Summary INEE- World Bank Workshop

Towards a New Generation of Standardized Student Assessments

JUNE 23 and 24th 2014

Mexico City, Mexico
Unidad Cultural Jesus Silva Herzog, Fondo de Cultura Económica
Carretera Picacho-Ajusco #227

Contents

Introduction

The Mexican education system has the constitutional mandate to guarantee quality of education to children and youth. Therefore civil society must have tools like student standardized assessments to monitor and assess quality of education. However, the design of standardized assessments varies depending on the objectives bestowed upon it. For example, the assessment can either be sample or census based, it can be administered in specific grades or in all grades, and it can be administered once a year or periodically.

In general, student standardized assessments are meant to accomplish at least three objectives: (i) to inform educational authorities and society in general on student performance in a school, region, or country, for an improved accountability. This function can generate a positive dynamic between the dissemination of results, citizen participation, and the quality of public service delivery; (ii) to diagnose student misconceptions to improve pedagogic practices in the classroom; and (iii) to improve decision-making by promoting better alignment between policy makers from the federal, state, and local level.

Objectives of the Workshop

1. Support the National Institute for the Evaluation of Education's (INEE) in the development of the second generation of standardized student assessments in Mexico.
2. Share lessons learned from different countries and different viewpoints, both from academics and policymakers.
3. Analyze new trends in the design, use, and goals of standardized student assessments.

Participants

The workshop gathered decision makers, in educational policy in Latin America, U.S.A, and Europe, experts in the design and implementation of standardized student assessments from the World Bank, and renowned academics, along with authorities from the Secretariat of Public Education and the INEE in Mexico.

Format of the Workshop

The workshop had a roundtable format to facilitate discussion; each session included a moderator. Each session began with a 30-35 minute presentation by a specialist, followed by 5 minutes of commentary by the moderator, and a 40-minute open discussion between the members of the roundtable. The rest of the guests participated passively and were able to contribute with questions or comments when time allowed.

The workshop lasted two days. The first day began with presentations by representatives from INEE and the World Bank, where they discussed the objectives of the workshop, the most recent changes regarding standardized student assessments in Mexico, and a general framework that helps center the debate behind standardized testing. Four sessions followed, each addressing a country experience. The first day concluded with a session discussing the relationship between large-scale and classroom assessments. The second day began with insights on data driven decision-making strategies for school-based management. A proposed design of the second generation of standardized student assessments in Mexico was also presented and discussed. And finally the second day concluded with a session on the future of standardized student assessments in Mexico.

Session 1: Welcome, context and objectives of the workshop.

Sylvia Schmelkes, president of INEE's Board of Governors

This is a critical time for INEE, since the 2013 education reform declared its full autonomy and enacted the Institution as the main responsible for providing the national guidelines for the evaluation of education in the country, including guidelines for schools principals', teachers and. In this context, INEE needs to comply with several roles including the accountability role but more importantly the role of providing information to improve education systems as a whole, as well as, each of its parts. In this context INEE is particularly interested on addressing the perverse incentives created by the design of previous large-scale student assessments in Mexico (ENLACE).

The Institute's main goal is to improve learning. To achieve educational goals it is necessary to develop public policies to improve the curriculum and reduce the learning gaps among students. In this sense, the results from student assessments should be used in schools to develop improvements plans that tackle the learning challenges of their students.

Mexico is one of the few countries in the world that, until last year, implemented a census based student assessment, in all grades (starting with 3rd grade of primary school) and in an annual basis through the ENLACE test. However ENLACE had several negative results that lead to its interruption. To give some examples: the results of ENLACE were linked to teacher's bonuses leading to several perverse effects such as teaching to the test, fraud and corruption (letting students copy or cheat or dictate results to them). Also, the census based tests contained limited curricular content, thus making teaching to the test even less desirable.

In the current process of defining a second generation of standardized tests, we would like to learn from our previous experience, and better understand international experiences on what has worked and how. The main questions we are addressing are: can sample tests result to be more desirable than census type tests? Are international experiences in measuring value added viable and cost-effective? Is there a way to control cultural background and avoid stigmatizing marginalized schools as a result of poor learning outcomes? How can the most common perverse effects be prevented (teaching to the test, cheating, reducing the areas of the curriculum that "matter" to those included in the test, among others)? How can these systems be used to improve education at the school and at the classroom level?

Reema Nayar, Education Manager for Latin America and the Caribbean, World Bank

There are many lessons from international experience. The World Bank has engaged in student assessments in low and middle-income countries. It is useful to work with a framework that illustrates the key elements to consider over when deciding what would make sense for Mexico in the design of the next generation of student assessments.

Marguerite Clarke, Senior Education Specialist, World Bank

What is a student assessment? It is an active process of gathering information, and about making decisions. One theory of change, taken from the US standards-based learning model, suggests that clear expectations for students and schools, will strengthen accountability and will motivate everyone to do a better job. This in turn, will help inform professional development, teaching, and ultimately increase levels of learning. This theory of change may or may not hold up in reality, but is one alternative.

Types of assessments.

1. Classroom assessment (formative)- When done well it can have a significant effect on learning outcomes (Bennett, Black and William). More work needed to define and isolate characteristics that lead to improved learning.
2. Examinations – (high stakes assessments assessing individual students) Most important for students and their parents (for instance they can affect college acceptance rates). Studies have shown that curriculum based exams do promote better learning. High stake examinations can have negative impacts for students, in contexts where there are limited alternatives for those failing to meet the requirements.
3. Large scale assessment - Simply reporting information about average school scores can lead to improved student performance. There is a weak positive link between accountability, uses of data, and better learning outcomes. However, there is much to learn about optimal design for accountability models.

A study in the State of Colima in Mexico suggests that softer use of results from ENLACE, the national standardized test, can produce better learning outcomes and therefore, this kind of approach should perhaps be used more often than punitive approaches. As part of the intervention the results from ENLACE where used to identify lower performing schools and provide them with greater support, while ensuring that teachers know students' scores and their learning gaps.

Countries and different approaches

FINLAND has classroom-based formative assessments, and a limited use of high stakes examinations (only one at the end of secondary school); and a regular schedule of low stakes large-scale assessment activities for system monitoring. A key constraint for adopting this model is quality of teachers: it is essential to have good quality motivated teachers.

CHINA has high-stakes examinations, and an emergent focus on classroom assessment, and large scale assessments. This model fits the context of a large number of students and scarce amount of resources.

USA has large-scale assessments (NAEP) and has less emphasis on classroom assessment and examinations.

Ideally we would like to see a system that has good quality classroom assessments, but there are also important constraints to this model.

Student assessment purposes. Different assessment designs meet different information needs and serve different purposes. For instance they could help improve teaching and learning or introduce high-stakes for individual students (large-scale).

Quality drivers. Whichever type of assessment chosen, it is less important than the quality drivers behind the design and implementation of the assessments. It is important to consider:

- Enabling context - leadership, policies, institutions, human capital and fiscal constraints.
- System alignment - learning goals, curricula, opportunities.
- Assessment quality - design, administration, etc.

Key trends and issues.

- Lines are blurring between classroom assessment and external assessments (technology helps).
- Blended cognitive and psychometric models.
- More focus on measuring 21 century skills = cognitive and non-cognitive.
- Key role of technology.
- Key role of teachers.
- Accountability uses of assessment data.

Session 2: The Brazilian experience: standardized student assessments as a tool for the application of the Basic Education Performance Index (IDEB). Francisco Soares.

The IDEB was created to generate standards, and measure quality in the Brazilian education system.

Features in Testing:
Content: Brazil does not have a unique national curriculum. However, national exams measure the common national core in Language, Math, Science, Social Studies, and Written Language. Children are expected to respond to 45 items and to write an essay.

Frequency: Tests are carried out every two years at a national level.

Quality in Education: It is important to reflect on the quality in education. There are two dimensions of quality in education: learning process and learning attainment. In a broader sense, quality in education allows students to be active citizens. Differences among social groups should be small since excellence of a few cannot represent a true sense of quality.

In terms of their trajectory, in Brazil there is an annual student census for basic education. Since 2007 Brazil can construct an individual trajectory or learning progression for each student, and there is a follow up at an individual basis.

It is a priority to reduce inequality. We would like to have learning attainment with high means and less variability. Inequality is very important in Brazil. Before, inequality was something studied by the social scientists, but now it is very important in different realms including the government.

It is important to strengthen pedagogical research. The pedagogic interventions to teach students in some cases are not resulting to be adequate, because students are not learning. Before we did not count with data but now we do. There is very limited expertise in pedagogy. We need to talk more about it and about the curricula, and of the education system.

Why is the IDEB so important? IDEB has helped assign a goal for the system and a goal for each school. There are goals for 2021, and by 2095 all states of the country should have the same IDEB. Inequality can be eliminated in one century. IDEB managed to put learning at the center of the education policy debate. The new national plan in education selected IDEB as an indicator that should be constantly monitored for the improvement of basic education in Brazil. It's an indicator for accountability.

There are many critics. Some are in the position to eliminate the index, and they propose that resources should be transferred from evaluation to those required to improve quality in education. Another group can be called the 'new reactionaries' and they believe that measuring results in education is not enough to improve results. This is a right wing critique, the government supports IDEB and it's a left wing government. We have to learn to be aware of these opposing views.

Evaluation is something much broader than just measuring performance. We give society figures that mean something, and society and other groups can take them into account to focus on progress.

The right to education must be something measureable along with learning attainment, school condition and teaching quality. It is not enough to measure students' results. In addition, results should be inclusive, broad, and relevant. There is no point in measuring without monitoring, and without the proper follow up improvement policies.

People should be guaranteed the right to education. What does this mean? For example, should children be required to learn, or should they be able to distinguish a fact from an opinion, the latter is difficult without a curriculum. Results should be used within a context, with indicators describing students' and schools' socioeconomic status, level of teacher training, and spending per student. Currently we see these indicators separately, but we never see the complete picture in education.

We have to avoid having figures that do not mean anything. We have to show teachers what the figures mean. We have to give them tools so that they can link results from the tests to the lessons in the classroom. We have to highlight the need to bring rigor into test making, from a pedagogic point of view.

To promote the understanding of large scale testing, there should be item mapping associated to the curricula and to empirical testing. Currently we are not spending enough on this. It is not clear what is going to happen to the indexes in Brazil. In an overall balance, we are better than before. Currently we can get a sense of the levels of learning of a particular school without having to visit it. We are better with the current system than without it.

Discussion

Bernando Naranjo

Question: What is the impact of high stakes testing vs. low stake testing, based on Brazil´s experience?
Answer: ENEM the high stakes test is the most visible evaluation system, almost 8 million students taking the test. People from the general public talk more about ENEM than IDEB, but the results from IDEB have a greater effect in the education policy debate.
Question: What mechanisms are in place to avoid cheating?
Answers: There are quality control mechanisms to avoid corruption; external agencies are involved in implementing the test. Every exam has the name of the student printed to give better control.
Question: What are the reactions from the results?

Answer: There is not much political opposition as compared with the technical and academic opposition.

Rafael de Hoyos

Question: What types of measures are taken to incentivize better results? Can policy makers from states or localities with IDEB gains be rewarded through greater public funding?
Answer: The connection that associates IDEB and socio economic status in not an easy one. The correlation between IDEB and socio economic status is very high, almost 0.6, therefore IDEB cannot be used in this sense.

Eduardo Backhoff

Question: Why is the test implemented every two years?
Answer: Two years were taken into account because the older system only took into account testing every two years. There isn´t a greater part to that answer. It’s not as good to have the exam every year because it would be very intrusive to schools.

Sylvia Schmelkes

Question: How to choose the common core curriculum?
Answer: there is a wide diversity in Brazil because there isn´t a unique national curricula, therefore there is no way of associating these tests with a particular curricula. The nice thing about Mexico is that you do have a common curriculum.
Question: How are standardized tests developed in Brazil, in face of enormous cultural diversity?
Answer: A lot of test items are used (over 160), and these cover the common core. This can help connect what happens in the classroom with the common core.

Gilberto Guevara

Comment: Evaluation should be linked to teaching and to the right to education.

David Calderon

Comment: We have to ask ourselves what are we doing to guarantee the right to education of each student.

Session 3: The Chilean experience. Improved results, greater accountability, and greater autonomy in school management. Lorena Meckes.

Standardized testing in Chile is a result of an evolving process. Currently evaluation consists of assessing skills and knowledge related to a curriculum, in different grades, or domains of knowledge, for every student in the country. The government allows for greater levels of school-based management for schools that achieve good results in the SIMCE.

Context

There are more than 12 thousand schools, 3.5 million students. Students receive funding and school financing depends on the number of students in the school. Schools with more vulnerable students receive more resources.

There are three types of schools: public schools fully funded with government expenditures; private schools that receive government funds but are privately owned and managed; and private schools that are privately funded, managed and owned. Municipal schools offer free education and cannot choose students. Private schools can select students. Results from national standardized tests are published at school level and this modifies the incentives.

Institutional context

There are three institutions that are in charge of quality assurance of education in Chile:
The Education Ministry, they design the curriculum, the standards, and the supporting policies. The National Education Council approves the proposals from the Minister and reviews them with experts; their role is to approve the curricula, the standards, and the selection method in schools. The Education Quality Agency carries out the SIMCE evaluation and any other international assessments; it classifies schools and carries out monitoring visits to assess the processes, the quality of teaching and parent participation.

The characteristics of the evaluation system in Chile: SIMCE

The evaluation system started in 1988 since then, there has been a lot of learning by doing and experimenting on diverse strategies. SIMCE assesses learning attainment placing education quality in the center. SIMCE is designed to align with curriculum and standards. The tests are designed to cover the entire curriculum. It focuses on language, communication skills, mathematics, science, history, geography, social studies, english, P.E. (physical education) and ICTs. Chile has a combination of census and sample based assessments. The following grades are evaluated: 2, 4, 6, 8, 10, and 11. There are teacher, student, and parent questionnaires in order to better interpret the test's results. Tests are applied either every one or two years. Results allow for value added measures of learning.

External agents apply SIMCE, and these are hired by the Ministry. Student's individual scores are not public, only scores at the school level are public. Tests incorporate the use of categories, and the socio-economic background of the student is considered. Tests are adapted for students with disability (for the blind and the deaf). During the first years of application of SIMCE, results were stagnant and there were many questions regarding the use of testing in Chile. From 2000 to 2009 the results have improved and this same pattern is evident in TIMSS and other international tests.

A lot of information has been gathered regarding the large discrepancies in test scores among students with different socio-economic background, (students in upper grades present a wider gap) with this gap reducing over time. Public information on the gap has helped inform and promote the debate about quality and equality in education in Chile.

Background of SIMCE

SIMCE took the role of a monitoring system for quality control from ‘the distance’. Test results measured school performance, and it guided parents’ decision of which school was best for his/her child. In addition, results allowed citizens to monitor compliance with the individual right to education and to reduce inequality.

- 1988-1990 SIMCE took off with the decentralization and privatization of education.
- 1995- SIMCE started published the results.
- 2000- curriculum was re-designed so that it was aligned to SIMCE
- 2003- there was rethinking of SIMCE´s purpose to use the results for pedagogic purpose.
- 2008-2014 - accountability measures were strengthened.
- 2008 - *Ley de Subvención preferencial* was introduced for vulnerable students. Schools are not allowed to choose which students to keep, and they should meet the goals of SIMCE, and carry out an improvement program.
- 2005-2009 - Tools were developed to analyze the test results.
- 2011 - *Ley de aseguramiento de calidad* was introduced to deal with the overburden of testing.

Currently there is no connection between external testing and the tests applied in the classroom or those applied by the teachers, even if they do align with the standards and the curriculum. With SIMCE there is more access to information, there are special reports that are issued on a regular basis for teachers and principals informing them about learning levels at different grades.

Discussion

Teresa Bracho

Question: What happens to schools that do not reach the learning levels that are written in the law? What happens to schools that do manage to improve learning levels?
Answer: The law defines different categories, schools can initially present an improvement program and immediately they can get the resources. The action plans of schools with low and regular performance are monitored, as they can receive greater pedagogic support, and more supervision. They receive initially a third of the resources and an action plan is approved so that they can receive the rest of the economic support. It is commonly the case that good performing schools improve their performance even more. The low performing schools can see their permit to function withdrawn if they do not improve within a certain period of time (although in practice this extreme measure rarely happens). Protests now around SIMCE are directed to reengineer

the closing of schools. Pedagogic support must be of good quality. Universities not always prepare teachers adequately, and this lowers the opportunities for improvement.
There is a contradiction inherent in the Chilean idiosyncrasies. This contradiction has to do with the measuring abuse, the over burden of testing, and the narrowing of teaching, and there is limited use of the results to inform better decision making by parents.
The weakest link in the Chilean evaluation system is testing student learning. There is a good opportunity in improving the evaluation capacity of teachers. New tests are developed but they do not necessarily uncover the learning processes of students.

Session 4: The Denmark Experience. Metrics for a better public administration. Jakob Wandall

Summary of Presentation:

In general the differences between the Danish and the Mexican context are striking in almost all terms: geographic, demographic, climatic, socioeconomic, cultural, etc. In the World Value Survey, students from Scandinavian countries value independence more than other things, while Latin American students value obedience. It is not surprising that the Danish education system is very different from Mexican. The Danish education system is small with about 60,000 students and 600,00 teachers in 2,000 public schools run by local governments. All Danish students have a computer connected to the internet. There are significant differences between municipalities and how they manage their schools.

The objectives of the Danish education system have not changed much since 1814 and they continue to aim at educating good citizens who are motivated and productive to coexist in a democratic system. Achieving these goals means enabling cognitive development of students.

Before Denmark's participation in the Program for International Student Assessment (PISA), society, parents, teachers, students and education policy makers in general believed that Denmark had a good education system. However, after the publication of the first PISA results, Denmark entered a stage of 'crisis' due to the low results achieved in the international test.

The recommendations of the OECD (2004) resulted in a mandate for the creation of a national evaluation system and for making final examinations compulsory for all schools in 2006. This measure started to develop a new evaluation culture. Since then, Denmark has focused on the design and implementation of an evaluation system of academic achievement based on an adaptive model. The system took around 4 years to be developed and was first implemented in 2010 (with an investment of approximately US$20 million).

The National System of Student Assessment is comprised of 10 mandatory tests: Reading for grades 2, 4, 6 and 8; Math for grades 3 and 6, English for grade 7; Physic, Geography and Biology for grade 8. Each test has three content areas. All students respond the tests on their computers (and on-line). The software used to implement the tests allows getting students' information quickly and reliably. It is also very easy to generate tables and graphs for each student, and to understand where each is compared to the rest of their peers. Tests do not have direct implications for any school actors (or for teachers) consequences.

Although the test results can be disaggregated by student and by item, teachers are just learning how to use test results to improve student learning in the areas where they face greater difficulties. After several consecutive years of implementation the results from the adaptive standardized test are more useful.

Features of the National Tests: detailed design of items, confidentiality of results are (only teachers can know the results of their students); administration by teachers, under the guidance of the Ministry of Education. Results are not entirely comparable at the individual level, but are highly reliable when used at an aggregated level. Its purpose is to evaluate each student and the system, not schools or teachers.

The adaptive nature of the assessment allows for a targeted and differentiated instruction according to the specific needs of each student, especially caters to the differing levels of students' academic achievement (that can be 6 to 8 years different for the same age group). The information provided by the electronic evaluation system permits comparing students with their peers nationally (the system has three types of charts / graphs to display information in percentiles.)

In addition from 2006, final exams focus on the evaluation of learning acquired during the school year. The results are public and have moderate consequences. These exams have a greater variety of items, they are multiple choice, there is an oral component, and they have open questions and essay writing.

Main lessons learned

Results should be used formatively by teachers in order to correct their students' flaws. It is therefore essential that teachers understand how to interpret and use the information from national standardized tests. Furthermore, teachers should be convinced of the importance of standardized tests to improve the learning of their students.

More research is needed with regards to teaching practice to help explain what drives a positive impact on learning outcomes. Research is well advanced in the design of standardized tests, but more research is needed to ensure improved learning. Teachers should combine data provided by standardized tests, with pedagogical practices to make a difference in the classroom.

It is important to supplement the information in schools with information and assessments outside schools. There is plenty to do from ages 0 to 6, before students begin primary and present standardized tests. Finally, adaptive testing can also be used as a tool for learning.

It is a good idea to separate the formative and summative evaluations and provide teachers with the opportunity to access information on their students test scores while allowing the Ministry of Education access to information at the locality level (not at the school or student levels).

Discussion

Rafael de Hoyos.

Comment: It is interesting that despite the differences between Mexico and Denmark one of the biggest challenges for both countries is to convince and train stakeholders (especially teachers) of the importance of using data to support weaker students as detected by the tests. It seems essential to train teachers to use the test findings (item by item analysis) to address the needs identified by standardized tests and to use the test results. Bernardo Naranjo has conducted some efforts in this respect with ENLACE results in Mexico.

Question: If the objective of the Danish educational system is to train better citizens, why was it decided to assess cognitive aspect through standardized tests? Are there no assessments to measure 'citizenship' and what it entails (eg socio-emotional skills)?

Answer: The Danish education system is under significant political pressure due tothe PISA results. That is why efforts have focused on measuring the cognitive ability of the students, as a means to achieve the goal of education.

Daniel Koretz

Comment: The standard deviations in the levels of student learning are very similar in different countries and contexts; most students achieve similar levels of learning but a few are well below or well above the average. It is important to be able to acknowledge differences and adapt measures for students with disabilities; however, the US is far from a system of students' adaptive evaluation. In the United States all students of the same age/grade take the same test, regardless of their strengths and weaknesses (including those with learning disabilities). Therefore the system cannot truly measure what students with learning disabilities have or have not learned/achieved, since usually they are not able to answer the exams at their grade level.

Marguerite Clarke

Question: Can you explain more about the profile of teachers in Denmark? Teachers seem to have much more autonomy and power than in other countries like Mexico?

Answer: Teachers are one of the main differences between Finland and Denmark. Students that want to enter the teaching profession are not the best students in Denmark (salaries of the teaching profession are lower than for other professions). However, they go through four years of college where they receive a solid foundation.

Question: Where did the leadership to improve education, through a national system of evaluation, come from? From a political platform, academics/researchers, parents? What can change the dynamics of school autonomy to create a culture to improve learning outcomes?

Answer: In Denmark we believe that the best way to help schools is by giving them the freedom to act according to their specific needs. Accountability of teachers, principals and schools is not based on the Danish national system of evaluation; rather this occurs through the democratic culture that exists within the schools. For example, directors are elected directly by schools and teachers can be fired by the same teachers. Parents are also highly involved in the education of children and demand results. In addition parents have the right to dismiss teachers if they do not fulfill their responsibility. Schools must also submit reports on the quality of education they provide to their students.

Bernardo Naranjo

Question: Why are the results of the National Tests in Denmark not public, and only for the exclusive use of the teacher? Is there any pressure to make these results public?

Answer: No, the results of national standardized tests are considered necessary for teachers and classrooms, not for accountability purposes. Thus, the evaluation is intended to be useful for teachers to identify the needs of their students, but not as a tool of accountability for them or schools. However, it is important to note that the results of the final exams in Denmark are published.

Carlos Mancera

Question: What differences have you noticed after the implementation of the national adaptive and confidential system of standardized tests in Denmark?

Answer: This evaluation system does not exist anywhere else (only in Norway) and we think it can work for Denmark. Although we have not yet seen significant changes in the PISA results, we believe that this is because teachers have failed to manage the information provided by the standardized tests to meet the needs of their students in specific areas of the curriculum. The tests have only been applied since 2010 and teachers need to further familiarize themselves with the results and learn to use them better. In addition, with accumulated years of application of the test, the information becomes richer and more useful for teachers. It is important to improve classroom diagnosis and to identify pedagogical practices that are more efficient to improve the learning outcomes of students.

Sylvia Schmelkes

Question: How can teachers best use the results from summative tests, while developing their own formative assessment skills?

Answer: Teachers should be trained in teaching techniques and methodologies including formative assessment of students by the relevant authorities (though initial teacher training and continuous professional development provided usually by the Ministry of Education). Formative evaluation should be complementary to summative rather than oppose to each other, or seem one as the substitute for the other. The important thing is to ensure that teachers use standardized testing information correctly through practice.

Session 5: The USA Experience. A national policy interpeteded and implemented differently at each state. Daniel Koretz.

Problem statement

Mexico is not unique. There is a growing interest worldwide in using large-scale assessments to improve education. In this respect the US has substantial experience with high stakes testing that has led to a wealth of research that can provide useful guidance for better evaluation systems elsewhere and in particular in Mexico.

Key Aspects

Policies using test-based accountability to improve instruction have been used since the 1970's at the state level in the US. Since the introduction of the Federal Law 'No child left behind' in 2002 all states needed to administer a test every year.

Different kinds of assessments have been tried in different states from performance assessments; standards based assessments, current/raw scores (status) to value added mechanisms and different types of accountability (sanctions and incentives). The commonality around all these different ways of assessment is that American teachers feel pressure to raise students' scores. Performance targets for student assessments have been set arbitrary, uniform and often at a very high level, introducing distortions in the system.

What we know about high-stakes testing

The effects of high-stakes testing on educational practice are mixed. These kinds of tests have contributed to moderate improvements in students learning (although there is a weak correlation and effects that vary across contexts); but high-stakes testing has also created undesirable effects such as teaching to the test, inflation of marks, cheating and other "gaming" behaviors. For instance, tests scores can become very seriously inflated going up 3 to 6 times more than they should be (e.g. in NY State grade 8 math score gains were 6 times bigger in the state's high-stakes test that in National Assessment of Educational Progress (NAEP); and with much larger exaggeration among blacks, showing an apparent narrowing of gaps between races).

This is not unique to education, literature in health care shows similar results, particularly when a numerical target is imposed; the distortion of results grows while making the overall improvement badly exaggerated. The inflation of results is a problem since relative effectiveness is estimated incorrectly; teachers, schools and systems are ranked incorrectly. Therefore, high-stakes student testing can contribute to an incorrect and unfair ranking of teachers.

What we do not know about high-stakes testing.

After more than 40 years of students testing in the US we still don't know: what are the net effects on student achievement (not of "gaming" strategies); and despite some evidence of small improvements in math scores there is no evidence of improvement in other subjects. Moreover, there has been a weak, non-systematic research design to assess the impact of student assessment and we still don't know what really works to improve students learning.

The sampling principle of testing

Tests are only a small sample of behavior. The inflation of the results happens for many reasons. First, tests are predictable. For instance, in math we have been using the same 42 test items for 10 years. Tests only work if the test sample of items can accurately infer what students have learned overall. The problem is that tests usually don't take a random sample of items from a large and relevant 'item bank' that corresponds to what all students should know. In most occasions, the content that can be most easily tested ends up being emphasized in standardized tests relative to other contents that are harder to test/measure, which tend to get de-emphasized.

In addition, many education goals cannot be tested through standardized exams. Also partly for technical reasons and for ensuring standardization and comparability of the tests from year to year, items tend to be repeated constantly. In addition, writing good items is complicated. The consequences of incomplete item sampling with low pressure are: systematically incomplete evaluation of education (tests not sufficient); modest effects on scores, measurement error, fluctuations in scores, and differences in results among tests. Once pressure (accountability) is added the consequences are: distorted (very large) effects, coaching (focusing instruction on presentation, rubrics, and incidental test content), narrowed instruction to focus only on the content that is emphasized on the tests at the expense of other relevant content, additional time dedicated to test preparation (focus on specific items) and cheating.

So what to do to get less of these negative effects of high-stakes standardized test in the US (since we cannot avoid them all): (i) anticipate score inflation, understand that scores are meaningful only if they generalize to the domain; (ii) make tests broader and less predictable. What we want is students that leave schools able to work or continue with their education, and tests that can help to achieve this if results reflect what student really know and are able to do. However, tests usually show predictable emphases, omissions and forms of representation over time (some intentional, for technical reasons and some accidental or to save time and money).

Recommendations. What we have learned from the US experience.

Recomnedation 1. Couple evaluation and accountability with training and support, particularly for teachers.

Recommendation 2. Evaluate other outcomes (make student assessment broader). Evaluate practices/processes as well as outcomes. May need to use subjective measures with objective measures (human judgment).

Recommendation 3. Use summative tests appropriately. Set realistic targets, less incentives for teachers to distort results, report in scale scores not performance standards (i.e. insufficient, regular, excellent, etc).

Recommendation 4. Design test for accountability uses. Avoid excessive narrowing and predictability in summative tests, design formative tests differently from summative tests and to serve formative purpose (if formative looks like summative you are preparing for summative test not for students learning).

Recommendation 5. Evaluate the evaluation and accountability system. There is no proven system for doing all right. All accountability systems cause some undesirable responses. Scores improvements overtime are not enough to indicate success; it is essential to monitor the effects of the accountability systems. Avoid very expensive and burdensome systems without a careful monitoring and evaluation of these systems.

Discussion

A.J. Vischer.
Question: Would you still recommend going for a high-stakes student standardized assessment system? Does the problem have to do with high-stakes or the evaluation system in itself? Wouldn't it be better only to focus on improving teacher practices?
Answer: For the US, accountability is different in different contexts but overall it is too high. In India, for example, just getting teachers to show up and prepare students for a test 'teach to test' may be a good option. Maybe there is more room in Mexico than in the US for balancing the accountability aspects of national student assessments. The ideal would be to have tests that provide teachers with more information about what their students are learning. It is true that in the US there are lots of individuals who should not be teaching, but there is very little evidence that shows that high stake tests has helped improved students learning.

David Calderon

Question: If the high-stakes are removed from the standardized tests, will problems continue to exist? In Mexico we all acknowledge that ENLACE should have not be linked to teachers incentives, but if this is removed, many of the problems with cheating and results inflation might be reduced substantially? Also, is there any hope to improve the assessment systems to make them reflect students learning? For example Colombia has started to explore options to test similar content in very different ways, to make tests less predictable. Finally, is there evidence that monitoring students' progress with other tools, besides standardized testing, can improve learning?
Answer: There is a long history of high stakes exams in the US and European countries. To avoid perverse incentives, some tests have recurred to open questions/essays. Nevertheless, students learned what was expected, and teachers knew what the tests would look like, which could continue the pattern of teaching and learning to the test; so essays will not solve this problem. In addition, the assessment issue is also a technical issue. To make a test comparable overtime items should be quite similar. In the US, Educational Testing Services (ETS) has done some work to make tests as comparable as possible without making them so predictable. Another option is to use different tests for monitoring vs. accountability.

Francisco Soares

Comment: We can improve tests to some degree, but not necessarily to the extent that the problems will go away. The key is the design of the accountability system. More than tests, reasonable pressure and rewards are essential.

Daniel Koretz

Comment: We also need to reward other important things outside the tests (Arne Duncan). We need to "observe fourth grade teachers" and see how a good teacher looks like in the classroom. It is amazing to see a math lesson where students are competing to explain the errors in their reasoning. These things do not usually show up on a grade 4 test. The question then is what will you do to reward this excellent teacher and how will you support others to learn from him or her?

Eduardo Backoff

Comment: Although NAEP performance agreements created a lot of controversy, NAEP census assessment works because it is not a high-stakes test, and it has not been tied to accountability measures. It also provides a state report but it doesn't publish the results at the school level. Although the NAEP test system has allowed for a broader approach in testing, it is hard to evaluate which schools have improved over time vs. those that haven´t. An option might be to create a test that allows for comparability over time, while implementing the other summative assessment. However, if schools are held accountable for crossing a threshold, teachers will tend to focus on students that are near the threshold.

Daniel Hernandez

Comment: Tests can be useful to bring all students on the same page as long as they are not corrupted.

Answer: In the US the fact is that we have done tests for decades but this has not worked as well. So the main question is how do we modify testing to make them do what we want them to do?

Marguerite Clarke

Question: What is your view on NAEP 12 grade? As you mentioned test 4th and 8th are fine and have not been corrupted, because these are not high-stakes test. But grade 12 NAEP is quite unreliable because of the same reason: students and teachers could not care less about it. Where is the sweet spot between high and low-stakes tests?

Answer: In the US it is hard to get 12 graders in the last semester to even show up to school because they have already completed their exams to enter university, and so NAEP grade 12 is not that useful or reliable; but NAEP for previous years is. However, it is true that lack of motivation of students might also affect their results.

Session 6: Navigating between large-scale assessment and classroom assessment. Mark Wilson.

The presentation seeks to understand the linkages between classroom assessment and large scale assessment, and will discuss the current challenges in the US context, as well as possible solutions. In particular the idea of learning progression is introduced and a three part strategy is given as a potential solution.

Context

There is a key difference between large scale assessments and classroom assessments; the latter has the intention of modifying teaching and learning activities. However, there is a tendency to use large scale assessments as a diagnosis of teachers' practices in the classroom, which end up influencing the curriculum and instruction, and creates a vicious triangle of learning (where curriculum affects assessment and instruction). The end result is that teaching becomes very narrow, and there is evidence of teacher burn-out.

Solution

A solution considers a formative approach towards teaching, where theories of learning affect the linkages between curriculum, instruction and assessment. This will enable a long-term view of student growth, and it relates to learning progression. This requires a debate about what learning progressions are and how are they related to the curriculum, and it relates to a perspective of how learning takes place in the classroom. This approach disentangles the process of learning by studying the different components of learning progressions: construct maps, redesigned curriculum, and conjecture-based class discussion. This perspective covers more than one school year and it involves a multidisciplinary collaboration among teachers. Each subject area should only have a few learning progressions a year.

Furthermore, a three-part strategy is outlined to connect the learning progression approach with assessment design. First, build classroom assessments involving one construct at a time. Second assemble constructs into learning progressions; and third build large-scale assessments by having a structured sample of learning progressions. The results will allow experts to identify knot points in learning progressions, and help teachers identify which constructs typically depend on others in order to gain knowledge on a specific domain. Finally, with this approach teachers will be able to know where each student is in the learning progression, and he or she can help the student go further by focusing on bridging specific constructs.

Conclusions

Learning progressions can provide a way to have a healthy relationship between large scale assessments and classroom assessments. However, this requires outcome definition of learning progression (which can be heterogeneous among students), and a construct map. To reconcile large scale with classroom assessments it is necessary to rethink blue prints for large scale assessments. Challenges involve having a curriculum debate, and converting it into testable hypothesis so that experts can test alternative hypothesis of structuring curricula. This opens a whole new research agenda on learning progressions, test alternatives, and alternative outcome progressions.

Discussion

Gilberto Guevara Niebla.

Comment: It is important to think about the right mix between rudimentary and sophisticated tools to support the development of competencies of basic skills.

Daniel Hernandez Franco.
Comment: This is a good framework to help teachers learn how to evaluate, and how to establish instruments that can help assess skills in a group of students.

Marguerite Clark.
Question: How to balance the science behind learning progression and the experience the learner in the classroom?
Answer: Classroom experience is built around argumentation/debates within small groups. The emphasis of argumentation is aside from the learning progression approach towards teaching. The latter focuses more on specific content and outcome progression.

Daniel Koretz.

Comment: I am interested in the debate between specialists that think that teaching has to be done in a certain order, vs. the idea that students can follow alternate progressions. Mark`s proposal is in a middle ground, suggesting that there are some constructs that are dependent from others but there is room to test which patterns are most successful in constructing learning progressions. There might be differences in the types of subjects, statistics vs. math, etc. and differentiating among topics and the importance of sequencing might be helpful to look at.

Teresa Bracho Gonzalez.

Question: The curriculum debate is an important one. How can a construct map be created in a way that can help design the entire curriculum? Another question, how can adaptive evaluation help in the learning progression approach?
Answer: An order was already tested, but one can certainly play around with the effect on changing the order. Adaptive evaluation might be part of the proposal (multidimensional adaptive) to test where the student lies in a particular learning progression.

A.J. Visscher

Question: Do you think teachers should know all the learning progressions children should follow in a particular year? In the Netherlands, teachers study around 100 hours to learn the

material needed to teach. Can it be feasible to expect teachers to learn all the constructs needed?
Answer: It would be much easier to know the constructs needed in a learning progression with the use of a construct map. Teachers should be seen as apprentices, and not expected to be experienced teachers, and gain experience through better exposure with the material.

Lorena Meckes

Question: In Chile, learning progressions have been tried in order to trace student progression through different constructs. However, a lot of the teaching quality depends on the discussion, questions, and prompts made in class. How has this been dealt with in the previous experience? Can learning progressions be a good way to design large scale vertical scaling? And can they give information to schools about progress in a domain of knowledge?
Answer: To solve the problem it is important to test parts of the curriculum that have been proven to work. About vertical scaling, you might want classroom assessment and large scale assessment to look very different with varying number of dimensions. First solve the logical problem about content, before solving the technical issue.

Eduardo Backhoff

Question: It is good to see that there is a useful linkage between classroom and large scale assessments. Can there be a unique learning progression, and if there is, how do you know which one is it. Cognitive map and construct map should be linked to assessments, given that a cognitive map should be related to empirical testing.
Answer: There is an iterative relationship between conceptual, cognitive maps and empirical testing. A unique learning progression might sometimes be inevitable, given that certain concepts depend on others, but there might be other circumstances where there are many ways of learning. In the case of this variety, there is a deeper question to solve: can we test fairly given this heterogeneity?

Session 7: Teaching and school based management based on information. A.J. Visscher.

Theory of change

It is about evaluating and analyzing results, setting SMART and challenging goals, determining strategy for goal accomplishment, and executing for goal accomplishment. Objective is to find relevant quality data, analysis & diagnosis of problems; formulating (improvement) plans deliberately, execute them, and evaluate their effects.

Where do differences in performance come from? Variability in performance is 4 times higher at the classroom level than at the school level. Most differences in student performance have to do with the students and then teachers.

Proposal

There are four parts to a data driven decision-making strategy for education:

1. Analyzing the relevant data - value added, achievement gains, and various analysis of student performance, that allow specialists to separate content that is mastered by students and compare ability growth in students compared with national average. It is a good idea to complement this with classroom observations, students' perceptions of education quality.
2. Goal setting - Ideally numerical goals.
3. Choose a strategy for goal accomplishment - Its important to differentiate between basic didactical skills and complex ones, in order to differentiate instruction.
4. Implementation - its important to see what happens in the classroom and how data is used.

Its important to move to a scenario where tablets can be used to ask questions, and teachers can receive real time information to immediately know where to put more effort in instruction. An open question is, what happens in a classroom that is able to use all information, does it improve quality in education? Is there a strong link?

Conclusions
Its important to think about the theory of change that explains how evaluation will lead to improvement, in other words its important to think through what information will cause change. Improvements will come with solutions provided by the school and not external to the school. More emphasis should be given in creating the systems that support teachers to find problems in instruction and help them to solve the problems. Good teachers are a necessary ingredient, and technology can help us in these endeavors.

Open questions
In Mexico there is an emphasis toward greater teacher accountability and informing policy makers in order to better choose appropriate resource allocation. This is different from the Netherlands approach, which intends to inform teachers of the results to improve the classroom experience. This may be a model in contexts where you have very good teachers. Can it work for low performing schools? Also how to think about improving accountability under this focus and not necessarily with the traditional focus of carrots and sticks?

Discussion

Jakob Wandall

Comment: Message was depressing and uplifting, it is good to see what can be done with the use of data. Even with a lot of data readiness, there are still many challenges in using data to improve learning.
Answer: Big lesson is that teaching is an uncertain enterprise. Classroom management is everything, in an ideal world; teachers have a chance to check who is learning and who is not. The current frequency of evaluation is too low. Experiments have been carried out with a digital monitoring system. Has this lead to better student achievement? The answer is yes, some factors are promoting success, and it has to do with frequency of tests, although there must be

a good distance between tests to measure value added. Also if you tell teachers what to do with different results, this can contribute to success.

A good idea is to use tablets to design questions and answers and hence provide teachers with real-time information. In that case teachers can immediately see where to put more effort. The most important question to be asked is: will teachers be able to do much with all the data they are getting?

Daniel koretz

Comment: Data displays are very rich and very interesting. Profiles of strengths and weaknesses are particularly interesting. This is one way of showing that reporting scores is much better than reporting standards (proficiency levels) so that you can have a flavor of the degrees in learning. It is helpful to use norms to compare across skills within classes in a school, to see where children stand relative to the average or other norms.

A particular display shown that compares growth for low performing and high performing students is very risky, because slopes are very easy to change depending on the scales used. The safest thing to do is to compare across similar students, for instance compare the growth in my students with similar kids in another group. It is worse to compare across teachers. In Tennessee they used value added modeling. With a period of 5 years they did a basic skills test. The slope would be flat even if there were good teachers, because the test wasn´t able to detect higher skills.

Marguerite Clark

Question: To what extent does this model generalize to other contexts, given differences in scale, differences in the education system, and the differences in governance systems?
Answer: yes, there are many differences between Holland and Mexico, even with decentralized vs. centralized systems. What is important is that people that have to use feedback to have the right set of skills. The question is can this work in Mexican schools?

Bernardo Naranjo

Comment: In Mexico there is an emphasis toward greater teacher accountability and a need to inform policy makers of education attainment in order to better allocate resources. This contrasts to the model in Holland that focuses on providing teachers with information to improve instruction. The end user should be taken as the student, the teacher, and the school director. For a moment we should think that decision makers have secondary importance. The system has reacted to external factors, that are not necessarily aligned with the end-users' needs. Europe has both accountability and formative first order purposes. However there is great difference in Europe, for example England has a more punitive system.

Carlos Mancera

Comment: The graph that shows where variance in student performance comes from can explain why the emphasis in evaluation is placed in affecting different actors. While most of the variance is explained by variations of teaching, teachers in Mexico might have a very limited role to play in affecting performance, in a context of high initial variability in performance. Teachers themselves have very little say of how to teach, there is limited autonomy. Technology can play a great role, and the success from standardized tests can be expanded through the use of technology and this can unleash many constraints at the school level.
Answer: There is much variability in how teachers use data, based on the data, some teachers are fired or they leave voluntarily because it is clear who is improving and who is not. First time it might be scary for teachers but then they find it very useful to have data and to compare. The model might not work for weak performing schools because even with presence of data, there are not many alternatives.

Teresa Bracho
Comment: Schools have to be the center of all processes and decisions, and results should always be contextualized. The teacher is not necessarily the end user of results; it could be the collective group of teachers that create an environment for improvement. Accountability is not just about providing traditional incentives (carrots and sticks), but it is much broader. It is useful to think about how evaluation can improve accountability in a much broader sense than just thinking about punitive versus rewarding systems.

Daniel Hernandez
Question: Should there be exams for all disciplines in primary school? How frequently would you like to see data? How to combine different sources of information when thinking about using student testing and student perceptions questionnaires? How frequent would you want other types of data besides student testing.
Answer: Yes, for primary schools it is important to focus on math, language, reading comprehension, and not necessary in history or geography. How often, how many times? There is no clear answer, but we need to have significant distance between tests to allow for improvement. In Holland, student perceptions have been collected twice a year.

Session 8: Proposed design for the second generation of standardized student assessments in Mexico. Eduardo Backhoff
Mexican experience

An account of the Mexican experience, roughly since 2002, starts with the creation with INEE, which is equivalent to NAEP. Standardized testing, through sample based testing was implemented in the end of each level of education, (pre-school, primary, secondary and high school) every four years in subjects like math, and reading comprehension.

In addition SEP implemented annual census based testing, ENLACE, in every grade, starting from third grade. In many cases there was an overlap between these two tests (census and sample based tests). ENLACE had an important effect in promoting a culture in evaluation. This was partly due to the fact that test results had an effect on teacher pay, affecting extrinsic teacher incentives. Sample based tests then lost relevance, especially when the attention of teachers had an effect in the media. Sample based testing was no longer an input in decision making among policy makers.

Standardized testing

The purpose of standardized testing is to assess levels of learning across students, for various levels of aggregation (student, classroom, school, municipality, district, state, national level).

The main objectives of evaluation is to improve student learning and teacher practices; improve decision making; and give timely information to schools, teachers and parents to strengthen accountability of the education system.

Guided by the Gordon´s Commission

The most valuable purpose to hold in evaluation systems is to improve pedagogic practices. Evaluations should be designed to give information to support students along their learning progress and to personalize the learning experience. Evaluations should not be used to hold students and teachers accountable for the results.

Standardized testing can only evaluate some measurable skills, they are not designed to evaluate domains of knowledge of equal value like writing, artistic expression, and active citizenship skills. Nonetheless whatever the subject, it becomes the main focus of attention for teaching in the classroom.

INEE´s position on evaluation

Evaluation should be used to assess if the right to receive high quality education is granted to children, youth and adults in Mexico. The underlying purpose should aim towards improving processes and results, considering Mexico´s cultural diversity.

Design implications for the second generation of standardized testing

Strong efforts should focus on avoiding corruption in the evaluation process, and in the potential inflation of results. It is strongly recommended to have better quality control in the implementation of standardized testing. Also eliminating high-stakes for key stakeholders will help reduce perverse incentives.

INEE proposes a shared responsibility between the Institute, authorities, and schools to carry out the implementation of the various evaluations such that census and sample based testing can be combined in a cyclical throughout the different school years.

A new proposed design

1. Measure the same constructs in both census and sample based testing.
2. Sample based testing should focus on strengthening accountability in the educations system by illustrating learning outcomes. INEE proposes that these tests should be applied at the end of the school year.
3. Census based testing should focus on improving the pedagogic and formative processes in instruction. INEE proposes that these tests should be controlled and administered by the schools at the beginning of the school year in forth, seventh, ninth, and twelfth grades.

End users of evaluation

- Sampling based tests: public authorities, and society in general.
- Census based tests: school principals, teachers, students and parents.

Sample based tests

INEE should be wholly responsible of the entire process. Authorities and schools should support INEE. INEE should document all the stages of the process, and prove the validity of the results.

Census based tests

These tests should contain the same constructs as sample based tests. Public authorities should be responsible of the printing, implementing, guarding quality control, grading the tests, analyzing and dissemination.

INEE and SEP should be responsible for controlling the quality of all stages of the evaluation through statistical validation or supervision in situ. INEE will give guidance through out the different phases of implementation.

Schools and evaluation

- The item bank in the census based tests should be published to be used as a diagnostic instrument for students.
- INEE will produce a manual for implementation, grading, and use of the results for teachers, and school directors.
- Schools will be able to reproduce the results and implement instruments.
- Results should be known to teachers, students, parents, but not the general public.

Strengths of this new design

Three big goals are being fulfilled

1. Evaluate the education system while improving implementation.
2. Evaluate all the schools with three grades in between each test, combining information from tests with supervision in situ.
3. Annual evaluation to all students, through a controlled census based test, while delegating testing to the schools. Census evaluation will be done at the beginning of the year which will alleviate school responsibilities.

Weaknesses in the design

- Census bases tests will not be completely controlled by a central authority, unless authorities agree to finance and plan for this.
- Also results in the census based test might not reach schools on time to fulfill its formative purpose.

Conclusions

There is much evidence that using standardized test with the purpose of informing and promoting accountability has had limited impact on learning outcomes. INEE will like to design an evaluation system that will improve learning, by giving greater emphasis to a formative assessment.

Ideally the new generation of standardized testing should provide two big levels of information: (i) information to strengthen accountability, while giving timely feedback to improve education policy; (ii) information that allows school principals and teachers to improve management and schooling, and better adapt teaching practices to improve levels of learning and reduce inequalities. This new information should be an input for planning and designing programs that meet different needs, and also to adjust instruction in the classroom. Also teachers should be better trained to improve their evaluation skills in the classroom, in order for them to complement standardized evaluation with their own evaluation.

Discussion

Mark Wilson

Comment: The fact that the new design contemplates two different tests and having them both focus on the same content is a positive sign. Aiming at the school is a good choice. It is important to include schools and teachers content development of the test, so that they are up to speed with what happens in the test. You have to think of the content and how the results will be used. It is useful to think about the reporting before rolling out the tests so that this actually turns out to be a useful exercise.

Carlos Mancera

Comment: Thinking of two different types of tests is a good suggestion. There might be negative effects from implementing a census based test at the beginning of the school year. It is clear that once teachers get the results back (in the best case scenario around December), it might be already too late to improve instruction. A recommendation would be to think about applying the census based tests at the end of the school year, in order to provide a diagnosis every year, and for every grade for formative purposes.
Answer: Tests at the end of the school year do not help to improve instruction, they are only used for accountability purposes. If tests are applied in the beginning, then tests can serve as a diagnostic purpose. Timely information will be a technical issue.

Rafael de Hoyos

Question: Could we think of a census based tests at end of school year capable of providing the detailed school by school, classroom by classroom results to be used at the beginning of the following school year to fulfill the formative purpose of evaluation?

Bernardo Naranjo

Comment: Its important to think about scoring starting from the average score of 500. This way each student can be compared with others, and it is a relative measure, and not an absolute measure. It is questionable to apply the census based test at the beginning of the school year, since the effect of vacations would make results less reliable. It is also better to test fifth grade instead of sixth grade in order to have one year to react. A good policy to push would be to make sure that students shouldn´t go past 8 years of age without knowing how to write or how to read.
Answer: Testing in the end of the school year would correspond to a summative assessment, rather than a formative assessment. It would be good to test in the beginning of the forth grade, seventh grade, and ninth grade.

Daniel Hernandez

Comment: If the census based test is applied at the beginning of the school year or in the end, it is not as important in the steady state. What matters is to have systematic information allowing comparisons over time. Whether its summative or formative, it refers to the use of the information. Its important to clarify what will the use of information be, and who will have access to the information. At the very beginning, ENLACE was thought as a test to inform parents and teachers only. However it became a politically sensitive topic and the information had to be disclosed.

Daniel Koretz

Comment: Many problems arise when many purposes are behind the design of the tests. Its important to assess the capacity of teachers to make use of this information, therefore thinking about professional development of teaches and ways in which they can use the data becomes instrumental in shaping the success of evaluation efforts. For instance, it’s good to ask the following question: is there any capacity to use these results in a context were there are limited resources to train people. For example, in the Israel proposal, special credentials were given when teachers were able to use data to help train others. This failed, and government support failed.

Francisco Soares

Comment: In the presentation there was no mention of the curriculum. Mexico has a national curriculum. The curriculum is the expression of what should be taught in schools, and therefore it is an essential ingredient for the evaluation system. Evaluation is part of the system - even if we do a perfect job, it is not a solution to the problem. As a result of trying to satisfy different stakeholders, we are mixing two objectives and this is not ideal. School and teachers will care for formative assessments whereas the society will care for summative assessments. Nonetheless, they should measure the same construct, which should be a reflection of what the children should know. We should have a very large number of test items that concretely tests constructs within the curriculum. How many evaluations should the system consider? Should they be applied once a year, twice a year, which time of the year? These are all decisions that should be taken by the teacher to fulfill the formative purpose of evaluations. We should not have a test that has a formative purpose, and at the same time, use results for summative assessments. Teachers should spend much more money with the tests that can help teachers to be better.
Answer: Formative and summative assessments are very different in nature. Purposes are very different and the use of test results should adapt to each one of these purposes.

Follow-up comment by Francisco Soares: Curriculum is unique, text books are unique, and what is key is to evaluate according to the curriculum. The idea of which nodes are important to test is another issue. Its important to think about how big your item bank will be. How many items will be assigned to each construct?

Rafael de Hoyos
Question: Why is the proposed design deviating substantially from the original design of ENLACE? Why can't we improve from what we already have with a clear identification of its drawbacks and limitations --for instance eliminating the link that ENLACE results had with teacher incentives -- as opposed to starting from scratch?

Jakob Wandall
Comment: In Denmark there is an important distinction among summative and formative tests. It might be good to think about different constructs in these two tests. One way of approaching the design is to base the general construct as connected to formative purposes, and use the curriculum to design the summative test. A good way of designing this is to test how the student performs in theses two.

Carlos Mancera

Comment: It is very important to make sure that the national curriculum makes sense. The current syllabus is available with an official format (a legal document) which is not user friendly. It is strongly recommended to redesign the curriculum, in a more coherent matter, while making it more accessible to end users. Before thinking about evaluation, a first order question should be to think about curriculum redesign in order to differentiate between summative and formative standardized tests. It could be reasonable to design an evaluation system where standardized evaluation for formative purposes is implemented internally by the schools at the time and frequency that they consider adequate, and have externally implemented, controlled, evaluations for summative purposes.

Arcelia Martinez

Comment: It is important to think about standardized testing as linked to a broader picture of the evaluation system. The Netherlands example can point towards the value of thinking about the causal relationship between evaluation, information and, eventually, improvement in results.

Lorena Meckes

Comment: A lot of support is needed in the use of the results in standardized testing. What will be done to monitor each student as he or she goes through a learning process of understanding the various constructs? Also it is important to think about the collective use of the results to affect the school community, and overall affect pedagogic practices in classrooms.

Daniel Koretz

Comment: An important note: the word formative is being used a lot in many different ways with different meanings: sometimes referring to a benchmark test, sometimes to a diagnostic test, and sometimes referring to tests that can show teachers why is it that students have certain misconceptions. It is a problem to confuse these purposes, as the design implications are very different. For example PISA uses complex problems to assess competencies. This design is actually terrible when used for diagnosis purposes, because the ability of the student to perform certain tasks requires a complementarity between different domains of knowledge, therefore this test will not tell you if a student performs well in a particular skill. The design of diagnostic tests involves items that test specific skills. In a formative tests, items will help explain why a student doesn´t have a particular skill, and what are the misconceptions contributing to the misunderstanding of a specific skill. Therefore it’s important to first know what purpose will guide the design of standardized testing and what the implications of this decision are.

Gilberto Guevara

Comment: Evaluation has a political and moral perspective; it is an instrument that helps citizens to have the right to education. Assessment is part of the educational process. When we compare Mexico to Netherlands, Denmark, US, Chile, Brazil, we see that those countries are ahead of us. As a consequence, how can we make Mexico go where the rest are, to foster improvement in education?

Assessments have to be an engine to improve education. This might require us to be humble and face the reality of what the status is. We have to also consider the context. The context tells us that 30 million individuals are in extreme poverty, there are many different languages, there is enormous inequality, and a privileged minority. Also in terms of the educational context, around 40-50% of schools are “incomplete schools” because they count with a limited number of teachers. Also there is a major difference between public and private schools. Private schools tend to have greater teaching expertise that is lacking in public schools, for example there are many private schools that have experimental pedagogic practices (like Montesorri, Piaget, etc.) that can be generating important gaps in learning.

Bernardo Naranjo

Comment: The end user is very important. It is highly recommended to form engagement groups with school principals, directors and teachers regarding the design of testing, to promote the legitimacy of the evaluation system.

Session 9: Round table about the future of standardized testing. Daniel Koretz.

In thinking how to structure this round table, it turns out that there may not be one trend in the future of standardized testing, instead it is useful to think about what is likely to occur in the future design of assessments in Mexico.

It turns out that there are more questions than answers, and the framing of the discussion will be around the main issues and questions that are important to consider. From the discussions made so far, there are four major themes: questions about everything, questions about goals, questions about design, and questions about checks and balances. Within this universe of questions, the broader question is how to maximize learning and minimize the negative effects of testing.

Thinking about goals

From what has been said, there are many goals, and the conclusions get to be confusing. Part of the problem is that there is not one system that can serve all purposes. We have heard from Brazil, US, and Chile evaluation systems that try to reach 3 major objectives: assessments made to inform policy makers, with a diagnosis of the education system as a whole; assessments with diagnostic and formative functions that can guide better instruction in the classroom; and internal and external assessments that strengthen accountability in order to give better incentives for better performance, respectively. The important point is to decide what will be the priorities, and how will this affect the design and the implications of what to take and what to leave out.

What should the larger evaluation system be?

It doesn't make sense to think about one test with out thinking about what the system will look like as a whole. You can´t decide a particular design without taking into account other tests. It is important to think about how to make results from tests count? And what support mechanisms and training systems should be in place in order to make this process effective. Finally, the overarching question is with respect to checks and balances: how can we evaluate the evaluation system?

For years policy makers have worked with the assumption that if test scores go up, then the educations system is improving, however this is not a good assumption to make, and it is important to monitor the evaluation system as a whole. Jim Heckman worked on the topic of job training programs in the US, and the results showed a great amount of gaming, and what agents can do to change the numbers. Therefore it is useful to think about dynamic accountability, and how monitoring and evaluations systems should always evolve. This last point should address the following question: what evaluation system will be put in place to evaluate gains and scores, and what are the teachers doing differently in the classroom due to the effects of the evaluation system?

Discussión

A.J. Visscher

Comment: Formative purposes are most important in my perspective, but with respect to the accountability purpose, it is important to think of the type of information needed to strengthen accountability. One challenge can arise from the limited number of observations at the school level. This makes it hard to make certain claims from the data collected. What governance measures will be put in place along with the accountability information received? How high will the stakes be? What will decision makers do with the information?

Rafael de Hoyos

Question: There is one more issue to be discussed and this refers to the use of standardized testing for evaluating education interventions. Can this other purpose be accounted for in one test or multiple tests? Is this an objective included in the design of student assessments in other parts of the world? Can you follow students over time to see how the students evolve, and under certain conditions, attribute some of those changes to certain policies?

Carlos Mancera

Comment: A comment on the accountability: how can checks and balances enter in the design of standardized testing. Diagnostic and formative issues can be tested in other ways; besides standardized testing, there are other cheaper and more productive ways of reaching the formative objectives of evaluation. However giving information to society is an integral part of the equation and of the evaluation process. Therefore standardized testing has to give information at the school level, and parents have the right to get that information. It is hard to think of only one census based test in 4th grade and one more in the first year of secondary school. Finally, when designing standardized testing, it is important to agree on the type of information needed for accountability purposes, this will allow citizens to assess school improvements through time.

Eduardo Backhoff

Comment: The evaluation systems that were created from the beginning were not designed for accountability purposes; they were always designed to improve learning. Accountability is important, and the information from testing can be used for this purpose. However it's important to decide at what level are we going to push for better accountability systems, and how extreme will this be driving the design of evaluation. Do we want to be so extreme as to evaluate all grades, all students, every single year? The level of intensity in which you can use test results is very important; too much testing and information can paralyze the education system. The accountability purposes underlying the design of the evaluation system should be moderate.

Lucrecia Santibañez

Comment: There needs to be a debate about the optimum frequency/intensity of testing. This is a very important question. Who makes the decision, what information is supporting this decision? This will shape the decision making process, and it should be based on evidence and on a coherent thought processes. Today the system is rather closed with limited transparency of the quality of education in each school. There is little engagement from parents, and limited access to information. There is a tradeoff between greater transparency with greater information vs. better instruments for better teaching and better performance. However it is important not to leave the issue of transparency and accountability to society off the table. Civil society nowadays demands information for every school, making it hard to go back to a system that does not disclose the results at the school level.

Marguerite Clark

Comment: Lessons learned at the Bank- Success in large scale assessments depends on a range of factors: technical rigor, leadership (who are the champions), institutions that will be guiding and supporting the process, cost/budget, and ways of financially supporting efforts, training to teachers, etc.

Its interesting to look at the case of Singapore, it's a small system, reforms are very easy to implement, and any decision is politically feasible. However a good and stable assessment system took 10 years to put in place, because efforts started through small pilots that were scaled little by little.

Gilberto Guevara

Comment: There is a serious issue regarding the pressure to give more importance to accountability in the design of standardized tests in Mexico. The Gordon Commission´s report showed that large-scale evaluations that privilege accountability have not produced important improvements in education. Policy decisions on resource allocations are important, but it is more important to think about how to protect the constitutional right of guaranteeing quality education for all children and youths in Mexico. Pressures for an improved accountability might come from civil society groups that represent the voice of the private sector conglomerates. Therefore we should take these issues with a grain of salt. Too much transparency leads to punitive consequences for teachers and this damages the learning process a lot.

Daniel Hernandez

Comment: Thinking more about inflation of results, not everyone is inflating properly because there is still a lower limit. Why are some schools cheating and others not? We have to delve into this issue if we are to design a really controlled test.

As a response to Gilberto and Eduardo: an important goal of the standardized test is to promote a culture of evaluation. Parents might not ask for it, as when democracy was not asked for in the past. It is important to keep distributing the results at the school level to keep parents engaged. Transparency is a broader part of the evaluation agenda. We must be thinking about which spaces to open in order to promote parental engagement.

Question: From a technical point of view: how are we going to do a robust longitudinal analysis of the results in presence of an accumulated measurement error.

Teresa Bracho

Comment: Accountability is important to inform society. It should not be seen as a minor purpose since it is important to tell citizens in what ways the education system is supporting the population and in what ways it isn't. It is important to use instruments that can help identify which interventions in education are helping improve learning and which are useless. Also it is important to distinguish between responsibility and blame in the case of teachers. Teachers are not to be blamed for bad scores but they are responsible for improving learning of students. Teachers should be able to design and implement evaluations and to use the results for improvement. Also parents should be aware of what their children should be learning. There is a good case for promoting a culture of evaluation to generate an enabling environment for learning. Experts and designers should take into account value added measures for these purposes.

Eduardo Backhoff

Comment: We have to think about the distribution of roles and responsibilities among different stakeholders in the public administration. This has implications for who has access to what type of information. This design can end up posing a threat to school autonomy.

Daniel Koretz

Comment: We should re-think the Singapore example as a way of pilot testing our way towards a better system. This requires answering the question: is there something we can do to get more information on individual schools? It means delaying 'progress' for a while, but there is a significant advantage of going this route. School level testing is very hard to do. We are stuck in a conundrum, the more we want to know from individual schools, the less we actually know. Theory has not advanced to a level where there are clear conclusions on how to do things; best recommendation is to test things out empirically via pilot programs.

Bernando Naranjo

Comment: Pay for results is not the only way of improving performance, there could be a pay for reputation. Shedding light on worst performing schools can actually generate a 'poverty trap' through the stigmatization of these schools pushing them further down instead of helping them perform better.

Carlos Mancera

Comment: Large scale census based standardized testing has to take place at least between 3rd grade and 6th grade (testing every 3 years at least), or else society would not have enough information to know where are the important gaps and how the education system is performing. Not disclosing the results of poor performing schools is the best way of generating a poverty trap and exacerbating inequalities.

Daniel Hernandez

Comment: Eduardo had mentioned that too much government intervention could go against school autonomy. School autonomy is a tricky issue. Today we do not have the necessary building blocks to promote school autonomy. We still need some level of government intervention to ensure minimal standards of quality.

Jakob Wandall

Comment: Accountability strengthening is not strongly related to improve learning. There is limited evidence of the effect that accountability has in improving learning. If you measure correctly, the only way to see system performance is to improve individual performance. Improving individual performance can be a result of the quality of teaching. Very few studies can tell us which mechanisms can help to improve Mexican´s performance, and how to maintain democratic control over resources.

Rafael de Hoyos

Comment: There is consensus when saying that assessments should be diagnostic for improved performance. There is not a consensus when thinking about the accountability aspect of assessments. There is no reason for choosing just one, perhaps the two objectives can be met. One test can serve two purposes: a diagnostic one and a summative one. On a different note, it is important to stress that not everything related with ENLACE is wrong; for instance, to certain extent due to ENLACE, we are having a debate about how to improve learning outcomes. Quality of education is now at the core of the education debate and without ENLACE we would have no starting point.

Daniel Koretz

Comment: One test cannot be used for two purposes, a diagnostic and a summative test. It is important to pick the purpose of what matters most. In presence of thick trade-offs we have to set priorities. An older version of testing allowed for better diagnostics, and currently testing helps assess if children can solve hard problems.

Rafael de Hoyos

Comment: However if you have to design one test, assuming it will be imperfect, its best to accommodate for both purposes (accountability and diagnostics).

Daniel Koretz

Comment: Once you have accountability at the school level, teachers are teaching to the test. This is why you need to think about what you want to measure, or else teachers will change their behavior.

Marguerite Clark

Question: What has been the experience when countries go through costly large scale testing?

Daniel Koretz

Comment: Tests can end up being very costly and very time consuming. It is hard to produce high quality performance testing.

Lucrecia Sanibañez

Comment: We have no idea what happens in the classrooms. If teachers were teaching to the test, then this would be an improvement from the starting point of no teaching at all. In one way we need standardized tests to allow for better results. But we also need to improve standardized testing, nevertheless it is still better to have information even if testing has its limitation.

Carlos Mancera

Comment: Evaluation can influence teacher practices. Evaluation has to have an accountability component, one test every three years, not necessarily at the student level, but it should produce a starting point for improvement. Daniel Koretz has documented the negative consequences in large scale testing. It is important to prioritize the goals, the use, etc. I suggest very few tests, only key tests, and the rest of the efforts should be directed to support teachers. No stakes testing can be arranged by INEE. There are many options in the design of evaluation system, and INEE can be very selective of the large scale standardized components.

Session 10: Concluding thoughts. Sylvia Schmelkes and Reema Nayar.

Reema Nayar

A clear success of the session is that there is great food for thought. The key aspect is how you bring these lessons to bear while designing a concrete evaluation system. There is not one solution, and it would be good to know if there is still room for piloting.

Sylvia Schmelkes

Evaluation is not developed as a proper science, in some ways the discussion is confusing and in some ways it gives a better perspective.

In the end we want to focus efforts to improve learning, and this means focusing on what happens in the classroom. This is the first priority. Secondly, it is important to know what happens with learning to guide future decision-making. Where are the gaps and how can we affect these gaps? This means supporting teachers in using results and how to change the curriculum. Thirdly it is important to strengthen accountability. We should allow society to know what is happening in education and to hold stakeholders accountable. In fourth order the priority can be given to the use of standardized test to evaluate and monitoring education policies.

These are the pending thoughts:

1. Evaluation is not going to be an automatic process.
2. It is important to think about how information will be used for public policy decision-making.
3. How will teacher training be affected by evaluation?
4. How this agenda will change the curriculum?

Annex: Bio of each participant

Sylvia Irene Schmelkes del Valle
Presidential Advisor of the Council Governing the National Institute for the Evaluation of Education

Sociologist, Master's in Educational Research and Develop- ment from the Universidad Iberoamericana. Researcher in education since 1970. She has published over 150 works, between books and articles, on the topics of education quality, adult education, training in values and intercultural education. She founded and was General Coordinator of the Intercultural and Bilingual Education of the Secretariat of Public Education (2001-2007). She served as president of the Education Research and Innovation Center of the OCDE (2002-2004). She received the Joan Amos Comenius medal, awarded by the Czech Republic and UNESCO, in 2008. She directed the Research Institute of Research for the Devel- opment of Education in the Universidad Iberoamericana in Mexico City from 2007 to 2013. She is currently the Presi- dential Advisor of the National Institute for the Evaluation of Education. She is a level III National Researcher.

Marguerite Clarke
Senior Education Specialist, The World Bank Group

Marguerite Clarke is a Senior Education Specialist in the Ed- ucation Global Practice at the World Bank. She leads the Bank's work on learning assessment, including assisting countries around the world to improve how they measure and use information on student learning. A former Fulbright Scholar, she received her PhD in Educational Research, Measurement, and Evaluation from Boston College (2000), and is a member of the Learning Metrics Task Force Advi- sory Committee, the advisory board of UNESCO Institute for Statistics, and the editorial board of the journal, Theory into Practice. She is the author of "What Matters Most for Student Assessment Systems" (World Bank, 2012).

José Francisco Soares
President of INEP, Brazil

Jose Francisco Soares is retired Professor of the Federal University of Minas Gerais. Since February is the president of INEP – Brazilian National Educational Evaluation Authority. Before he was a member of the National Education Council (CNE), and of the Board of Governance of the movement "Education for All". Has a PhD in Statistics from the University of Wisconsin – Madison and postdoctoral studies in Educa- tion from the University of Michigan – Ann Arbor. He received in 2012 the Bunge Foundation Award for his lifelong con- tributions to Educational Evaluation in Brazil. His academic work is focused on measures of educational outcomes and calculation and explanation of the effect of schools of basic Education in Brazil.

Lorena Meckes
Former Director of Evaluation, Education MInistry of Chile

Currently she is a researcher for the Center for Policy and Practice in Education in Universidad Católica de Chile (CEPPE), with a focus area on teacher training. Under this role, she has lead projects related to the development of national standards for graduating professional teachers. She is a professor in the Education Division of the same University, specializing in Education Evaluation. Between 2003 and 2008 she actively participated in the national measurement system of education quality or Sistema Nacional de la Medición de la Calidad de la Educación (SIMCE) where she lead the development of national standards and promoted improvement in learning levels in the country. She also lead initiatives related to the publication of results for an improved accountability. She has been a consultant, and an expert in the development of standards and in standardized testing for the OECD, World Bank, and various governments in Latin America and in the Middle East.

Jakob Wandall
Executive Director, NordicMetrics Aps

JW has a background as a researcher in social science, pri- marily evaluation and education (1984-1994).

From 1994 to 2011, JW was employed by the Ministry of Education as a Head of Division/Chief Adviser, covering dif- ferent areas (lifelong learning, adult education, youth ed- ucation, and primary and secondary school). In 2004, he was the key architect behind an adaptive test-design and project manager of the development process, tender, and implementation of the IT-based National Testing system in Denmark. Until 2011, he worked with development, dissemination of knowledge and QA of testing, assessment, examination, and IT-based learning.
Since 2011, JW has been providing independent research and consultancy services for development of efficient as- sessment and training with clients, including research and educational institutions, publishers, municipalities, and other authorities. Furthermore, he is an external senior lecturer at University of Aarhus (DPU) in educational science. Today, JW is partner and executive director (R&D) at NordicMetrics Aps.

Daniel Koretz Professor of Education, Harvard University

Daniel Koretz is Henry Lee Shattuck Professor of Education at Harvard University. His research focuses on educational assessment and education policy, particularly the effects of high3 stakes testing on educational practice and the problem of score inflation. His research has also investigated the assessment of students with disabilities, international differences in the variability of student performance, alternatives to traditional college3admissions testing, and the application of value3added models to educational achievement. His current work focuses on variations in score inflation across types of students and schools, the relationships between test scores and later outcomes, and the design and evaluation of 'self monitoring assessments.'

Mark Wilson
Professor of Education at the University of California, Berkeley

Mark Wilson is a professor of Education at UC, Berkeley. He received his PhD degree from the University of Chicago in 1984. His interests focus on measurement and applied sta- tistics, and he has published just over 100 refereed articles in those areas. Recently, he was elected president of the Psy- chometric Society, and also became a member of the US National Academy of Education, and a Fellow of the Ameri- can Educational Research Association. In the past few years, he has published three books: one, Constructing measures: An item response modeling approach (Routledge Academ- ic), is an introduction to modern measurement; the second (with Paul De Boeck of the University of Ohio), Explanato- ry item response models: A generalized linear and nonlin- ear approach (Springer-Verlag), introduces an overarching framework for the statistical modeling of measurements; the third, Towards coherence between classroom assessment and accountability (University of Chicago Press—National Society for the Study of Education) is about the relation- ships between large-scale assessment and classroom-level assessment. He has also recently co-chaired a US Nation- al Research Council committee on assessment of science achievement—Developing Assessments for the Next Gener- ation Science Standards.

A.J. Visscher
Associate Professor at the University of Twente, The Netherlands

A. Visscher is an associate professor at the University of Twente where, over the years, he has been involved in and has led a large number of European and national research projects. The central focus of his research has been on the school organizational characteristics explaining differences between schools, in terms of the average learning gains of their students. During the last five years, 3.8 million euro of research funding was acquired from a variety of sourc- es (e. g., Onderwijsbewijs, NWO, Kennisnet, the Ministry of Education, the Dutch Inspectorate). His volume “Managing Schools towards High Performance,” in which school orga- nization theory is linked to the school effectiveness knowl- edge base, has been used at Dutch and foreign universities for many years.

During the last 10 years, he has focused his research activ- ities on the potential of school performance feedback (e.g. quality assessment data from CITO student monitoring sys- tems fed back to teachers and schools to show them the results of their efforts) for improving teacher and school qual- ity. The volume School Improvement through Performance Feedback (with prof. Coe from the UK) on this topic is widely known. During the last five years, he has developed and im- plemented professionalization interventions for school teams and individual teachers to promote data-driven teaching. The interventions have been implemented in more than 200 Dutch schools and, in parallel to the training activities, data has been collected longitudinally on the impact of the profes- sionalization activities.

He has published widely on the aforementioned topics in sci- entific journals and books (140 publications).

At the request of the Kenniskamer of the Ministry of Edu- cation, in cooperation with Dr. M. Ehren, the analysis “De eenvoud en complexiteit van opbrengstgericht werken” was completed and received wide recognition. He currently supervises 6 PhD candidates and chairs (with Dr. A. Tim- mermans) the Research Theme “Schools and the Societal context of Education” of the Interuniversity Centre for Edu- cational Research.

Eduardo Backhoff
Advisor to the Council Governing the National Institute for the Evaluation of Education

Psychologist and a PHD in Education. Level II Research Fellow in the National Research System (SNI). He is a co-author of various Standardized Assesments that are used by public universities around the country to select and assess students from higher secondary levels to join tertiary education, these include knowledge and skill testing and basic competency testing. He has lead the development of the test Excale (Exámenes de Calidad y Logro Educativos), designed by INEE, that was used to assess quality of basic education system. He has been an advisor of international standardized large-scale student assessments in collaboration with the Organization for Economic Co-operation and Development (OECD): PISA-2012 and TALIS-2013. His latest research projects include validity of standardized student testing both with national and international scope.