doi: http://dx.doi.org/10.7203/RASE.10.2.10058
Págs. 115-127

Testing and the promise of academic progress

Las pruebas estandarizadas y la promesa del progreso académico

Nelly P. Stromquist[1]

Abstract

In light of the constantly expanding use of standardized tests (STs) not only in industrialized countries but also in developing nations, this article seeks to offer some reflections on the assessment path that educational systems across the world are taking. In unpacking the STs we must not take for granted a number of aspects of standardized tests-from their construction and application to their unintended consequences. The article proceeds in four distinct but interrelated parts: (1) understanding the assumptions and process in the development of STs, (2) probing the assumptions underlying the use of STs, (3) recognizing that STs are limited in what they can tell teachers about how to improve their instructional practices, and (4) forming a broader philosophical critique about how knowledge transmission should be measured and by whom. There are multiple procedures and features that render STs an easy-to-use assessment tool for student learning. However, assumptions used test design create considerable distance between their construction and any measurement of ability to teach. This distance impedes the use of STs as practical tools to improve instruction. Nonetheless, the testing industry in the US is solid and there are global efforts underway not only to expand the application from pre-Kindergarten to secondary schools but also to the university level. The article concludes with a few challenging ideas for further debate-ideas such as the need to rely more on teachers' evaluation of their students and to prepare teachers to fulfill better their instructional and assessment functions.

Key words

Standardized testing, normal distribution curve, teacher performance, instructional practice, academic achievement, testing industry.

Resumen

A la luz del uso en constante expansión de las pruebas estandarizadas (STs) no solo en los países industrializados, sino también en los países en desarrollo, este artículo pretende ofrecer algunas reflexiones sobre la trayectoria de evaluación que los sistemas educativos de todo el mundo están tomando. La deconstrucción de las ST requiere que no se den por sentado varios aspectos de las pruebas estandarizadas, desde su construcción y aplicación hasta las consecuencias imprevistas de tales pruebas. El artículo se desarrolla en cuatro partes distintas pero interrelacionadas: (1) comprender las suposiciones y el proceso en el desarrollo de las STs, (2) sondar las suposiciones subyacentes al uso de STs, (3) reconocer lo que las STs pueden y no pueden decir a los maestros en cuanto a cómo mejorar sus prácticas de instrucción y (4) formar una crítica filosófica más amplia sobre cómo debe medirse la transmisión del conocimiento y por quién debe ser hecha. Existen múltiples procedimientos y características que hacen que las STs sean fáciles de usar como herramientas de evaluación del aprendizaje de los estudiantes. Sin embargo, los supuestos del diseño de la prueba estandarizada crean una distancia considerable entre la construcción de la prueba y la medición precisa de la capacidad del docente para enseñar. Esta distancia impide el uso de las STs como herramientas prácticas para mejorar la instrucción. Sin embargo, la industria de las pruebas en los Estados Unidos está fuertemente arraigada y hay esfuerzos globales en curso no solo para ampliar la aplicación las mismas desde los niveles de pre-jardín de infancia hasta la escuela secundaria, sino también a nivel universitario. El artículo concluye con algunas ideas desafiantes para el debate posterior: ideas como la necesidad de confiar más en la evaluación realizada por los profesores de sus estudiantes y la preparación a los profesores para que desempeñen mejor sus funciones.

Palabras clave

Pruebas estandarizadas, curva de la distribución normal, desempeño docente, prácticas de instrucción áulica, logro académico, industria ligada a la construcción de pruebas estandarizadas.

Recibido: 30-03-2017
Aceptado: 02-05-2017

1 University of Maryland, stromqui@umd.edu

1. Introduction

In the era of the «knowledge society» a growing and global interest among policymakers is to monitor the performance of public schools so that the contribution of public investment may be justified. With advances in test design, standardized tests (STs) have become ubiquitous. They range from those used merely for national purposes to those deployed for international comparisons. They also vary greatly in purpose, from whether they are used to assessing school performance to serving high-stakes decision-making, in which case student performance is used to determine teacher competence and, thus, to reward or punish individual teachers.

Reliance on STs is significant across the world, but perhaps greatest in the US, where high-stakes tests have been mandated by successive school reforms, with the greatest growth in usage occurring in the mid-1980s (Stecher, 2002). Today, STs are in their third generation. They began as part of the legislative mandates for the Elementary and Secondary Education Act (ESEA) of 1965, then continued as part of high-stakes testing with the No Child Left Behind (NCLB) Act of 2001 and its reauthorization of ESEA (Hamilton, Stecher & Klein, 2002), and now with the Every Student Succeeds Act (ESSA) of 2015. The NCLB of 2001 mandated universal testing in reading and math in grades 3 through 9.

The Council of the Great City Schools, which represents large urban districts in the US, found that students take an average of 113 standardized tests from pre-K (before Kindergarten) to the 12th grade (the last year of secondary school) (Hefling, 2015). Notwithstanding the reliance on test scores as instruments of educational policy reform, major professional associations dealing with education in some form, have expressed concern with the use of testing to reward or punish individual teachers, arguing that «No single test score can be a perfectly dependable indicator of student performance and high-stakes decisions about individuals should be based on factors other than the score on a single test» (AERA, American Psychological Association, and Council on Measurement in Education, 1999).

By 2012, the K-12 testing had become a $1.7 billion industry, an amount that will increase as tests are mandated by the Common Core Standards of ESSA, which represents a continuous strategy by the US federal government to use testing as a major policy reform mechanism (Resmovits, 2012). In the US, to comply with NCLB alone, some 68 million students were estimated to take STs in 2007 (Scherer, 2005). States have the flexibility to develop their own testing accountability system as long as they have «an appropriate feedback mechanism» (Hamilton, Stecher & Klein, 2002: xiii). The fact is that many education systems rely on STs developed by only a handful of test developers, all major national firms.

STs are unquestionably an American model of learning assessment that is attaining widespread diffusion across the world. We have today a growing array of cross-national tests: Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), the South African Consortium for Monitoring Essential Quality (SACMEQ), Progress in International Reading Literacy Study (PIRLS), Programme for the International Assessment of Adult Competencies (PIACC), Programme for the Analysis of Education Systems of the CONFEMEN (PASEC) and, more recently, efforts by the Organisation for Economic Co-operation and Development (OECD) are in place to measure student learning at the higher education level in the future through the Assessment of Higher Education Learning Outcomes (AHELO).

The growing diffusion of international large-scale assessments through STs has been compared to old colonial practices of exporting their goods as the best possible to the colonies. One observation catches this well:

> *«Thus, while in previous centuries Christian missionaries followed in the wake of the military to promote their brand of education across the European colonial empires, we now face the prospect of the West exporting its vision of schooling around the world through the auspices of cross-national tests supported by the modern missionaries and camp followers of our time: the think tanks and multinational companies who specialize in identifying and delivering "what works"»* (Morris, 2016:2).

An international institution that is heavily invested in the promotion of STs is OECD. This organization is engaged in two initiatives. The PISA for Development Initiative seeks to support the Education 2030 agenda of the Sustainable Development Goals (SDGs) and thus to increase the number of participant countries engaged in STs beyond the over 70 that participate in PISA today. The second initiative refers to OECD efforts underway to develop STs to assess student learning and institutional performance across the world at the higher education level. Its Assessment of Higher Education Learning Outcomes (AHELO) is now undergoing test development in 17 countries participating in a feasibility study to develop standardized tests in three areas: «generic skills,» economics, and engineering. OECD hopes to have the project initiated sometime in 2017. The American Council on Education and Universities Canada have expressed opposition to AHELO, stating that «the AHELO approach fundamentally misconstrues the purpose of learning outcomes, which should be to allow institutions to determine and define what they expect students will achieve and to measure whether they have been successful in doing so» (Redden, 2015).

In light of the growing presence of STs, this article proposes to engage in reflection about the value of such tests, the appropriateness of their use for different purposes, and their major consequences for the production of knowledge. To do so, the article proceeds in four parts: (1) understanding the assumptions and processes in the development of STs, (2) probing the assumptions underlying the use of STs, (3) coming to terms with what STs can and cannot tell us about how teachers might improve their practice, and (4) forming a broader philosophical critique about how knowledge acquisition should be measured and by whom. This article then concludes with a few challenging ideas for further debate. It should be noted that I am a sociologist of education, not an expert in STs. This article is based primarily on a review of English language sources, comprising books, articles, and several reports and presentations from reputable education institutions and research organizations. It is in the English-speaking world that most of the pertinent literature has been produced. I make an effort to synthesize the pertinent literature on STs and to offer some ideas of my own based on the argument that, in the same way war is too important to be left to the generals, knowledge is too important to be left to test developers and politicians.

2. Assumptions and processes in the development of STs

The design of a norm-referenced achievement test starts with the consideration of content: an examination of the curriculum and textbooks (Hamilton & Koretz, 2002). But soon other factors enter into play producing particular effects and consequences. Decisions on test design (test length, item format, content coverage) require trade-offs with respect to reliability, validity, and costs. Some of these considerations tend to take us far away from classroom realities.

There are many ways to assess the acquisition of learning: multiple choice, true/false, student portfolios, self-reporting inventories, among others. Further, the assessment can be formal (grades) or informal (Popham, 2009). STs derive their popularity from the relatively easy procedures they create: uniform administration, uniform scoring, convenience in reporting, and – with increasingly sophisticated digital computation – results highly amenable to quantitative analysis. STs are timed by design: usually, text administration is limited to one or two hours, to reveal what has been learned over a considerably longer period. Most STs are based on multiple choice (an American invention par excellence), although some include true/false responses.

Some reflections on the nature of knowledge are pertinent here. Multiple choice questions require precise answers (e.g., choosing response a, b, c, d, or e). Life, however, is not black and white and frequently engages us in situations where recognizing degrees of grayness is crucial, where a, b, c, d, or e choices are not so neatly laid out. Certain skills may not be measured with multiple-choice items (Le & Klein, 2002). I would add: certain skills *should not* be measured with multiple-choice items. It is paradoxical that STs are so prevalent in a country (the US) that places a premium – at least, discursively – on student-centered learning. Such learning calls for exploration and discovery, which means that the knowledge being developed does comprise degrees of ambiguity and contradiction. Student-centered learning implies that children are encouraged to pay attention to a variety of experiences; student-centered learning also implies that the student is a key agent in the knowledge acquisition process. So, it is the process of self-determination that should be functioning, not the forced selection of items in a multiple-choice format.

From a construction perspective, STs are based on two samples: a sample of the knowledge being tested and a sample of the population upon which the test will be normed. Regarding the first sample, a test contains only a fraction of all questions that could be asked on a given subject area (Le & Klein, 2002). In the second sample, norm-referenced means comparing the standing of the students being tested to «a larger group,» which really means a sample of students of similar age and levels of education (Hamilton, Stecher & Klein, 2002).

The logic of ST construction gives priority to the test's ability to discriminate among the test takers. It does so by employing the normal curve as the referent for the distribution of test scores. The normal distribution curve – whose precise name is normal distribution *probability* curve – assumes that the majority of people (64% to be exact) will have a satisfactory performance and that a few will have an excellent performance while another more or less equal few will have a poor performance. Basing ST scores on the normal distribution necessarily places half the students above the median score and the other half below it (McMillan, 2001). Examining this logic more closely, what the distribution of scores according to the normal curve implies is that satisfactory/complete knowledge on any discipline/area can never be acquired by all. This assumption is plausible, but an additional assumption of the normal curve is that while excellence is reserved for a few, failure is also created and assigned to one unfortunate group of students. This group is, of course, small, but nonetheless placed in the category of failure. So STs create both failure and excellence – warranted or not – by design. Further adding to the arbitrariness of test construction is the concomitant use of cut points to create categories. Hamilton and Koretz (2002:27) note, «The use of cut points is fundamentally judgmental and is often called into question.» Often, a difference of one point in ST performance can mean student placement in a different performance level. To be sure, teachers' judgments of student performance also have an irreducible element of arbitrariness, but it is the STs that claim full objectivity, while denying the considerable element of subjectivity behind the numbers.

From the perspective of test developers, these tests must be created to produce well-spread score distributions to attain high reliability indices. Popham (2001:48), a well-known US expert in instruction, evaluation, and measurement has observed that, «The more important the content, the more likely teachers are to stress it. The more that teachers stress important content, the better students will do on an item measuring that content. But the better students do on an item, the more likely it is than an item will disappear from the test.»

One may assign to a test all the good purposes one wishes but if the test has not been designed to meet those objectives, it becomes a victim of wishful thinking, and even dangerous. One reason tests cannot be used to measure teacher effectiveness or education quality is the frequent mismatch between local curriculum content and standardized test content. Popham (2001:43) remarks, based on his long experience with test development, «If you look carefully at what the items in a standardized test are actually measuring, you will often find that half or more of what's tested wasn't even supposed to be taught in a particular district or state.»

The logic of ST construction – actually, any form of testing – reduces knowledge to a few competencies. This is to be expected and does not exclude standardized testing from the array of valid assessments of knowledge. What this logic should is to invalidate any claim made by educational policy makers that standardized testing is the sole measure of student knowledge and that teachers are the sole agent responsible for student performance in tests. The dominance of the testing regime is characterized by intellectual contradictions. Despite the strong support by policymakers on the use of STs to monitor schools and learning, it has been observed that, «Many educators and educational policymakers are largely untrained in test design and validation.» (Hamilton, Stecher & Klein, 2002: xiv).

3. Assumptions about the use of STs

The ultimate purpose of testing students is to improve educational practice: Teachers will adopt more affective practices and students will be motivated to work harder (Stecher, 2002). This objective makes three assumptions of how learning occurs: (1) learning is based on what the teacher teaches, (2) all students have the same possibility of learning regardless of differences in personal context, (3) the distribution of test scores along certain subjects (e.g., math, reading) or, even better, certain specific test items, enables the teacher to figure out how to improve her instructional strategies or her content knowledge about the subject.

While by no means arguing that teachers do not make a difference in learning, we have accumulated massive research findings that indicate that family influences on learning are powerful and surpass school/teacher influences. Some observe that family factors can explain up to 70% of the observed variance in test scores. A key assumption in the use of STs is that the determinants of student performance lie primarily with the education system and that the determinants of superior performance can be isolated and policies based on these causes can be transplanted into different contexts (Morris, 2016).

For teachers to use test scores to improve their practice with some disadvantaged students, scores would have to be given at both the item level and the specific student level. Test score distribution at such a molecular level seldom happens, even though it is possible with current digital analysis capabilities.

Hamilton *et al.*, (2002) ask a critical question: how do teachers use test data to change instructional practices? In other words, how would a teacher move into corrective action given a student's poor per-

formance in a given subject? Let us assume for a moment that teachers have access to the most detailed of test performance data: If I were the teacher and saw that Jim, my fifth grade math student, did very poorly on item three of the ST for his grade, how can I tell what went wrong? Did I not explain it well to him? Did he need more time to understand it? Was my instructional approach not effective with him? And if the majority of students in the class did better than Jim on that item, should I change my practice? In other words, what information does the ST item give me to personalize my practice vis-à-vis Jim? I would argue that the best source of data to improve my teaching skills would have to come from my daily observations and my knowledge of specific students. ST data – given their relative distance from the curriculum covered in class, the time elapsed between my teaching of the relevant module and the time I got the test results, and the unspecified connection between my teaching and the test results – would make the whole exercise a difficult piece of detective work for which I, as a teacher, would have few clues or the time to piece together those classroom memories into an answer. I would join other teachers in a common response: smile, accept the test results, and shelve them.

How many teachers receive feedback on their students' performance – a meaningful feedback that enables them to trace which aspects of a specific discipline might have been better explained or treated in the classroom? Hamilton and Koretz (2002:25) remark, «If performance is what the student knows and can do, neither the difficulty of test items nor the items' ability to differentiate between high and low performing students is necessarily important.» The same authors add, «If all students show maturity, then that is what the test should reveal.» (2002:25).

McMillan (2001:132) holds that one of the most serious misuses of STs is to use them to evaluate teachers. He states:

> *«This is a misuse of the results for several reasons: (a) Standardized tests are not designed to evaluate teaching or teachers; (b) the content and skills tested will not have a perfect match with local curriculum or what individual teachers stress in the classroom; (c) each year brings a unique group of students to a teacher, with knowledge, skills, motivation, and group chemistry that may be different from other years; (d) it is difficult, if not impossible, to isolate the influence of differences between teachers (most of what students experience is common); and (e) a standardized test provides only one indication of student performance.»*

Porter (1995) and Apple (1996) were the first to observe a disjuncture between test developers/proponents and test users. Porter, a philosopher of science, has noted that in the production of STs there are two dissimilar expert groups: those who are experts in the development of tests and their quantification and those who are the experts in the content being measured. Morris makes a similar comment when he remarks that there is an «increasing detachment of policymaking from academia and the public sector.» In this view, «Those advocating reform are now directly and heavily involved in the policy-making process, primarily as the providers of independent "evidence-based research," which serves to define both the nature of policy problems and their solution.» (Morris, 2016:8). Describing this state of affairs, the Graduate Institute of International and Development Studies (2016) puts it this way: «The epistemic community of education statistics and measurement is distant from the objects of measurement.» In this view, these communities are so distant that the experts in content seldom influence the experts in test development. It is to be noted that the UNESCO Institute for Statistics (UIS) has launched the Global Alliance to Monitor Learning (GAML) initiative, which seeks to ensure «technically sound and reliable

approaches to measure learning» (Graduate Institute of International and Development Studies. 2016). It does not seem clear to what extent this initiative will bring the two epistemic communities together.

Positive teacher responses to test results include: providing more instructional time, working harder to cover more material, working more effectively (Stecher, 2002). Despite the good intentions of teachers and school administrators, it is not clear how tests can lead to more equal opportunities or outcomes for students from different backgrounds. The distribution of scores may tell us that certain minorities are doing less well than the dominant ethnic group, but to move from there to improved practices would require institutional responses – not teacher responses alone. Unfortunately, there often emerges less an institutional than an individual response. In some US states, teachers are spending large portions of class time on test preparation, particularly in urban, low-income, and high minority school districts (Stecher, 2002)[2].

Psychometricians develop tests but soon afterwards statisticians take over. The role of statisticians is to analyze the distribution of scores across national and various subnational levels and, often, to examine the correlates of test scores with various factors, such as student location (urban, rural), sex, age, and – when available – education and occupation of parents. What is constantly found is that family background significantly affects student performance and, thus, that economically and socially disadvantaged minority students persistently perform less well than wealthier students. Here, statisticians play a confirmatory, reproductive role.

At global levels today, the role of the statistician is growing in complexity and authority. Statistical experts from UN international agencies have been assigned to work on the development of indicators to measure progress in the implementation of the Sustainable Development Goals (SDGs), a global policy that will influence education in all countries between 2015 and 2030. Consequently, statisticians are spending considerable amounts of time searching for and defining indicators of «learning.» «Expert groups» have been appointed by governments and international agencies and are working hard on the identification of global indicators. Since the SDGs include education – i.e., goal number 4: «Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all» – efforts have been underway to design STs to measure learning across the world. Needless to say, the knowledge to be measured by the tests will be greatly decontextualized and subjected to standard features of ST construction, which include the identification of excellence – but also that of failure. What will come out of this? Both outcomes may be expected, but both are greatly influenced by family, community, and national circumstances that lie beyond the scope of any standardized educational testing and, therefore, beyond the capacity of such tests to judge teachers.

4. What STs can and cannot tell teachers that will improve their practice

Popham (2009) makes a distinction between assessment *of* learning and assessment *for* learning. He is much more in favor of assessment for learning. According to him, there is strong evidence that instructionally oriented assessment, *if effectively implemented*, will improve student learning. Since STs are designed to discriminate among students, to assess student performance more completely (showing more of what the student knows), other forms of learning evidence should also be allowed.

2 To see a full discussion of unintended consequences of STs on learning, see the study by Barrenechea (2010), which makes six additional critiques to standardized tests, i.e., the tension between multiple intelligence levels and the standardized form of evaluation, lack of attention to the real curriculum, risks to teaching to the test, incentives to cheat in the production of results, failure to consider student socioeconomic differences, and test limitations in predicting entrance to the labor market. Also relevant is an earlier work by Hamilton, Stecher, & Klein, 2002.

Several educators argue that STs cannot serve teachers because these tests do not have an instructional component, i.e., they are not designed to improve instruction; they are designed to determine system performance in general (accountability) (DePascale, c. 2004). Popham (2006:82) has remarked that most of the key state standardized tests are «instructionally insensitive – that is, they are unable to detect even striking instructional improvements when such improvement occurs.» This shortcoming arises in part because scores are forced to fit the normal curve (see above) and in part because the tests are so strongly linked to socioeconomic status that they measure what students bring to school rather than what they are taught there. If a student did not respond well to an item, it is not clear that this happened because the knowledge measured by that item was never taught by the teacher, badly taught, or the student learned it but subsequently forgot. Looking at aggregate data (whether national or even classroom) would not tell us this. Teachers would need to go item by item to identify response patterns.

Teachers are expected to be beneficiaries of testing, not by their own request but rather as the intention of policymakers. By having access to student standardized test results teachers are supposed to learn how to improve their teaching in terms of covering the content better, improving their instructional strategies, or both. The core assumption underlying testing is that by enabling teachers to confront the actual distribution of scores among their students, this constant monitoring will lead to constant professional development by education authorities and through the individual initiatives of the teachers themselves.

Let's examine the role of teachers in greater depth. To create a comprehensive conceptual framework to understand learning, teachers would have to be assigned a key function. To expect teacher performance that is satisfactory or higher, teachers would have to be:

- credentialed (initially trained in the details of pedagogical skills and in the subject to be taught),
- provided constant professional development, to hone their skills in the face of diverse and new socioeconomic contexts and challenges,
- well remunerated and recognized as professionals, to ensure high levels of professionalism and permanence in the educational system, and
- allowed access to an appropriate classroom environment, in terms of student class size, amount and quality of education materials, amount and quality of furniture and equipment.

In developed countries, teachers' conditions are generally satisfactory, although their training regarding «diverse and new socioeconomic contexts and challenges» could be much better. In many developing countries, the four conditions above are a remote fantasy. If many teachers operate under poor working conditions, how realistic is it to expect that learning should improve by confronting them with ST results? An analogy that comes to mind is to expect people to gain weight by weighing them regularly but without altering their diet. As noted earlier, one of the strongest findings we have gained from research on schools is the great influence of family wealth (or lack of wealth) on the academic progress of students. To neutralize this influence requires special measures and not just constant testing that ranks students and schools in highly predictable ways.

A relatively recent review of the use of STs by a prestigious commission appointed by the National Research Council in the US (Hout and Elliott, 2011) summarized contributions from economics, psychology, and educational measurement on the use of STs for educational improvement. It concluded that not only had these tests not functioned as incentives to improve student achievement but that they had actually decreased the rate of high school graduation. The study noted somberly that «despite using test-

ing for several decades, policy makers and educators do not yet know how to use test-based incentives to consistently generate positive effects of achievement and improve education.» (2011:92).

The global application of STs will bring additional difficulties. Such tests compare very different populations: some education systems include children with special needs or children of migrant workers, sometimes cities (e.g. Shanghai) are compared with countries (Morris, 2016). Moreover, rankings across countries introduce an element of unreality. Under what criteria can we expect students in developed countries that spend on average $212 annually per primary school student and about $368 annually per lower secondary student to register higher ST results than students in OECD countries, where investment per student in combined primary and lower secondary is about $9000, not to mention $11,000 in the US? (International Commission on Financing Global Education Opportunity, 2016). We should also not forget that STs are an acquired practice, which means that some students are more acquainted with such tests than others. For instance, STs are less frequent in developing countries than industrialized nations. Given the multiplicity of factors, Morris (2016:19) concludes, «The reasons underlying different levels of pupil achievement are inherently complex and explanations are conditional.» Further adding to the imprecise nature of STs are pressing logistical conditions. It has been noted that the Programme for the Analysis of Education Systems of the CONFEMEN (PASEC), which comprises 13 countries and aims to develop a common qualification framework, has a lag of two years between assessments and their publication. A similar time lag has been detected in the case of the South African Consortium for Monitoring Essential Quality (SACMEQ) (Graduate Institute of International Development Studies, 2016). Such delays likely remove test results as practical information that teachers might consider to improve their practice and situate the tests as general barometers of national education performance.

5. How knowledge should be measured and by whom

Through the widespread use of STs, we have seen a drastic reduction of knowledge transmission at the classroom level across the world as STs focus mostly on two disciplines. It is estimated that 99% of national assessments today measure math and reading, while topics linked to culture and art are covered in 36% of assessments, socio-emotional issues are treated in 18%, and physical wellbeing and learning approaches/cognition are covered is less than 1% (Anderson & Winthrop, 2016). In other words, important topics dealing with the joy and meaning of life as well as knowledge of how attain and negotiate relationships and conditions are not assessed. It is well understood that STs are not capable of measuring skills that have been learned but are not susceptible to easy quantification (Beck, 1976). It is encouraging that some knowledge and skills are *not* tested, yet it must be acknowledged that in heavy testing environments, what is not tested is simply taken out of the curriculum. In the US, a well-known critic of educational reforms laments that, despite the importance of fields such as history, science, literature, the arts, and politics, attention to them has diminished considerably (Ravitch, 2016). Echoing this concern, Gordon and Rajagopalan (2016) fear that the simultaneous measurement of decontextualized knowledge, fully deployed in STs, and the strong claim to precise quantification by measurement science may actually impede students' development of their own capacity to analyze, document, and appraise.

We live in a world of contradictions. On the one hand, we recognize that «education… should be directed to the all-round development of the human personality and to the spiritual, moral, social, culture and economic progress of the community, as well as to the inculcation of deep respect for human rights and fundamental freedoms.» (ILO/UNESCO, 1966). On the other hand, we have embarked on a

course of testing regimes that promote knowledge and skills that focus on labor market productivity, to the exclusion of more humanistic endeavors. Another significant contradiction is that, again on the one hand, there is growing evidence that improving the education quality in developing countries has not correlated with the increasing use of ST-based assessments, which has led several researchers to assert that large-scale assessments are weak tools for changing the classroom experience (Graduate Institute of International and Development Studies, 2016); yet on the other hand industrialized nations, led by the US and those in OECD, are resolutely committed to the use of STs and global comparisons.

No one disputes the authority of a university professor in grading her students. After watching student performance over a semester, we consider the professor as qualified to assess student learning. Why should we not give the same trust to teachers working at primary and secondary education levels, who spend not a semester but a year with their students? This lack of trust in teachers would be justified if the teachers are not credentialed or not given constant professional development. If this is the case, would not teacher training be the point at which to begin academic improvement efforts?

6. Conclusions

Standardized testing and the accompanying test-based accountability have taken solid root in the US since 2001 and are now integral to the evaluation of teacher performance. And, as noted, under the influence of organizations such as OECD, the march of STs across the world seems inexorable. It would be difficult to remove standardized student testing from the policy arena, but what can be done is to question their properties and thus erode their pervasive influence in determining the fate of teachers and dictating school behaviors (Hamilton, Stecher & Klein, 2002).

Administering STs is not inexpensive. In 1996, testing cost $165 million and was projected to reach $300 million in 2000 in the US (Hamilton, Stecher & Klein, 2002). By 2015, these projections had been notably exceeded when the testing industry was reported to be generating about $2 billion annually (Strauss, 2015). There are additional costs linked to test administration and related activities. The General Accounting Office (1993) estimated the value of class time required for test administration to be seven hours per student or $516 million during the academic year 1990-91 (Hamilton, Stecher, & Klein, 2002). So, why is standardized testing so prevailing and increasingly endorsed by policymakers? Several plausible reasons emerge, not necessarily mutually exclusive. One is that «Tests are cheaper than changing classroom practice though direct intervention.» (Linn, 2000:8). Asserting that tests will produce the information needed to improve the system tends to increase the government's (i.e., the backer's) legitimacy. Another possibility is that the current test regime involves so many firms and professionals that it would be difficult to erase their influence on policymakers. In 2015, the large test developers spent $20 million in lobbying activities in the US (Strauss, 2015). A third explanation is that STs have generated their own research constituency: Some advocate the use of international tests for the opportunity they open to create data that enable all sorts of related quantitative analyses. Multiple regressions that use test scores as the dependent variable dominate studies of school effectiveness and these regression models generally look for causality, which is a noble endeavor in most sociological research. Many of these analyses, unfortunately, continue to be devoid of clear conceptual frameworks to explore such causality. Yet a fourth explanation would argue that as we move into the «knowledge science,» we enter the delusion that all things ought to be – and *can* be – quantified.

The logic of STs, as discussed above, incorporates several assumptions that intervene between teacher practices and *a true* assessment of student learning outcomes. It makes sense to want to know how teachers and schools are doing, but we have to remember that there is no direct link between this and ST test scores. Teachers observe ongoing student performance from the beginning to the end of the class or academic year. Because of this, teachers know their students' strengths and weaknesses more than any ST could discover, but the trend today is not to trust teachers' judgments. Conversely, teachers do not think much of STs. A teacher survey conducted by the Teachers Network (2007), based on a nationwide sample of teachers covering all levels of schooling, probed their opinions about NCLB testing. It found that, while 37% of the teachers saw such testing as «somewhat helpful,» a larger proportion (42%) found it «not at all helpful.» About 40% of the participating teachers also stated that those tests encouraged rote drilling and eliminated curriculum materials not tested. Most damaging of all, was the teachers' assertion, with 69% strongly agreeing, that the Adequate Yearly Progress goals (a provision of NCLB and that was based on ST performance) had «contributed to teacher burnout.» ST objectives, therefore, may end up functioning as highly oppressive requirements.

Educators and policymakers alike must continue to address three crucial questions: How to encourage learning? How to encourage teacher professionalism? And, what is the relationship between measurement and quality of education? Standardized testing is incapable of measuring with any precision levels of critical thinking and high-order reasoning. Further, educators fear that STs remove, if not erase, teachers' judgments, all the more so since tests are designed by experts distant from the classroom environment (Apple, 1996).

STs can play a useful role in understanding basic trends in knowledge acquisition by successive cohorts. They can also play a role in overall monitoring of the educational system and *societal* performance. But two points should be made clear: (1) the assumption that ST results can be readily used to improve teaching practice is simply erroneous, and (2) it is possible to monitor national performance without necessarily having a regional or global comparison. We should make sure that tests cover all important elements of the domain; for that reason, it is necessary to rely on content experts, not test developers alone (Le & Klein, 2002).

Poor performance of a country's education system gives rise to expressions of alarm and, in some cases, to serious concern over teaching quality. The unintended consequences of STs, and particularly of their misuse, have been «a decline of trust in public education and increased pressure from both users and business firms to turn to the private sector partially or wholy for solutions.» (Graduate Institute of International Development Studies, 2016). In developing countries, teacher training, whether pre-service or in-service, continues to be in need of reform and improvement. Enabling teachers to become better professionals rather than expanding the use of STs – this is where our priorities should be directed.

References

American Education Research Association, American Psychological Association, & Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington: American Education Research Association, American Psychological Association, & Council on Measurement in Education.

Anderson, Kate & Winthrop, Rebecca (2016): "Building Global Consensus on Measuring Learning" in Simon McGrath and Qing Gu (eds.): *Routledge Handbook of International Education and Development*. New York: Routledge.

Apple, Michael (1996). *Cultural Politics and Education*. Buckingham: Open University Press.

Beck, Robert (1976): "Equal Time: Standardized Test Results: Answering Our Critics and Ourselves". *The English Journal*, 65 (3), 65-71.

Barrenechea, Ignacio (2010): "Evaluaciones Estandarizadas: Seis Reflexiones Críticas". *Archivos Analíticos de Políticas Educativas*, 18 (8), 1-27.

DePascale, Charles. (2003). *The Ideal Role of Large-Scale Testing in a Comprehensive Assessment System*. National Center for the Improvement of Educational Assessment.

Gordon, Edmund & Rajagopalan, Kavitha (2016). *Testing and the Learning Revolution: The Future of Assessment in Education*. New York: Palgrave Macmillan.

Graduate Institute of International and Development Studies. (June 2016). *Learning from Learning Assessments: The Politics and Policies of Attaining Quality. Roundtable Report*. Geneva: Graduate Institute of International and Development Studies.

Hamilton, Laura; Stecher, Brian & Klein, Stephen (eds.) (2002). *Making Sense of Test-Based Accountability in Education*. Santa Monica: Rand Corporation.

Hamilton, Laura & Koretz, Daniel (2002): "Tests and their Use in Test-Based Accountability Systems" in Laura Hamilton, Brian Stecher & Stephen Klein (eds.): *Making Sense of Test-Based Accountability in Education*. Santa Monica: Rand Corporation.

Hefling, K. (17 January 2015): "Do students take too many tests?" Congress to weigh the question. The Rundown (a blog of news and insights).

Hout, Michael & Elliott, Stuart (eds.) (2011). *Incentives and Test-Based Accountability in Education*. Washington: The National Academies Press.

ILO/UNESCO (1966). The ILO/UNESCO Recommendations concerning the Status of Teachers. <http://www.ilo.org/sector/Resources/sectoral-standards/WCMS_162034/lang--en/index.htm?ssSourceSiteId=global>

International Commission on Financing Global Education Opportunity. (2016). The Learning Generation: Investing in Education for a Changing World. <http.//report.educationcommision.org/report/>

Le, Vi-Nhuan & Klein, Stephen (2002): "Technical Criteria for Evaluating Tests" In Laura Hamilton, Brian Stecher, & Stephen Klein (eds.): *Making Sense of Test-Based Accountability in Education*. Santa Monica: Rand Corporation.

Linn, Robert (2000): "Assessments and accountability". *Educational Researcher*, 29 (2), 4-16.

McMillan, James H (2001). *Essential Assessment Concepts for Teachers and Administrators*. Thousand Oaks: Corwin Press.

Morris, Paul (2016). *Education policy, cross-national tests of pupil achievement, and the pursuit of world-class schooling.* [Based on an Inaugural Lecture delivered December 2015 at the UCI Institute of Education]. London: UC Institute of Education.

Popham, W. James (2009). *Instruction that Measures Up: Successful teaching in the age of accountability*. Alexandria: Association for Supervision and Curriculum Development.

Popham, W. James (2006): "All About Accountability/Assessment for Learning: An Endangered Species?". *Educational Leadership*, 63 (5), 82-83.

Popham, W. James (2001). *The Truth about Testing: An Educator's Call to Action*. Alexandria: Association for Supervision and Curriculum Development.

Porter, Theodore (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.

Ravitch, Diane (2016). *The Death and Life of the Great American School System: How Testing and Choice are Eroding Education*. New York: Basic Books.

Redden, Elizabeth. (5 June, 2015). "Objections to the OCED's AHELO", Inside Higher Ed. <https://www.insidehighered.com/quicktakes/2015/06/05/objections-oecds-ahelo>

Resmovits, Joy (29 November 2012): "School Testing in US Costs $1.7 Billion But That May Not Be Enough" [Report]. *The Huffington Post*.

Scherer, Marge (2005): "Reclaiming Testing". *Educational Leadership*, 63 (3), 9.

Stecher, Brian (2002): "Consequences of Large-Scale High-Stakes Testing on School and Classroom Practice" in In Laura Hamilton, Brian Stecher, & Stephen Klein (eds.): *Making Sense of Test-Based Accountability in Education*. Santa Monica, California: Rand Corporation.

Strauss, Valerie (March 2015): "Big education firms spend $20 million lobbying for pro-testing policies" [report]. *The Washington Post*.

Teachers Network (2007): Teachers Network Learning Institute: Survey: No Child Left Behind: What people are saying. Teachers Network. <teachersnetwork.org/tnli/survey_highlights.htm>

Nota biográfica

Nelly P. Stromquist es profesora de políticas educativas internacionales de la Facultad de Educación de la Universidad de Maryland. Especializada en temas relacionados con el género, las organizaciones dirigidas por mujeres, la educación popular y el impacto de la globalización en el profesorado, que examina desde una perspectiva crítico-política. Algunos de sus libros más recientes son: *Critiques and alternatives* (coeditado con S. Klee y J. Samoff) y *Feminist Organizations and Social Transformation in Latin America*. Fue premiada con el galardón Krestin Hesselgren al profesor invitado 2012, otorgado por el Consejo de Investigación de Suecia. Su último libro, publicado en 2017, es: *Women Teachers in Africa. Challenges and possibilities* (coeditado con S. Klee y J. Lin).