but no school (or district) sanctions or rewards (no or weak external pressure) get a 1. Those states that test at the elementary and middle school levels and have moderate school or district accountability sanctions/rewards or, alternatively, a high school exit test (that sanctions students but pressures schools to improve student performance) get a 2. Those states that test at the lower and middle grades, have moderate accountability repercussions for schools and districts, and require an exit test in high school, get a 3. Those that test and place strong pressure on schools or districts to improve student achievement (threat of reconstitution, principal transfer, loss of students) but do not require a high school exit test get a 4. States receiving a 5 test students in primary and middle grades, strongly sanction and reward schools or districts based on improvement in student test scores, and require a high school minimum competency exit test for graduation."

greater the gains states made in raising the percentage of eighth graders functioning at or above the basic level in mathematics. A two-rank increase in the accountability index was associated with about a one-half standard deviation gain, which is statistically significant. The results were significantly positive for black, white, and Hispanic students, and held up after controlling for how many students each state excludes from NAEP testing" (Loveless, 2005: 9).

In a similar study, John Bishop of Cornell University examined systems targeting both students and educators, using the 1996 and 1998 NAEP scores of eighth graders in states with different accountability regimes—i.e., "for students, meeting basic course requirements, passing minimum competency exams, and passing a curriculum-based external exit exam; and for schools, receiving rewards or sanctions based on test scores" (Loveless, 2005: 9-10). He found that "students in states requiring curriculum-based external exit exams (New York and North Carolina) exhibited the highest levels of achievement, with an advantage of 0.45 grade levels in math and science, followed by states that reward and sanction schools, with gains of 0.20 grade levels. Minimum competency tests had a positive but insignificant effect" (Loveless, 2005: 9-10).

Accountability systems create incentive effects, which Bishop detects in his regression analyses, and ~~to~~ include "an increase in knowledge and skills caused by students studying more, paying better attention in class, and taking tougher courses, in a high-stakes testing environment" (Phelps, 2003: 239).

Bishop studied high-stakes exit exams in Canada in the early 1990s, when, as today, only some provinces had curriculum-based exit examinations. He found that the examination systems had pervasive effects on school administrators and teachers and students. In the provinces with external exams,

- "*Schools* were significantly more likely to: employ specialist teachers in math and science; employ teachers who had studied the subject in college; have high quality science labs; schedule extra hours of math and science instruction; assign more homework in math, in science and in other subjects; have students do or watch experiments in science class; and schedule frequent tests in math and science class.
- "*Teachers* were significantly more likely to: give more homework, cover more difficult material; schedule more quizzes and tests; reduce the time students spend doing group problem solving; increase the time students work alone; and schedule more experiments in science class.

- "*Students* were significantly more likely to: watch less television (40 percent less), report that their parents want them to do well in exam subjects (4 to 6 percent more likely) and talk to them about what they are learning in school, read more for pleasure, and choose educational programs when they did watch television.
- Perhaps even more importantly, the improved academic performance brought about by testing and accountability has effects that last far beyond graduation: "When other factors that influence academic achievement are controlled for, students from states, provinces, or countries with medium or high-stakes testing programs score better on neutral, common tests and earn higher salaries after graduation that do their counterparts from states, provinces, or countries with no- or low-stakes tests." (Bishop, 1994: 22-26 as summarized in Phelps, 2003: 241)

A study of curriculum and instruction systems across 30 countries that participated in the 1994-95 TIMSS revealed a strong relationship "between the number of decision points—high-stakes selection points[8]—which serve as quality controls and student performance on the TIMSS 8th-grade mathematics exam. The more high stakes selection points, the better the country performance" (Phelps, 2003: 221). The study also suggests that there is "an exponential relationship between quality control and student achievement; after a certain critical mass of quality control measures are implemented, student achievement can really take off" (Phelps, 2001; Bishop, 1997).

[8] Decision points (selection points) are defined as "an occasion when a student performance standard is actually applied: a judgment is made—for example, that a student achieves or does not achieve a standard—and an appropriate consequence results." Most often, decision points consist of high-stakes tests or selective admission to certain schools or curricular tracks.

Conclusions

Modern standardized curriculum-aligned educational testing is important for maintaining a meritocratic system with high standards. While norm-referenced, off-the-shelf tests that were frequently used in the past were disconnected from actual curricula, this is no longer true for curriculum-aligned tests. Further, modern tests are carefully checked against group biases that were present in some older tests. Attacks on standardized testing by interest groups are fed by outdated and incorrect information, by reluctance to reward merit, and by resistance to change when indicated by those tests.

Curriculum-aligned standardized tests with stakes for students promote both higher student achievement and school improvement by sending a clear and powerful signal to all stakeholders—students, parents, teachers, and administrators—about what is working and what needs improvement. It's no wonder that the public strongly supports such testing.

References

Bishop, John H. (1994). *Impact of Curriculum-Based Examinations on Learning in Canadian Secondary Schools.* Working Paper #94-30. Center for Advanced Human Resource Studies, Cornell University.

Bishop, John H. (1997). *Do Curriculum-Based External Exit Exam Systems Enhance Student Achievement?* Working Paper 97-28. Center for Advanced Human Resource Studies, Cornell University. <https://ecommons.cornell.edu/handle/1813/77025>, as of May 18, 2022.

Bishop, John H. (2004). High School Diploma Exams: Explaining High Achievement Levels in Students of Some Commonwealth Countries. *Fraser Forum* (July). Fraser Institute. <https://www.heartland.org/_template-assets/documents/publications/15762.pdf>, as of May 18, 2022.

Bishop, John H. (2005). *High School Exit Examinations: When Do Learning Effects Generalize?* Working Paper 05-04. Center for Advanced Human Resource Studies, Cornell University. <https://ecommons.cornell.edu/bitstream/handle/1813/77276/WP05_04.pdf>, as of May 18, 2022.

Carnoy, Martin, and Susanna Loeb (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Education Evaluation Policy Analysis* 24, 4.

Cizek, Gregory J. (1996). Grades: The Final Frontier in Assessment Reform. *NASSP* [National Association of Secondary School Principals] *Bulletin* 80, 584 (December): 103-110.

Cizek, Gregory J. (1998). *Filing in the Blanks: Putting Standardized Tests to the Test.* Thomas B. Fordham Foundation. <https://fordhaminstitute.org/national/research/filling-blanks-putting-standardized-tests-test>, as of May 18, 2022.

Education Next (2018). *2018 EdNext Poll Interactive*. Education Next. <https://www.educationnext.org/2018-ednext-poll-interactive>, as of May 18, 2021.

Evers, William (2001, August 20). What Do Tests Tell Us? *Hoover Daily Report*. Hoover Institution. <https://www.hoover.org/research/what-do-tests-tell-us>, as of May 18, 2022.

Farkas, Steve, and Ann Duffett (2010). *Cracks in the Ivory Tower? The Views of Education Professors Circa 2010.* Thomas Fordham Institute (September). <https://files.eric.ed.gov/fulltext/ED512010.pdf>, as of May 18, 2022.

Foorman, Barbara, Jack M. Fletcher, and David J. Francis (2019). Chapter 3: Early Reading Assessment. *Part Two: Constructive Uses of Tests.* In Williamson Evers and Herbert Walberg (eds.), *Testing Student Learning, Evaluating Teacher Effectiveness*. Hoover Institution. <https://www.hoover.org/sites/default/files/uploads/documents/0817929827_79.pdf>, as of May 18, 2022.

Hanushek, Eric A. (2006). Alternative School Policies and the Benefits of General Cognitive Skills. *Economics of Education Review* 25: 447-462. <http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%202006%20EEduRev%2025%284%29.pdf>, as of May 18, 2022.

Hess, Frederick M. (2022, March 24). When 'Rigorous' Courses Aren't. *The Dispatch.* American Enterprise Institute. <https://www.aei.org/op-eds/when-rigorous-courses-arent>, as of May 18, 2022.

Holmes, Mark (1998). *The Reformation of Canada's Schools*. McGill-Queen's University Press.

Hoxby, Caroline M. (2002). *The Cost of Accountability.* Working Paper 8855. National Bureau of Economic Research. <https://www.nber.org/system/files/working_papers/w8855/w8855.pdf>, as of May 18, 2022.

Kober, Nancy (2002). *What Tests Can and Cannot Tell Us.* TestTalk for Leaders 2 (October). Centre on Education Policy. <https://web.archive.org/web/20030821213952/http://www.cep-dc.org/testing/testtalkoctober2002.pdf>, as of May 18, 2022.

Lincoln, Edward Andrews, and Linwood L. Workman (1935). *Testing and the Uses of Test Results*. Macmillan.

Loveless, Tom, Robert M. Costrell, and Larry Cuban (2005). *Test-Based Accountability: The Promise and the Perils*. Brookings Papers on Education Policy 8. Brookings Institution. <https://nicspaull.files.wordpress.com/2011/04/loveless-2005-test-based-accountability.pdf>, as of May 18, 2022.

MacPherson, Paige (2022). *Strong Parental Support for Standardized Testing across Canada*. Research Bulletin. Fraser Institute. <https://www.fraserinstitute.org/sites/default/files/strong-parental-support-for-standardized-testing-across-canada.pdf>, as of May 18, 2022.

Phelps, Richard P. (1999). *Why Testing Experts Hate Testing.* Fordham Report 3, 1. Thomas B. Fordham Foundation. <https://files.eric.ed.gov/fulltext/ED429089.pdf>, as of May 18, 2022.

Phelps, Richard P. (2001). Benchmarking to the World's Best in Mathematics: Quality Control in Curriculum and Instruction Among the Top Performers in the TIMSS. *Evaluation Review* 25, 24 (August): 391-439. <https://richardphelps.net/BenchmarkingArticle.pdf>, as of May 18, 2022.

Phelps, Richard P. (2003). *Kill the Messenger: The War on Standardized Testing.* Transaction Publishers.

Phelps, Richard P. (2007). *Standardized Testing*. Peter Lang Primer.

Phelps, Richard P. (2012). The Effect of Testing on Student Achievement, 1910-2010. *International Journal of Testing* 12: 21-43. <http://edmeasurement.net/MAG/Phelps%202012%20effects%20of%20testing.pdf>, as of May 18.

Phelps, Richard P. (2019). Test Frequency, Stakes, and Feedback in Student Achievement: A Meta-Analysis. *Evaluation Review* 43 (3-4): 111-151.

United States, General Accounting Office [US GAO] (1993). *Student Testing: Current Extent and Expenditures, with Cost Estimates for a National Examination*. Report to Congressional Requesters (January), number GAO/PEMD-93-8. GAO. <https://www.gao.gov/assets/pemd-93-8.pdf>, May 18, 2022..

About the Author

Ze′ev Wurman

Ze'ev Wurman is the Chief Software Architect with MonolithIC3D Inc. and a Research Fellow with the Independent Institute in Oakland, California. He has served as a senior policy adviser with the Office of Planning, Evaluation and Policy Development at the U.S. Department of Education. Throughout the development of the Common Core standards in 2009-2010 Wurman analyzed the mathematics drafts for Pioneer Institute and for the State of California. In the summer of 2010, he served on the California Academic Content Standards Commission that evaluated the suitability of the Common Core standards for California. Wurman holds over 45 US patents and earned his BSc and MSc degrees in Electrical Engineering from the Technion, Israel Institute of Technology, in Haifa, Israel.

Acknowledgments

This study was generously funded with support from the Lotte and John Hecht Memorial Foundation. The author thanks the anonymous referees for their useful comments on an earlier draft. Any remaining errors are the sole responsibility of the author. As the researcher has worked independently, the views and conclusions expressed in this paper do not necessarily reflect those of the Board of Directors of the Fraser Institute, the staff, or supporters.

Publishing Information

Distribution

These publications are available from <http://www.fraserinstitute.org> in Portable Document Format (PDF) and can be read with Adobe Acrobat® or Adobe Reader®, versions 8 or later. Adobe Reader® DC, the most recent version, is available free of charge from Adobe Systems Inc. at <http://get.adobe.com/reader/>. Readers having trouble viewing or printing our PDF files using applications from other manufacturers (e.g., Apple's Preview) should use Reader® or Acrobat®.

Ordering publications

To order printed publications from the Fraser Institute, please contact:

- e-mail: sales@fraserinstitute.org
- telephone: 604.688.0221 ext. 580 or, toll free, 1.800.665.3558 ext. 580
- fax: 604.688.8539.

Media

For media enquiries, please contact our Communications Department:

- 604.714.4582
- e-mail: communications@fraserinstitute.org.

Copyright

Date of issue

June 2022

ISBN

978-0-88975-698-4

Citation

Ze'ev Wurman (2022). *Why Educational Testing is Necessary*. Fraser Institute. <http://www.fraserinstitute.org>.

Supporting the Fraser Institute

To learn how to support the Fraser Institute, please contact

- Development Department, Fraser Institute
  Fourth Floor, 1770 Burrard Street
  Vancouver, British Columbia, V6J 3G7 Canada
- telephone, toll-free: 1.800.665.3558 ext. 548
- e-mail: development@fraserinstitute.org
- website: <http://www.fraserinstitute.org/donate>

Purpose, funding, and independence

The Fraser Institute provides a useful public service. We report objective information about the economic and social effects of current public policies, and we offer evidence-based research and education about policy options that can improve the quality of life.

The Institute is a non-profit organization. Our activities are funded by charitable donations, unrestricted grants, ticket sales, and sponsorships from events, the licensing of products for public distribution, and the sale of publications.

All research is subject to rigorous review by external experts, and is conducted and published separately from the Institute's Board of Trustees and its donors.

The opinions expressed by authors are their own, and do not necessarily reflect those of the Institute, its Board of Trustees, its donors and supporters, or its staff. This publication in no way implies that the Fraser Institute, its trustees, or staff are in favour of, or oppose the passage of, any bill; or that they support or oppose any particular political party or candidate.

As a healthy part of public discussion among fellow citizens who desire to improve the lives of people through better public policy, the Institute welcomes evidence-focused scrutiny of the research we publish, including verification of data sources, replication of analytical methods, and intelligent debate about the practical effects of policy recommendations.

About the Fraser Institute

Our mission is to improve the quality of life for Canadians, their families, and future generations by studying, measuring, and broadly communicating the effects of government policies, entrepreneurship, and choice on their well-being.

Notre mission consiste à améliorer la qualité de vie des Canadiens et des générations à venir en étudiant, en mesurant et en diffusant les effets des politiques gouvernementales, de l'entrepreneuriat et des choix sur leur bien-être.

Peer review—validating the accuracy of our research

The Fraser Institute maintains a rigorous peer review process for its research. New research, major research projects, and substantively modified research conducted by the Fraser Institute are reviewed by experts with a recognized expertise in the topic area being addressed. Whenever possible, external review is a blind process. Updates to previously reviewed research or new editions of previously reviewed research are not reviewed unless the update includes substantive or material changes in the methodology.

The review process is overseen by the directors of the Institute's research departments who are responsible for ensuring all research published by the Institute passes through the appropriate peer review. If a dispute about the recommendations of the reviewers should arise during the Institute's peer review process, the Institute has an Editorial Advisory Board, a panel of scholars from Canada, the United States, and Europe to whom it can turn for help in resolving the dispute.

Editorial Advisory Board

Members

Past members