Session 1: Welcome, context and objectives of the workshop
Reema Nayar, Sylvia Schmelkes, Marguerite Clarke

A perspective of what shapes the design, implementation, and use of student standardized assessments is presented. Elements like purpose, context, and the theory of change are fundamental building blocks in this process.

There are many types of standardized tests that rest on different premises. It is important to consider a country's performance in a series of quality drivers: enabling context, systems alignment, and quality of tests. Depending on the context, a country can choose among classroom assessments, high stakes examinations, or large-scale assessments. Ideally we would only have classroom assessments to improve learning, but there are important constraints for this, for instance quality and motivation of teachers.

Previously, Mexico used a census-based test (ENLACE) to link teacher pay to results from student assessments. This led to unintended consequences, such as teaching to the test, corruption, and "gaming" of the system. INEE is now in a position to learn from experiences in Mexico and abroad on the use of sample and census based testing, the cost effective ways to measure value added, how to account for cultural diversity, and how they can be used to improve the teaching-learning process in classrooms.

What will be the guiding principles behind the second generation of standardized assessments in Mexico? The emphasis will be on improving learning outcomes in an environment that acknowledges the diverse and unequal contexts prevalent in Mexico. To achieve educational goals, curricula must be improved and learning gaps must be reduced. Student assessments should thus be used to develop improvement plans and tackle learning challenges.

New trends point toward a more important role for teachers and technology. Increased attention is also being given to the use of cognitive and psychometric models to inform the design of standardized testing.

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

The main goal underlying Brazil´s evaluation system is to provide a tool to promote the right to education and to make better citizens. Great emphasis is given to bridging the learning gaps that result from inequality and to better equipping the system for cultural diversity. Many questions are left concerning teaching and performance, given that greater monitoring and improved teaching do not always lead to a greater success rate in results.

Brazil´s evaluation system is characterized by the following assessments:

- ANA (National Literacy Assessment): tests reading, math, and writing every year, for all 3rd grade students.
- PROVA Brazil, SAEB (National System of Basic Education Assessment): tests common core knowledge in reading, math, and science every two years, among all public school students in 5th

and 9th grade.

- ENEM (National Exam of Secondary School): tests language, math, science, humanities, and writing in a large sample of students in various Government programs (around 8,700,000 candidates).

Under the new national plan in education, IDEB was selected as an indicator that should be closely monitored for the improvement of basic education in Brazil. IDEB is an indicator that multiplies performance, measured by the average student proficiency in PROVA Brazil SAEB, by the average promotion rate at each school grade. IDEB has helped set a goal for the system and a goal for each school, and is one of the bases for resource allocation among schools. There are goals for 2021 and all states in Brazil should have the same IDEB by 2095.

There are two key indicators in standardized testing: learning progress and learning attainment. In Brazil, data can now further illustrate this relationship. Since 2007, Brazil has been able to construct individual learning trajectories for each student. In this stage, a deeper analysis has to be made to understand how assessments interact with the curriculum and how teacher quality and regular monitoring can be important levers for achieving better results.

Brazil lacks a unique national curriculum. The evaluation system currently monitors performance and value added, but questions remain regarding the sustainability of the system and there is uncertainty on policies capable of improving the system's quality.

It is important to see the complete picture in education and to contextualize results (socioeconomic status, teacher training, spending per student, etc.). Indicators should mean something in order for teachers to link classroom practices with results. More rigor and attention should be given to pedagogy. Item mapping should be associated to the curricula and to empirical testing.

The right to education must be something measureable, along with learning attainment, school conditions, and teaching quality; measuring students' results is not sufficient. Moreover, results should be inclusive, broad, and relevant. Measuring should include monitoring and adequate improvement policies.

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

Lorena Meckes

Since 1988, Chile has had extensive experience in standardized testing. Using results for accountability purposes as well as for formative purposes is not easy. It requires a greater number of evaluations for greater accountability, which also limits and complicates the formative aspect of the test. Recently, intensifying accountability measures with high-stakes implications have been unpopular.

The Chilean education system includes twelve thousand schools and over three million students. Chile has a voucher system that provides around US$116 per student in primary and US$138 per student in secondary school, and an extra US$68 and US$45 are provided in cases of vulnerability in primary and secondary school, respectively.

The evaluation system has four main actors: the Ministry of Education, the Quality of Education Agency, the National Council for Education, and the Superintendence. The Quality of Education Agency is responsible for implementing SIMCE, Chile's standardized test, and other tests within the evaluation system.

Chile has a combination of census and sample based assessments. SIMCE is designed to align with the national curriculum and standards for subjects such as language, written and reading comprehension, math, science, history, geography, social studies, English, sports and information and communication technologies. The following grades are evaluated: 2, 4, 6, 8, 10, and 11.

Evaluations are complemented with context questionnaires answered by teachers, students and parents to better understand and interpret the results. Tests are applied every one to two years; tests that are applied every two years measure value added for greater accountability. SIMCE results are published at the school level and are used as a market measure to guarantee the right to access to quality education. Additionally, SIMCE allows for national improvement goals and complements other testing strategies used for impact evaluation of specific programs.

SIMCE began between 1988 and 1990 during the decentralization and privatization of education. SIMCE took the role of a centralized monitoring system for quality control. It was not until 1995, however, that SIMCE results were published. In 2000, the curriculum was re-designed to align SIMCE with the national curriculum.

The state classifies schools into categories depending on the percentage of students that meet the national standards, the relative ranking of schools compared to similar schools, students' performance over time, retention rates, percentage of vulnerable students, and teacher evaluations. Schools with a high proportion of vulnerable students can initially present an improvement plan and immediately receive technical assistance.

High performing schools receive complete funding and are granted greater autonomy – results show that the system has had the most effect in this type of schools. Improvement plans in schools with low and regular performance are monitored, as they can receive greater pedagogic support and more supervision. Part of the resources is withheld until schools show progress. The lowest performing schools run the risk of closure if they fail to show improvement within 4 years.

Support tools for teachers include online workshops that inform different ways of interpreting results from national and international tests. Regular reporting of school learning levels is also issued for both teachers and school directors.

Collective incentives are given to schools serving students from different socio-economic backgrounds. Much information has been gathered regarding the socio-economic learning gaps between students and these differences have been decreasing over time.

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

Jakob Wandall

Denmark currently has an "adaptive testing" system, which can assess individual learning gaps. The system was first implemented in 2010 with an investment of US $20 million. Research is well

advanced in the design of standardized tests and test results analysis, but further research is needed on the most effective ways to ensure that teachers use past results to improve future results.

The objective of education in Denmark is to create good citizens within a democratic system. The Danish education system includes 60,000 students in 2,000 public schools. All students above 3rd grade have a computer connected to high speed internet; for 3rd grade or below there are 3-5 students per computer. The evaluation system consists of a summative assessment after 9th grade, and a formative assessment within the national testing system, which includes the following mandatory tests: reading for grades 2, 4, 6 and 8; math for grades 3 and 6, English for grade 7; and physics, geography and biology for grade 8.

Each test has three content areas. Tests do not have direct implications for any school actors. This design requires very large item-pools with the right mix of items of high and low difficulty. There is a minimum of 540 items per test (180 per profile area), and they are evenly distributed for different difficulty levels. All items should be tested on 500-700 students.

The results so far tell us that the distance between the 10% best and the 10% worst performing students in the class is equivalent to 6-8 years of formal education. While schools have some effect on the performance of test scores, the student background determines, to a large extent, the final outcome.

Adaptive assessment allows targeted differentiated instruction according to the specific needs of each student, specially catered to the different levels of students' academic achievement. 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). Results are confidential – only teachers can know the results of their students. The teacher gets results online including students' scores and details of the tests and access to the items/responses in the individual tests.

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. It is therefore essential that teachers understand how to interpret and use the information of the national standardized tests. Teachers should combine data provided by standardized tests with innovations in pedagogical practices that make a difference in the classroom. Finally, technology can play a very important role for improving assessment systems through web platforms with assessment embedded in training and other learning platforms.

Teachers and school directors are not held accountable through the Danish national system of evaluation; rather, this occurs through the democratic culture that exists within the schools. For example, school directors are elected directly by students and teachers, and other teachers can fire their peers. In addition, parents have the right to dismiss teachers if they do not fulfill their responsibility.

Session 5: The USA Experience. A national policy interpreted and implemented differently in each State

Daniel Koretz

A perspective of the common pitfalls that arise with high-stakes student standardized testing was

presented, along with a set of recommendations that have been learned from the US experience. The US high-stakes standardized testing approach has not necessarily led to consistent improved levels of learning, however, many studies have shown common pitfalls that arise from the design of these systems. A clear message form this experience is that teacher training and other support systems can be instrumental in translating test results into improved performance. Additionally, measuring progress is as important as measuring outcomes.

There are important lessons to be gathered; accountability-strengthening measures in the evaluation system should not be linked to high-stake implications for teachers and school directors. Experience in the US has illustrated that this linkage can lead to distortionary incentives, thus affecting the theory of change.

Policies using test-based accountability to improve pedagogic practices have been used since the 1970s 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 for every grade and every child. Different kinds of assessments have been tried in different states, including performance based assessments; standards based assessments; tests with current/raw scores or value added measures; and assessments that trigger different rewards/sanctions for teachers or students. The commonality around all these models is that American teachers feel pressure to raise students' scores.

Performance targets have often shown to be arbitrary, they are set to be uniform across different contexts, and unrealistically high. Overall improvements tend to be exaggerated, and relative effectiveness is measured incorrectly. This system introduces unintended strategic behavior among teachers and school principals as they attempt to reach the targets, in the form of teaching to the test, score inflation, cheating and other "gaming" strategies.

Result inflation, for example, happens for many reasons. First, tests are predictable. For instance, in math the same 42 test items have been used for 10 years. In most occasions, the content that can be more easily tested ends up being over emphasized in standardized tests relative to other contents that are harder to test/measure, which tend to get de-emphasized. Partly for technical reasons and for ensuring standardization and comparability of the tests from year to year, items tend to be repeated constantly.

The consequences of incomplete item sampling with low pressure are systematically incomplete evaluation of education, modest effects on scores, measurement error, fluctuations in scores, and differences in results among tests. Once pressure is added, the consequences are: distorted effects, coaching (focusing instruction on presentation, rubrics, and incidental test content), reallocation (narrowed instruction to focus only on the content that is emphasized on the tests at the expense of other relevant content), and additional time dedicated to test preparation (focus on specific items) and cheating.

This led to the following recommendations:

1. Couple evaluation and accountability with training and support, particularly for teachers.
2. Evaluate other outcomes by making student assessment broader. Evaluate practices and processes as well as outcomes. Subjective measures may need to be combined with objective measures.

3. Use summative tests appropriately. Set realistic targets to lessen distortionary incentives for teachers and report scores in the form of scales, not performance standards.
4. Design test for accountability uses. Avoid excessive narrowing and predictability in summative tests.
5. Design formative tests differently from summative tests.
6. Evaluate the evaluation system.

Open questions still remaining: if high-stakes were removed from the standardized tests, would the problems referred in the presentation continue to exist? Where is the "sweet spot" between high and low-stakes testing?

Session 6: Navigating between large-scale assessment and classroom assessment

Mark Wilson

The presentation sought to understand the linkages that exist between classroom assessment and large-scale assessment. 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, thus leading to a vicious cycle 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.

A solution that is presented considers a formative approach towards teaching, where theories of learning affect the linkages that exist between curriculum, instruction and assessment. The proposed three part strategy involves: (i) build classroom assessments with one construct at a time, (ii) assemble these into learning progressions, and (iii) develop large-scale assessments by having a structured sample of learning progressions, and test the learning progressions that can become a derived measure of learning.

There are some constructs that are dependent on others, but there is room to test the patterns that become most successful in shaping learning progressions. 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.

The resulting learning progressions can give a perspective of how a typical student can most easily learn a domain of knowledge. Assessments can then help identify where each student lies in the learning progression (knot point) and what types of teaching practices can help the student progress through the next construct and improve learning.

Classroom experience is built around argumentation/debates within small groups. The argumentation practice side of this proposal is different to the learning progression, which focuses more on specific content and outcome progression.
Learning progressions can provide a way to have a healthy relationship between large-scale assessments and classroom assessments. However, this requires an outcome progression and a construct map and the need to rethink blue prints for large-scale assessments. Challenges involve having a curriculum debate and converting it into a testable hypothesis so that experts can test alternative hypotheses of structuring curricula. This opens a whole new research agenda on learning progressions, test alternatives, and alternative outcome progressions. An open question remains: can

we test fairly given large heterogeneity on learning progressions?

Session 7: Teaching and school based management based on information

Arend Visscher

The theory of change for data-driven decision making at the classroom level is the following: implement evaluations, analyze the results, set smart and challenging goals, determine the strategy for goal accomplishment, execute the strategy, monitor its implementation and, finally, evaluate again to see if there were improvements.

However, it is important to assess where differences in performance come from. In the Netherlands, for example, variability in performance is 4 times more at the classroom level than at the school level. Most differences in student performance are accounted for by students' characteristics, such as socio-economic background; the second most important determinant is teachers.

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

1. **Analyzing the relevant data –** this includes value added achievement gains and various analyses of student performance that allow specialists to separate content that is mastered by students and compare ability growth in students vis-a-vis the national average. It is a good idea to complement this with classroom observations and students' perceptions.
2. **Goal setting** – ideally focus on numerical goals.
3. **Choose a strategy for goal accomplishment –** it is key to differentiate between basic didactic skills and complex ones, in order to differentiate instruction.
4. **Implementation –** it delves into what happens in the classroom and how data is used.

Ideally schools can move to a scenario where small 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 the following: what happens in a classroom that enables teachers to use all information? Does it improve quality in education, is there a significant effect on learning outcomes?

The underlying theory of change of how data will lead to improved performance is key. This can illustrate how the additional information will cause change. It is important to trace the types of improvements that can come from within the school, and the types of improvements that will come from other stakeholders external to the school.

Greater emphasis should be given in creating systems that support teachers to identify problems in instruction and help them solve those problems. Good quality teachers are necessary for this approach to work. Technology can also be conducive to this approach.

Open questions: is there an emphasis towards providing greater teacher accountability and informing policy makers of the status of education in order to better allocate resources in Mexico? Does this go somewhat against the approach experienced by the Netherlands? The model for the Netherlands may be an effective model in contexts where there are very good teachers, but can this approach work for low performing schools (with low performing teachers)? Can it be useful for Mexico, knowing that quality of teachers is one of the most important challenges? Additionally, how

can the Netherlands's approach help countries build a culture that promotes data driven accountability without recurring to the traditional accountability strengthening measures like the design of carrot and stick incentives?

Session 8: Proposed design for the second generation of standardized student assessments in Mexico

Eduardo Backhoff

Mexican experience

INEE's current view is that evaluation should help guarantee the right to quality education for children, youth, and adults in Mexico, and this entails improving processes that lead to higher results, in the context of Mexico's cultural diversity.

The main objective of the evaluation strategy is to improve student learning, teacher practices, decision-making, and provide useful information to schools, teachers and parents to strengthen accountability in the education system. The most valuable purpose to fulfill in evaluation is to improve pedagogical practices and the experience within the classroom. Evaluations should be designed to give better information to support students' learning process and to personalize their learning experience.

A new proposed design

Sample based testing should strengthen accountability by assessing the status of education of the system as a whole. INEE proposes that these tests should be applied at the end of the school year. End users would be public authorities and the society in general. INEE should be responsible of the entire process behind these tests: design, application, grading and communication of results. Authorities and schools should support INEE throughout implementation. INEE should document all the stages of the process, and prove the validity of the results.

Census based testing should focus on improving the pedagogic and formative processes. INEE proposes that these tests should be controlled and administered by the schools at the beginning of the school year (forth, seventh, ninth, and twelfth grades). End users would be school directors, teachers, students and parents. These tests should contain the same constructs as the sample based tests. Education authorities should be responsible for printing, implementing, grading, analyzing, and communicating results. The Ministry of Education (Secretaría de Educación Pública, SEP) and the INEE should ensure quality control throughout the process and certify that the assessments are being applied correctly (both standardized student tests and supervision in situ).

Strengths of the design: The overall quality of the standardized test would improve by strengthening the control measures and protocols established by INEE. Schools will be evaluated once every three years in two different grades. A yearly evaluation will be applied to all students through testing administered by the schools.

Weaknesses of the design: Census based tests will not be completely controlled by a central authority, unless education authorities agree to finance and plan for this. In the years and grades of the census-based test, results take a long time to get to the schools.

There is evidence showing that the accountability effects of standardized testing has had limited impact. INEE seeks to design an evaluation system that will mainly serve to improve learning, giving greater emphasis to a formative assessment. Ideally, the new generation of standardized tests in Mexico should provide two types of information: (i) information to strengthen accountability, by giving timely feedback to improve education policy; and (ii) information that allows school directors and teachers to improve management and teaching practices.

Open questions include: how will the information from the assessments be reported to different stakeholders, and how will this cause change? Why is it important to test at the beginning of the year as compared to the end of the year? Would results of the census-based test reach the schools in time to help them identify problems and implement solutions? What support systems will be in place so teachers can make use of the information? How can INEE make sure that the results would positively affect pedagogic practices? What implications does the new evaluation system have with regard to curriculum design? Why is the proposed design so different from ENLACE, as opposed to taking what was already working and building from there? What measures will be put in place to ensure legitimacy of the new evaluation system? Will there be engagement groups that can champion the new initiative?

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

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.

Goals

Part of the problem is that there is not one system that can serve all purposes. Brazil, the US, and Chile provided three examples of different and important purposes: assessments made to inform policy makers while monitoring the performance of the education system as a whole; assessments with diagnostic and formative functions that can better guide instruction in the classroom; and internal and external assessments that strengthen accountability in order to give better incentives for better performance. The important point is to decide what will be the priorities of the standardized test and how this will affect its design.

The evaluation system as a whole

It is crucial to think about how the information from the standardized student assessments will improve learning. What support mechanisms and training systems will be put in place in order to make this process effective? How can we evaluate the evaluation system, given that it is useful to think about dynamic accountability?

When considering the accountability purpose of the test: What type of information will improve accountability? What decisions will be affected by the results? In what way can society make sure that there are gains at the school level? What is the underlying theory of change? How extreme will the design favor strengthening accountability, given that it holds a substantial trade-off with a

design that seeks to fulfill formative purposes? What happens if you design standardized testing taking into account these two main purposes simultaneously? How can too much information dissemination affect school autonomy and what is the best way to account for this in the Mexican context?

Other remaining questions: can students be followed through time for a longitudinal analysis? Can the design of the evaluation system be set through an intense period of pilot testing, lasting around 10 years like the case of Singapore? How can the design of the evaluation system promote a culture of evaluation? How can parents be properly engaged with the use of the results from standardized testing? Are results inflation and teaching to the tests real concerns for the Mexican context?

Session 10: Concluding thoughts

Sylvia Schmelkes and Reema Nayar

There is great food for thought at the conclusion of the Workshop. The key aspect to consider will be how to bring these lessons to bear when designing a concrete proposal. What was gathered from the sessions is that evaluation has not developed as a proper science, there is not one solution, and it would be good to know if there is still room for piloting.

When considering the second generation of standardized student assessments in Mexico, the first priority of the evaluation system will be to improve learning by focusing on what happens in the classroom. Second, assessments will be designed to identify gaps and to work through policies in order to eliminate these gaps? Third, assessments should be a tool to strengthen accountability; INEE should allow society to know what the state the education system is in and hold stakeholders accountable. Finally, assessments should enable impact evaluation of education policies.

Evaluation is not going to be an automatic process, and it is important to think about how information will be used for public policy decision-making. Important questions remain to be considered, such as how teacher training will be affected by evaluation and how this agenda will change the curriculum.