THEDIALOGUE
Leadership for the Americas
WORLD BANK GROUP
COMMISSION FOR QUALITY EDUCATION FOR ALL: BACKGROUND PAPER
SEPTEMBER 2015

WHAT HAS HAPPENED WITH LEARNING ASSESSMENT SYSTEMS IN LATIN AMERICA?

Lessons from the Last Decade of Experience

Guillermo Ferrer and Ariel Fiszbein

Introduction

The development of learning assessment systems in Latin America received a significant push starting in the 1990s as a part of efforts to modernize education systems—often with strong support from the World Bank and other international organizations.

The Inter-American Dialogue, through the PREAL initiative, both supported and monitored the growth and development of those systems. In a volume published by PREAL, Guillermo Ferrer (2006) took stock of the state of learning assessment systems in the region around 2004-2005. The report recognized the significant progress countries in Latin America had made in establishing evaluation systems, but also revealed strong disparities in the levels of development between different systems, depending on the nature of the organizations responsible for the evaluation functions and the political context in which they operated. The report suggested that institutional arrangements were more stable when responsibilities were outside the organizational structure of the Ministries of Education through institutes with greater administrative and technical autonomy. According to the report, an area of special weakness for the operation of learning assessment systems was proper dissemination and use of results. In particular, the report argued that low civil society participation in the evaluation processes failed to promote demand for higher levels of quality in education.

The decade since the report was published has seen many important changes in the region. First, many countries have undergone significant changes in institutional arrangements related to learning assessment functions. For example, Ecuador established, as part of a new national constitution, its Instituto Nacional de Evaluación Educativa (Ineval) in 2012, transferring responsibilities previously in the hands of the Ministry of Education (MOE). In Mexico, the Instituto Nacional para la Evaluación de la Educación (INEE), which was created by a Presidential Decree in 2002, acquired a stronger and more formal legal status in 2013 through a new national education law and a constitutional reform.

Second, there have been important innovations in the communication of results, including Brazil's index of education performance (IDEB) available at school level and, similarly, the results of the SABER tests in Colombia, both discussed below. And finally, the participation of Latin American countries in international tests has also expanded over the last decade.

Table of Contents

ACKNOWLEDGEMENTS

This work was made possible through the support of the World Bank and its Russia Education Aid for Development Trust Fund Program – READ (available at: http://www.worldbank.org/en/programs/read). The authors benefited from useful comments by Marguerite Clarke and Gilbert Valverde, and the support of Marianela D'Annunzio, Maria Oviedo, Ben Raderstorf, Laura Ruberto, and Federico Sucre during the preparation process.

Perhaps surprisingly, very little has been written about these changes. Moreover, not much is known about the effects that changes in evaluation systems may have had on education systems and quality of education more generally.[1] For example, while the so-called 'Metas Educativas 2021'—a list of goals set forth by members of the Organization of Ibero-American States (OEI)—included a specific target related to the 'strengthening of national evaluation systems', the first monitoring report (OEI, 2012) on those goals explicitly excludes that particular goal due to a lack of information. It should also be noted that none of the 21 countries for which a World Bank Systems Approach for Better Education Results (SABER) report on Student Assessment Systems is available are in Latin America.[2]

We believe there is great value in addressing this information gap. This applies both for countries in Latin America—which can learn from a decade of often-intense change—and those elsewhere in the developing world, particularly in light of how Latin America was the first region of the world to undergo a 'big, concerted, push' to establish learning assessment systems. The increased attention to measuring learning outcomes in the context of the post-2015 discussions on international development goals (United Nations, 2014) and the ensuing need to strengthen national efforts to assess learning makes this a timely and, hopefully, useful exercise.

This paper is the result of a (partial) update of the 2006 report, carried out over the period of three months, with the purpose of offering a critical review of progress made and lessons learned from the experience of the last decade. Rather than reviewing progress in all the seven dimensions covered in the original analysis[3] we focused only on three dimensions:

(i) the institutional and organizational framework of learning assessment systems;
(ii) the dissemination and uses of information (to/by authorities, school actors, families, civil society); and
(iii) participation in international tests.

Given the limited time available for this exercise, we focused our attention on the experience in five countries: Brazil, Colombia, Guatemala, Mexico, and Peru (see **Table 1**). Our selection of country cases seeks to capture the experience of some of those countries that have made significant investments[4] in learning assessment systems

TABLE 1: BASIC EDUCATION INFORMATION

	GDP PER CAPITA, 2013 (IN US$)	% GDP SPENT ON EDUCATION, 2013	NUMBER OF TEACHERS	NUMBER OF STUDENTS
Brazil	11,208	6.1*	2,515,648	46,512,407
Colombia	6,807	4.9	430,410	10,927,876
Guatemala	3,478	2.8	228,072	4,167,443
Mexico	10,307	5.1*	1,932,061	32,077,094
Peru	6,662	3.3	493,284	9,282,045

*Data from 2011.
Sources: GDP per capita and % GDP spent on Education: World Bank Data. Exception: Brazil's % GDP on education was obtained from OECD Education at a Glance Country. Number of Students: UIS Databank. Unless otherwise noted, data is for 2013. It includes students under ISCED 0 to 3. Number of Teachers: Peru: "Magnitudes" Database of ESCALE (Estadística de la Calidad Educativa) of the Ministry of Education of Peru. Includes teachers from EBR (Educación Básica Regular) in preprimary, primary and secondary school in 2014. Guatemala: CIEN and Inter-American Dialogue (2015). El estado de las políticas públicas docentes, Guatemala. Includes pre-primary, primary, and secondary teachers at all types of institutions. Mexico: SEP (2015). "Descarga serie histórica y pronósticos de la estadística del sistema educativo nacional." Web. Secretaría de Educación Pública. Includes teachers in preprimary, primary and secondary school in all institution types. Brazil and Colombia: "Teachers by teaching level of education." [Dataset]. UIS UNESCO Institute for Statistics.

during this period. At the same time, we had an interest in highlighting a diversity of country conditions. In particular, we sought to cover both countries in which evaluation functions are the responsibility of institutes that have an arms-length relationship with the Ministries of Education, and those in which the function is performed by units operating within the formal structure of the Ministry. The countries covered also represent some regional diversity (including countries in North, Central and South America). The case of Guatemala is useful as an example of a country with a history of heavy reliance on international financial and technical support.

Over the last decade large-scale, standardized learning assessment systems in Latin America have gained visibility, strength, and a prominent role in the overall discussion about educational quality.

For these countries we conducted an in-depth and comprehensive review of official materials (accessed through official websites)[5] and a set of semi-structured interviews with twelve experts[6] knowledgeable of the operation of learning assessment systems in these countries. An effort was made to interview both individuals that played a key role in the management of these systems (in all cases now not directly involved with those systems in an official capacity) and individuals that represent independent civil society, academic, and technical voices in education evaluation.

Beyond the five countries of special focus, our review involved analysis of information regarding participation in international tests for the entire region. The assessment also benefited from our familiarity with the experience of other countries in Latin America and the Caribbean for which we did not conduct a review of official documentation or interviews.

This was a short-term project, and we are conscious of the limitations of our analysis and the need to collect more information to further test our hypotheses. In all likelihood, the most consistent observation that can be made at this point, by way of introduction, is that over the last decade large-scale, standardized learning assessment systems in Latin America have gained visibility, strength, and a prominent role in the overall discussion about educational quality. Overall, the trend in Latin America is, beyond doubt, to keep supporting the work of national assessment units[7] and to strengthen their institutional and technical components.

However, this progress has not been automatic. In order to achieve this strengthening, countries struggled to firmly maintain both the form and content of what already existed. And in some cases, as described below, they made changes and adjustments to maintain stability and continuity in times of adversity. This observation applies principally and more directly to the first of the three major aspects of assessment dealt with in this report, i.e. the institutional framework for assessments. However, it also has direct implications for the use of data and results, as well as for countries' participation in international assessments.

Learning Assessment Systems in Latin America: Changes and Continuities, 2005-2015

In this paper we set out to assess the changes and continuities in three key dimensions of learning assessment systems in Latin America, taking as a reference the findings of the 2006 report. The first dimension we consider is the institutional framework behind the evaluation function and its effects on the credibility and legitimacy of assessment activities, as well as the technical and operational sustainability of those activities. This dimension of our analysis corresponds to what the SABER-Student Assessment framework refers to as 'Enabling Context'.[8]

The choice between a model that relies on autonomous (or semi-autonomous) agencies and one that uses specialized units within the MOE to lead and administer education evaluations is the most notable area of focus. The second dimension we consider is the reporting, dissemination, and use of data. The influence of learning assessments on the quality of education depends critically on the ways in which results are presented, information is shared and findings are used by schools, teachers, administrators, parents and civil society at large—and we seek to understand the changes that took place on those aspects. This dimension of our analysis corresponds to what the SABER-Student Assessment framework refers to as 'Assessment Quality'. Finally, we consider the participation in international tests both as an indicator of a country's commitment to evaluation and as a factor influencing in-country evaluation capacity.

INSTITUTIONAL FRAMEWORKS

In the five countries leadership and administrative responsibility for standardized assessment programs rests with two major types of institutions: national institutes and assessment units in Ministries of Education (MOE, or the Secretaría de Educación Pública, SEP, in the case of Mexico). As it happens, the case of Mexico currently presents a situation that can be characterized as a *mixed format,* which involves a national institute–the Instituto Nacional para la Evaluación de la Educación (INEE)–and the assessment unit within the SEP. The other two national institutes are the Instituto Colombiano para la Evaluación de la Educación (ICFES) in Colombia and the Instituto Nacional de Estudos e Pesquisas (INEP) in Brazil. MOE's units are the Unidad de Medición de la Calidad Educativa (UMC) in Peru and the Dirección General de Evaluación e Investigación Educativa (DIGEDUCA) in Guatemala.[9] All institutes and MOE's units already existed in 2005, with the exception of DIGEDUCA, which was already operating then but became officially constituted by law in 2007.

A number of changes have occurred in national institutes, mainly in terms of new legal arrangements which give institutes more administrative autonomy. ICFES, for example, now has more formalized responsibilities to offer testing services, collect revenue, and as a result of reforms introduced in 2009, retain surpluses to be reinvested in technical and program development.

Mexico's INEE underwent a major institutional change when the organization, originally created by a Presidential Decree in 2002, became ratified by law through a constitutional reform approved in 2013. This change gave INEE not only more autonomy but also more responsibilities, including the design and leadership of a National Evaluation System, which involves the assessment programs administered both by the Institute and the SEP.[10] It should be noted that, despite the fact that INEE's operations were supposed to replace the SEP's evaluations since 2002, what actually occurred is that the SEP became stronger, as well as much more visible, with the implementation of a census-based assessment that rapidly captured national attention, while diminishing INEE's role and influence in leading the educational evaluation functions. Due to a number of technical and political reasons, the INEE has now regained both power and legal authority to coordinate a National Evaluation System which includes all assessment activities, both within the Institute and those carried out by the SEP. How this arrangement—in particular the balance between autonomy and accountability—will play out during the next few years will be an important subject for future research.

Brazil's INEP has been an autarchic[11] organization for many years and, while attempts to establish it as a completely independent agency were not successful due to a legal vacuum concerning public agencies, the Institute still has vast discretional power for self-government and administration.

In all three cases, though, the head of these institutes is selected by the MOE and appointed by the national President (Colombia and Brazil) or by the Senate (Mexico). In short, all institutes have kept functioning during the last decade, and their autonomy and responsibilities have widened and strengthened either by sheer practice or through new legal provisions.

Units like UMC and DIGEDUCA have continued to operate within the organizational structure of the ministries. There is still some debate as to whether these units should be turned into autonomous agencies outside the MOEs. However, the technical capacity and transparency they have demonstrated over the years have made the question of institutional change an issue of lesser priority than could have been expected a few years ago. In both Guatemala and Peru evaluation units operate almost entirely with funding from the national budget rather than from international sources–something that did not happen in the past. As a result these units have become less and less dependent on international aid and credits. This growing independence from external agencies denotes the presence of national political will–for a number of possible

reasons–to keep these units fully operating, even when assessment results may be low and represent "bad news" for political authorities and policymakers. This was not an obvious prediction ten years ago.

Even though institutes' autonomy is not questioned, and is actually seldom politically challenged, there is an overt and permanent demand that their assessment work become more closely aligned with government educational policies, in particular those related to curriculum implementation and teacher professional development. However, it should be acknowledged that, despite having their own assessment units, MOEs also often struggle to manage an adequate internal flow of data to improve decision-making based on student learning results. In fact, it is fair to say that neither type of institutional configuration can guarantee *per se* an efficient use of the data for informed policymaking and educational improvement.

Even though institutes' autonomy is not questioned, and is actually seldom politically challenged, there is an overt and permanent demand that their assessment work become more closely aligned with government educational policies.

One aspect that does appear to differ between institutes and units is the possibility of designing and effectively carrying out mid- to long-term evaluation plans. Institutes tend to make long-term decisions about assessment programs (including population and content coverage, test administration cycles, etc.) and stick to them for as long as they consider those programs useful and pertinent. MOEs' units, on the other hand, find it much more difficult to maintain long term assessment plans and cycles due to frequent—sometimes erratic and arbitrary—demands by political authorities, particularly when government changes occur. Institutes are also likely to receive these types of demands, but they also have more power and legal authority to discuss the convenience of changes and, ultimately, decide whether or not they are to be implemented. In the case of Mexico, under the new legal and Constitutional framework, INEE may now not only decide freely on its own evaluation programs but also dictate what the SEP can or cannot do in terms of educational assessment.[12]

There are at least two more political factors that might—in principle—distinguish institutes from MOEs' units:

a. One of them relates to these organizations' "sensitivity" to political interference in handling public data, namely assessment results. However, during the past decade there has been no difference in this respect between countries with different institutional frameworks. Already in 2005 the results were always made public, and so it remains. The last two salient instances when results were censored were around the year 2000, when Mexico and Peru banned the publication of international or national assessment data. From 2005 to the present, there have been no reports on attempts to censor data dissemination in any of these countries and therefore the autonomy or government-subordination factors have not been put to the test in this regard.

b. Teacher unions, the other important political actor and historical detractor of standardized evaluations, have played the expected role of opposition and criticism. But, again, this was independent of the institutional framework of the assessment function. When they have acted or made public their position against evaluations, or when they have agreed to avoid confrontation in return for an increase in compensation, they have done so in direct conversations with higher political authorities. The interviews conducted for this report suggest it is Ministers of Education and Presidents who take the lead in negotiations; institutes or MOEs' assessment units are not the targets or objects of political disputes regarding educational evaluation.[13]

Another organizational aspect that could conceivably involve a distinction between institutes and MOEs' units is their differential capacity to create and develop solid and stable technical expertise. The size and the administrative autonomy of institutes might lead us to think that they have more opportunities and capacity to recruit and maintain a permanent and increasing number of experts in their ranks. The truth, however, is that, whereas institutes do have more administrative latitude and budget stability to recruit new personnel, the pool of qualified professionals for jobs in highly specialized positions is quite reduced in all these countries. Therefore, both institutes and MOEs' units face a permanent challenge of selecting training, and retaining adequate candidates. Eligible professionals usually come from the fields of psychology, sociology, statistics, and economics (seldom are they educational specialists with a strong pedagogical background). Many of them, after gaining knowledge and experience in educational measurement, move on to other areas of government, to private companies or to pursue graduate studies abroad. Additionally, schools of education in universities in the region lack sufficient (high quality) programs in statistics and educational measurement. Perhaps an exception worth mentioning is Brazil's INEP, which despite difficulties in recruiting new and well-trained staff, at least offers job stability under local public administration laws, an incentive appreciated by many prospective or recruited candidates.

REPORTING, DISSEMINATION, AND USE OF DATA

Institutional frameworks, as described in the previous section, are important to analyze inasmuch as they may affect the production and use of the data, as well as how they reach intended audiences and how they can be publicly accessed. This section presents a description of the different ways the data gathered through learning assessment are reported (i.e. the models and formats used for presenting the data), how they are channeled to different audiences, and how they are used or expected to be used to improve educational quality.

All the countries we consider in this paper have a combination of high-stakes and low-stakes evaluations. High-stakes evaluations, intended by definition to have direct consequences for individuals or institutions (students, teachers, schools), greatly determine how regularly the evaluations take place, and also how rigorous and timely the delivery of results can be. Clearly, evaluations (or rather exams) like SABER 11 in Colombia and ENEM in Brazil, both of which have consequences for students' access to higher education, are always administered according to a rigorous calendar, and their results delivered to schools and students on fixed, previously announced dates of the academic year. There are other assessments, which, although increasingly used in practice to provide economic incentives to schools based on educational progress, do not require such an exact calendar of administration. This would be the case of Peru's census-based evaluations in primary education, Brazil's Prova Brasil or Colombia's SABER 3-5-7-9.[14] The latter cases were not intended originally to be high-stakes assessments but have progressively turned into high visibility evaluations that put the spotlight on schools, with the potential to create social and political accountability.

Finally, there are assessments that have no direct consequences for schools or individuals, mostly sample-based or voluntary, such as Provinha Brasil, Guatemala's evaluations in primary education, or Mexico's EXCALE evaluations developed by the INEE.[15] In these cases, though, the administration of tests is also quite regular, in fact more so than could have been expected a decade ago. Nevertheless, the regularity of these tests depends on the assessment units' calendar, technical possibilities, regulations by law, and even political will.

Regardless of the nature of the evaluations the format in which results are presented follows a common pattern.[16] Ten years ago, most countries in Latin America had already transitioned from Classical Test Theory (intended for comparing, selecting, or ranking individuals) to Item Response Theory, which allows for the reporting of results in a way that is much more descriptive of what students know and can do, and how basic or advanced their knowledge is. In other words, regardless of whether today's evaluations are meant to compare, classify, or rank students in a high-stakes evaluation scenario, or whether they are meant to provide feedback about the quality of their learning, they always present achievement scales, split into performance levels, showing not only the percentage of students in each segment but also a substantial description of what students know and can do at each level of the scale. Certainly, this methodology and reporting format has much more potential for pedagogical interventions than a report on the average percentage of questions answered correctly. It is worth noting that a modern and much more sophisticated methodology has remained in place over the years, as well as the fact that the trend is attuned to the latest methodological developments in evaluation worldwide.

At the same time, high-stakes assessments are usually census-based. This has implications for the analysis of educational quality, especially in terms of curriculum coverage, i.e. how much of what is taught, or supposed to be taught, can actually be captured by the tests. As is well known, census-based assessments cover large populations of students, and therefore cannot go very deep in the coverage of curriculum content. They are thus characterized, in assessment jargon, as "thin" evaluations.

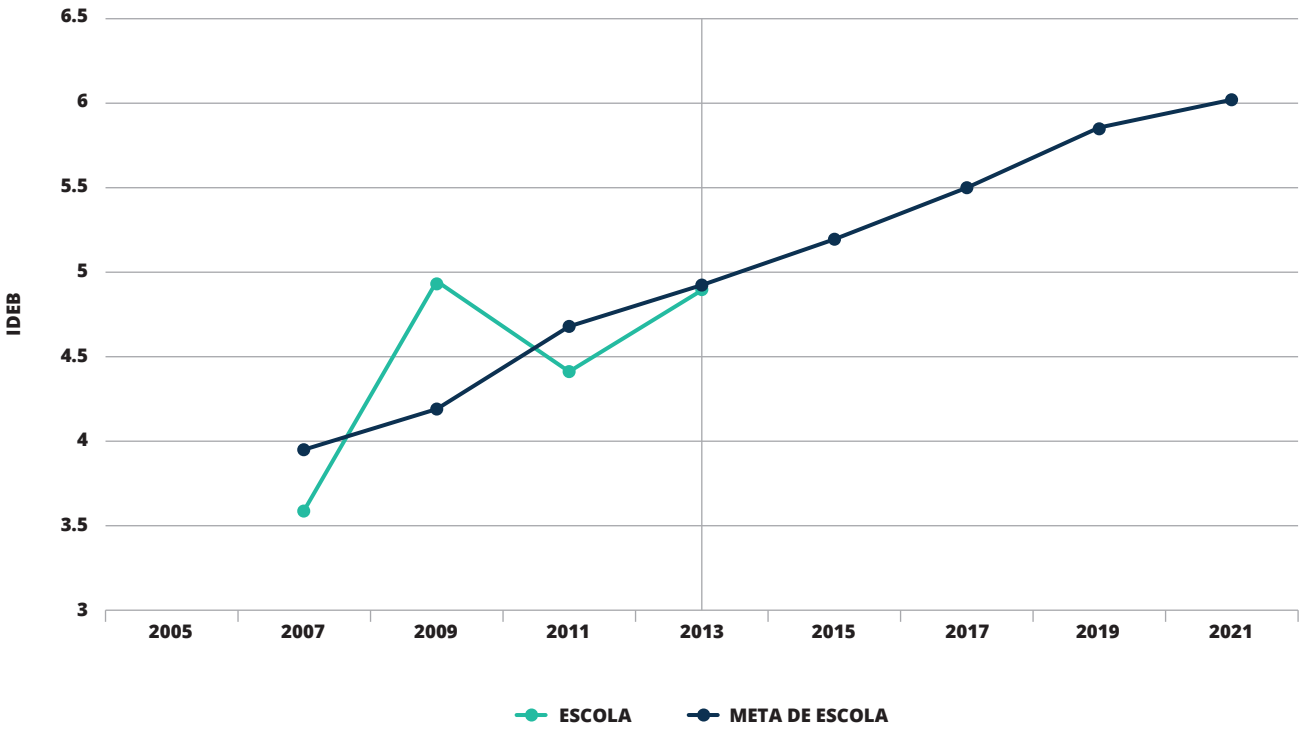
This carries negative implications for two reasons, both emphatically expressed by experts in the field. In the first place, as indicated earlier, the results do not allow for a deep analysis of the so-called attained curriculum. Data in this respect are simply too scarce. Second, census-based assessments, "thin" as they are, do not allow for rigorous and truly useful analyses of learning associated factors,[17] such as school characteristics, students and families' cultural capital or teaching methodologies.

Nevertheless, over the last decade there are a number of examples of how census-based evaluations can be used for the development of educational quality indexes, such as Brazil's Índice de Desenvolvimento da Educação Básica (IDEB) or Colombia's new Índice Sintético de Calidad Educativa (ISCE). These represent one of the major and latest developments in terms of data use in the region (See **Graphs 1** and **2**).

In both cases, indexes have been developed with two principal purposes. First, to attain a measure of educational quality which takes into account at least three different variables, not only learning achievement. Brazil's IDEB combines learning results with students' passing and drop-out rates, while simultaneously measuring progress in this index year by year, in every primary school of the system.[18] Colombia's ISCE considers student achievement together with passing and drop-out rates, as well as classroom environment. This index also measures school-by-school yearly progress. In both cases, learning achievement data are provided, respectively, by the national evaluations Prova Brasil and SABER. Indexes also allow policy makers to have a measure of progress when they establish mid- to long-term attainment goals. For example, Brazil has made public its intention of increasing annual results, in terms

GRAPH 1: IDEB EVOLUTION FOR A RANDOM SCHOOL

Source: Índice de Desenvolvimento da Educação Básica (IDEB) - Development Index of Basic Education. Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira. (http://www.qedu.org.br/escola/1186-escola-absolon-moreira/ideb)



GRAPH 2: IDEB RESULTS AND GOALS

Source: Índice de Desenvolvimento da Educação Básica (IDEB), Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira. (http://ideb.inep.gov.br/resultado/)

IDEB: 2005, 2007, 2009, 2011, 2013, Projections for Brazil

ANOS INICIAIS DO ENSINO FUNDAMENTAL										
	IDEB OBSERVADO					METAS				
	2005	2007	2009	2011	2013	2007	2009	2011	2013	2021
Total	3.8	4.2	4.6	5.0	5.2	3.9	4.2	4.6	4.9	6.0
DEPENDÊNCIA ADMINISTRATIVE										
Estadual	3.9	4.3	4.9	5.1	5.4	4.0	4.3	4.7	5.0	6.1
Municipal	3.4	4.0	4.4	4.7	4.9	3.5	3.8	4.2	4.5	5.7
Privada	5.9	6.0	6.4	6.5	6.7	6.0	6.3	6.6	6.8	7.5
Pública	3.6	4.0	4.4	4.7	4.9	3.6	4.0	4.4	4.7	5.8

ANOS FINAIS DO ENSINO FUNDAMENTAL										
	IDEB OBSERVADO					METAS				
	2005	2007	2009	2011	2013	2007	2009	2011	2013	2021
Total	3.5	3.8	4.0	4.1	4.2	3.5	3.7	3.9	4.4	5.5
DEPENDÊNCIA ADMINISTRATIVE										
Estadual	3.3	3.6	3.8	3.9	4.0	3.3	3.5	3.8	4.2	5.3
Municipal	3.1	3.4	3.6	3.8	3.8	3.1	3.3	3.5	3.9	5.1
Privada	5.8	5.8	5.9	6.0	5.9	5.8	6.0	6.2	6.5	7.3
Pública	3.2	3.5	3.7	3.9	4.0	3.3	3.4	3.7	4.1	5.2

ENSINO MÉDIO										
	IDEB OBSERVADO					METAS				
	2005	2007	2009	2011	2013	2007	2009	2011	2013	2021
Total	3.4	3.5	3.6	3.7	3.7	3.4	3.5	3.7	3.9	5.2
DEPENDÊNCIA ADMINISTRATIVE										
Estadual	3.0	3.2	3.4	3.4	3.4	3.1	3.2	3.3	3.6	4.9
Privada	5.6	5.6	5.6	5.7	5.4	5.6	5.7	5.8	6.0	7.0
Pública	3.1	3.2	3.4	3.4	3.4	3.1	3.2	3.4	3.6	4.9

of rising index values, up to the year 2021. This is clearly a sign of political accountability, unheard of in Latin America a decade ago.

The second purpose for which the indexes have been developed is to establish incentive programs for those schools that show progress in the index values. Whether or not these incentives have or will have a positive impact on learning improvement is yet to be seen. So far, there is not enough research evidence to prove how adequate or potent these incentives really are. However, the fact that more complex, i.e. less simplistic, measures are being developed is a positive sign of a more substantial use of assessment data in these countries.

What still remains a major problem is an insufficient capacity on the part of educational audiences–especially at the school level–to understand data presented in textual formats they are not used to reading.

To sustain this conclusion, if only by contrast, it is necessary to keep in mind that incentive programs based solely on learning achievement data are more much likely to encourage corruption and data manipulation, as has been the case with Mexico's Enlace assessment administered by the SEP. In the last five years, reports of cheating and corruption have surrounded the Enlace student examinations. Experiments conducted in Mexico by Zúñiga Molina and Gaviria (2010) suggest that the average percentage of probable cheating across all grade levels between 2006 and 2009 was high—ranging from 4.5% in 2006 to 7% in 2008 (OECD, 2011, p. 61).

Additionally, a Review Team from the Organisation for Economic Co-operation and Development (OECD) observed that in many of the schools they visited, some teachers asked their low-performing students to miss school on the day the ENLACE evaluations were administered, while others actively helped students complete the exam (OECD, 2012, p. 126). In 2014, the Mexican Secretariat of Public Education announced the suspension of the ENLACE exams, citing numerous instances where teachers provided students with answers to the exam in advance, or corrected wrong answers in order to boost scores (Martinez Carballo, 2014). The OECD attributes this to the fact that teachers stood to gain salary bonuses if their students performed well on the exam, and that the exam was often used by the media to rank individual schools (OECD, 2011, p. 85).

In spite of the limitations census-based assessments face to enable deeper analyses of learning associated factors, the experience indicates that the "simpler" factor data as collected by IDEB and ISCE are actually being used. In contrast, more detailed associated factors, such as those produced by Guatemala's sample evaluations in primary schools,[19] or Brazil's sub-sample factor data extracted from Prova Brazil, are not having the impact desired in terms of informing policy, or leading to the introduction of changes and improvements for better educational quality. In short, we observe a tendency to use more limited context data for high impact indexes, while deeper analyzes, though more substantial and potentially more useful, are being sub-utilized and reserved for academic audiences at best.

Use and dissemination of data can also be observed in the ways assessment units make information public through their websites. Here the scenario is quite heterogeneous among countries, perhaps not so much in terms of what and how is digitally published, but in the way those websites are organized. Regarding data publication, we observe at least eight different ways of reporting results and explaining how to interpret them:[20] (i) Full reports with national and sub-national results for each assessment administration; (ii) Reports showing longitudinal progress over the past 5-10 years; (iii) Individual reports for students and schools in the case of high-stakes evaluations (typically password protected to ensure confidentiality); (iv) Databases for academic use (research); (v) Secondary analyzes carried out by assessment units themselves or by external researchers; (vi) Pedagogical recommendations based on released items and evaluation results; (vii) In-site

search engines to find results (when not confidential) and to find examples of curriculum content evaluated; and (viii) A variety of audiovisual presentations, including PowerPoint, Prezi and videos.

In spite of the variety of reports found in digital format, it must be said that official websites are often not well designed: it takes time and search efforts to find reports, to distinguish reports of different types of assessments and oftentimes to open or download files due to failures in the system. Both general information and specific data are often repeated in different sub-sites, found under confusing link names, or put together without a clear specification of the audiences intended. These flaws may not be so critical for researchers, who tend to have the will and experience to carry out complex searches. However, if judged as instruments for regular, non-expert users such as teachers or interested citizens, these websites still need much improvement to become more friendly and accessible.[21] It appears as if units' authorities and staff order website managers to load all available information, while exercising little control or supervision over how that information is presented to users.

Our review found that official websites do not provide any data on how much of the information digitally reported reaches audiences in regular, paper editions, and whether they are simply post-mailed or delivered in the context of workshops, on-site teacher training sessions and other similar events.

Regardless of how reports are channeled to audiences, either digitally or in paper, what still remains a major problem is an insufficient capacity on the part of educational audiences—especially at the school level—to understand data presented in textual formats they are not used to reading, and for which they have not been prepared during their studies as prospective educators. The textual formats we refer to are those that include achievement data in a variety of graphic presentations, including charts, bars, scales and comparative tables, among others. Countries have tried different ways of presenting these formats in a reader-friendly manner, resorting to all kinds of drawings, pictures, color references, as well as simplified language and a profusion of examples, both static (as in paper) and dynamic (as in Prezi-like presentations and videos).

However, for several reasons, no format seems to yield satisfactory results in terms of interpretation and use.[22] At least two possible factors may be hindering progress in this respect. First, as mentioned above, local educators (teachers, principals, supervisors), as well as families, appear not to have enough reading and basic mathematical skills to interpret non-continuous, non-verbal texts. Not even on-site delivery and workshops seem to be enough to counterbalance this deficiency in cognitive preparation. Second, it can be assumed that low-stakes evaluations, i.e. when results have no direct consequences for individuals, provide little incentive to make an effort and make out the meaning and relevance of these texts. However, these are the authors' impressions based on first-hand experience and interviews conducted. More research on communication strategies would be warranted in order to understand this problem and find adequate solutions.

Finally, we consider the impact that assessment data currently have on public opinion and society, to a considerable extent mediated by the press. This is probably one of the major signs of evolution in the field of large-scale evaluations in Latin America. All interviewees consulted, in addition to the authors' knowledge of the learning assessment scenario in the region, coincide in their impressions about how much assessment results have helped to install the issue of educational quality–understood in terms of learning results–in national and transnational debates in Latin America. Most national education goals, as expressed in a variety of mid- to long-term governmental plans, as well as international agencies' guidelines and support measures for improving educational quality in the region, include learning results as a key indicator for measuring quality and progress over time.

Most national education goals, as expressed in a variety of mid- to long-term governmental plans, include learning results as a key indicator for measuring quality and progress over time.

Already a decade ago, the press had started to pay attention to these indicators, especially after Latin American countries began to participate in the Programme for International Student Assessment (PISA) and it became clear how far behind countries in the region are in terms of their learning results. Initially, international comparative results were received with great anticipation, mostly with an interest in publishing negative news, demonizing educational systems through headlines, which focused only on rankings, and rarely going into deeper treatment of the causes or the complexity of improving educational conditions. Something similar occurred with national results, especially in those countries that were implementing, or starting to implement, census-based assessments in the final years of secondary education.

Nowadays, although news about results continues to be mixed, media and journalists are becoming more knowledgeable and specialized in educational matters; therefore, they can now go beyond the simple scandal-provoking headlines so frequent and popular in the past. Bad news still makes for negative headlines and triggers harsh criticism of government policies, but at the same time it seems to contribute to longer lasting debates on educational quality in a variety of fora. Probably as a result of this mass coverage, politicians are becoming, if not necessarily more accountable, at least more sensitive to the importance and prominence of learning assessment results and the need to pursue more coherent measures conducive to improving learning outcomes. Backlash in this sense is also predictable, when not occurring already, as politicians tend to make rush decisions in order to show that actions are being taken to change the situation. Thus, the more recent wave of teacher and school incentive programs to increase results could actually have negative, unintended outcomes if not planned and implemented after careful thought and study of similar policies in comparable national contexts.

Probably as a result of this mass coverage, politicians are becoming, if not necessarily more accountable, at least more sensitive to the importance and prominence of learning assessment results and the need to pursue more coherent measures conducive to improving learning outcomes.

PARTICIPATION IN INTERNATIONAL ASSESSMENTS

Conceptually, it could be argued that international assessments should not be treated as a different category from national assessments. However, due to the fact that these assessments are designed and implemented by supranational agencies, their connotation and implications, both in technical and political terms, deserve to be dealt with in a separate section.[23]

Over the period under consideration, Latin American countries have continued to participate in international assessments, with some variations over time regarding the evaluations they choose to be part of, and also in the number of countries taking part in each of the different assessments (See **Appendix 4**). The trends show that (a) an increasing number of countries in the region have decided to participate in PISA and LLECE (UNESCO's Latin American Laboratory for Assessment of the Quality of Education); (b) just two to three countries have taken part in TIMSS between 1999 and 2011; (c) Latin American countries do not show much interest in PIRLS[24] (only one took part in the latest administration while the number of countries globally has increased); and (d) more Latin American countries have become interested in assessing citizenship competencies through IEA's ICCS evaluation (formerly CIV-ED).

The reasons for participating in international assessments, according to interviewees, have remained practically the same as in 2003. First, countries consider it important and necessary to compare themselves both with regional "peers", as in the case of LLECE, and

with highly industrialized countries in Asia, Europe and North America, mainly through PISA-OECD evaluations. Second, national assessment units have found in PISA a major opportunity for developing and strengthening their technical and methodological capacities. Even though Latin American countries do not have a significant role in test design and development, the process of participating in those programs provides multiple opportunities for technical exchange and for acquiring knowledge of increasingly complex measurement methodologies. This was also a strong motivation for joining the first LLECE evaluation in 1997. Nowadays, however, that perception has changed as LLECE does not seem to provide that kind of incentive. Participation in LLECE, though, has not diminished. The reasons for continuity might warrant further research.

One of the most important objections to participating in PISA, even today, is that it does not make sense to invest time and economic resources in an evaluation project where Latin American countries always come last in the ranking—the so-called assessment "horse race". These objections, though, face strong counterarguments on at least two fronts. For one, some upward movement in results has been observed, as in the case of Peru, suggesting that a combination of better educational policies, more investment and overall improved economic performance pays back in terms of improved results. Second, PISA offers a regular measure every three years—in theory with much more reliable instruments—which is highly valued by countries when interested in making longitudinal analyses of progress in educational quality. Additionally, coming last in the rankings may have positive effects in spurring national debate and gaining public support to introduce difficult or controversial policy measures, such as reforms in teacher training or teacher evaluation programs.

As evidenced by interviewees' impressions of PISA, in some countries like Mexico, Brazil, and Chile, participation in the test is part of a bigger, geopolitical interest—either because they are member states of the OECD (and therefore obliged to be assessed by PISA), or because they find the OECD an attractive "environment" to be part of in strategic, economic terms. This might explain why some countries like Argentina still participate in PISA, even when it disqualifies and makes no use of the data, or why Mexico allows the OECD not only to assess learning through PISA but also to discuss specific recommendations and guidelines for national educational policy.

In general, though, as is also the case with low-stakes national assessments, there is a perception that international evaluations provide a wealth of information that is not sufficiently utilized for deeper analyses of contextual factors affecting results, or to rigorously compare the country's performance with that of educational systems around the world to derive input for educational reforms or improvements. In this sense, Finland has become the paradigmatic "success story", generating interest throughout Latin America rather than sheer disqualification for being a highly developed country (and therefore not suitable for valid comparisons with Latin America).

In spite of real or alleged sub-utilization of data, it is noticeable how PISA's assessment framework, constructs, and structure have powerfully influenced Latin America's national evaluation models and contents. They have also influenced curriculum design, both in content and structure. This is perhaps an ambitious conclusion based on insufficient evidence in the context of this report; however, the authors have been able to observe this tendency in a number of national cases through previous studies and consultancy work in different countries. Just to provide an example, it was after PISA that reading competencies, or reading comprehension to be exact, came to be almost universally defined in three distinct though complementary cognitive processes: locating information in a text, making inferences, and evaluating the texts' quality to adequately

Coming last in the rankings may have positive effects in spurring national debate and gaining public support to introduce difficult or controversial policy measures.

convey meaning. Before PISA, one could find as many definitions of "reading" as tests or curricula were being implemented. This has now changed.[25]

Finally, and again referring to PISA—currently and clearly the most influential international assessment in the region—it is worth mentioning that a new version of the test, specifically designed for developing countries, is soon to be implemented. It will be called PISA for Development (PFD). Three Latin American countries (Ecuador, Guatemala, and Paraguay) have confirmed their participation. PFD's aim is twofold. In the first place, it will include items that measure very basic skills, which in turn will help discriminate better and reveal differences in performance among millions of students who are likely to fall in the "Below 1" level of the regular version of PISA. Provided more basic skills are measured and compared, there will be more chances for analyses to understand finer aspects of learning within these populations and therefore to obtain more relevant input for improvement policies. The second aim will be to give countries with weak or inexperienced assessment units the possibility of developing technical capacity until they can, if desired, join the regular global assessment program. It is in accordance with this second aim that PFD is planning to pair up inexperienced country teams with stronger technical partners in the region to promote exchange, collaboration, and technical development. The fact that three of the five countries that have initially signed on to PFD are in Latin America suggests the high level of interest PISA generates in the region.

Lack of realism can lead to frustrated expectations, and insufficient consistency in capacity building is most surely conducive to poor performance in evaluation systems.

Lessons Learned

Our review found clear progress was made over the last decade. While acknowledging that we only covered three of the seven original dimensions (in Ferrer 2006) and the limited scope of our review, overall we believe that countries in Latin America have continued building their education assessment systems. While the rate of improvement may vary across countries, we did not find signs of backtracking. On the contrary, there are many signs of maturing systems.

However, not all news has been positive. Our review found that, in spite of the progress observed, many weaknesses remain. While this may have been clear to experts all along, the experience of countries in Latin America reconfirms that building an effective education assessment system that informs and supports improvements in the quality of education is a laborious and complex task that requires aligning system learning goals, standards, curriculum, and other critical aspects such as teacher training (pre and in service). This is, in fact, one of the three dimensions of the SABER Student Assessment framework (Clarke, 2012). It demands efforts to build capacity and adaptation in tools and processes to respond to the changing information needs of the education system. It may be accurate to say that among the five countries we considered, only Brazil and Colombia are reaching the point where we can say they have developed assessment systems that articulate production and use of data, with users both within and outside the schooling system. We believe this is an area in which further research is needed to better document the progress and the barriers and bottlenecks countries in Latin America are facing to achieve an advanced level of system alignment.

These are important lessons for countries in other regions that are less advanced in developing learning assessment systems (many of which experience situations similar to the ones countries in Latin America had two decades ago) and donors supporting them. They should be both realistic on what can be achieved over relatively short periods of time, and persistent in their capacity building efforts. Lack of realism can lead to frustrated expectations, and insufficient consistency in capacity building is most surely conducive to poor performance in evaluation systems.

The choice between putting autonomous institutes in charge of evaluation or keeping those functions in specialized units within the MOE is less obvious than many people would have thought ten years ago. Among the cases we reviewed, units (in Guatemala and Peru) have been able to do a sound job overall. While, naturally, we do not know what the situation would have been had they pursued the alternative path of establishing autonomous institutes, it is simply not the case that having the evaluation functions within the MOE is necessarily a recipe for failure.

Having said that, there is some suggestive evidence that institutes have some advantages vis-à-vis units arising from their higher level of autonomy. First, in places where bureaucratic bottlenecks (e.g. hiring rules) make it difficult to sustain the presence of highly qualified technical people within the Ministry, autonomy can help create better conditions to build capacity. ICFES and INEP may have benefited from this. Second, if sustained, autonomy can provide a higher level of protection from political pressures than those faced by units that report directly to political authorities in the MOE.

Two caveats to the above characterization are worth mentioning. First, legal autonomy does not guarantee political autonomy: political pressures can weaken the resolve of seemingly autonomous bodies. Second, autonomy does not solve, *per se,* the problem of poor credibility plaguing many evaluation systems. Unless the governance arrangements for autonomous evaluation institutes provide adequate oversight of their activities, doubts may remain about the trustworthiness of their results. The lesson for countries that are considering the establishment of autonomous evaluation institutes is that they demand a strong legal framework—including the establishment of governance arrangements that ensure the proper technical oversight, dedicated funding, and the ability to make the necessary human resource decisions independently.

Transforming the data generated through learning assessments into information that is usable and used by various system stakeholders remains a challenge. In part this is due to continued weaknesses in how results are reported and used. In some cases there may be problems in terms of the way results are communicated (for example, too much data or complex websites that are hard to navigate).

But reporting is only part of the problem. Effective demand for, and capacity to analyze, data is also a binding constraint in many cases. The truth is that countries in Latin America still struggle to get teachers and parents interested and invested in the data being generated by learning assessment systems. This is an area ripe for an experimental approach that tries alternative ways of communicating results and engaging actors in the education system in using those results.[26]

Perhaps the most interesting innovation in how evaluation results are communicated in easily usable form is the adoption of indices, as in the case of Brazil and Colombia. Indices not only simplify results presentations but they can also present a broader perspective on results, one that doesn't rely on learning outcomes as a single measure of quality.

Unless the governance arrangements for autonomous evaluation institutes provide adequate oversight of their activities, doubts may remain about the trustworthiness of their results.

On the other hand, civil society and the media in many countries have demonstrated an increased analytical capacity to use the data. Building their capacity is probably a more direct task than building the capacity of state organizations with more bureaucratic and political requirements. The development of such capacities in the media and civil society in Latin America has been, most likely, part of the (slow) process of exercising democratic rights and, thus, hard to accelerate and manage through external efforts. Nevertheless, thinking about countries in which donors have been active in strengthening the capacities of civil society organizations and the media, it is worth exploring efforts targeted at building capacities to demand and analyze education data.

Countries in Latin America are increasingly participating in international tests. This is, by now, a well-established trend. It appears that, from the point of view of decision-makers, the benefits of participation exceed the costs, including the political costs of being identified as relatively poor performers.

Experience suggests that those benefits are both technical and political. On the technical side, the participation in high quality international assessments such as PISA is seen as an investment activity: the resulting capacity building provides a strong incentive for a large and growing number of countries in Latin America to participate. PISA for Development may further contribute to this trend.

For those countries that are starting to build their education evaluation systems (in Latin America and elsewhere in the developing world), the lesson is that it is a worthwhile trip that demands a long-term commitment and sustained political support.

On the political side, there are some indications that education authorities in Latin America are starting to value the opportunity to benchmark their performance with that of other countries. While in the past regional benchmarking may have felt safer and more comfortable, as several countries in the region are already OECD members or aspire to become members in the not-too-distant future (e.g. Colombia or Peru), the political value of participating in regional assessments may diminish relative to international ones over time.

A review of the experience of more countries in Latin America, and a larger number of interviews, would have probably allowed us to dig deeper into the factors helping and constraining the development of learning assessment systems. Nevertheless, the broad trends we identified in our review provide some valuable lessons both for countries in Latin America and in other regions of the developing world. Education evaluation is being consolidated in the region and there are clear signs of progress. Further consolidation is still needed. Making good use of data to improve the quality of education remains the Achilles' heel of these efforts. The good news is that there are good practices in several countries from which the entire region can learn. And for those countries that are starting to build their education evaluation systems (in Latin America and elsewhere in the developing world), the lesson is that it is a worthwhile trip that demands a long-term commitment and sustained political support.

Notes

1. Starting in the late 1990s countries in Latin America experienced a wave of education reforms that moved beyond the traditional, exclusive, focus on expanding coverage and instead sought to improve quality in various ways. That wave created the initial demand for learning assessment systems. A complete list of evaluation programs in the region is included in **Appendix 3**.

2. The SABER country reports (http://saber.worldbank.org/index.cfm?indx=9&pd=5, visited on June 7, 2015) use a common methodology to assess student assessment systems in developing countries. This paper considers two of three dimensions of the SABER conceptual framework for Student Assessment Systems (Clarke, 2012).

3. The dimensions in the original study not covered here are: Curriculum and Standards, Instruments, Reports, and Subnational Systems. Given the nature of this paper – based on interviews and reviews of documents available online, all performed over a short period of time— it was particularly difficult to analyze these dimensions which would have required field visits (e.g. to collect data on sub-national systems) or the collection and review of documents not always available online (e.g. instruments).

4. We are not aware of any studies measuring the costs of setting up and running learning assessment systems in the region beyond an early PREAL study (Wolff, 2007). Nevertheless, as a reference, the budget of Mexico's INEE increased from $300 million pesos in 2013, to $613 million pesos en 2014 and $1 billion pesos in 2015 (or US$69 million). In comparison, the long-established INEP in Brazil had a budget of $854 million Reais in 2013 (the equivalent of US$393 million). In per capita terms, INEP's budget is almost 4 times as large as INEE's.

5. A list of websites is included in the references section.

6. See **Appendix 1** and **2** for a list of people interviewed and the interview protocol.

7. In this report, we will use the expression "assessment units" to refer to those offices or divisions within Ministries of Education (MOE) devoted to the specific task of measuring learning achievement. When we refer to Institutes, it will be in reference to autonomous or semi-autonomous organizations outside MOEs, which have that same function of measuring learning achievement. In neither case are we assuming that learning assessment functions have actually become a system, for that would imply that the evaluation programs have managed to become truly articulated with other components of educational administration, such as curriculum development, teacher training, and other related programs and policies.

8. In SABER, the enabling context covers such areas as the legislative or policy framework for assessment activities; leadership surrounding the assessment activity; public engagement with the assessment activity; the institutional arrangements for designing, carrying out, or using the results from the assessment activity; the availability of sufficient and stable sources of funding and the presence of competent assessment unit staff and classroom teachers (Clarke, 2012).

9. Even though the unit is known as DIGEDUCA, the legal name under which it was established is Dirección General de Evaluación, Investigación y Estándares Educativos.

10. Under the new legal framework INEE is responsible for coordinating the National System of Education Evaluation, regulating all the components of that system, evaluating the system and issuing policy guidelines based on evaluation results. This paper considers only the evaluation of student learning and not the other evaluation functions for which INEE is responsible.

11. Autarchic: Having and exercising complete political power and control (Source: The Free Dictionary, web). In Latin America, public institutions often define themselves as autarchic, autonomous, or semi-autonomous. The difference is not always clear, even if guided by sociological definitions, so in this paper the authors have respected the terminology officially used by the institutions under study.

12. The SEP can actually object to INEE's resolutions, provided that adequate explanations and alternatives are presented to national authorities. The education reform of 2012 in Mexico declared the National Institute of Educational Evaluation (INEE) an "autonomous public entity with legal personality and own patrimony" with the purpose of guaranteeing the provision of "high-quality educational services" in the country, and ensuring the production of well-founded, pertinent, and reliable data on the education system (INEE, 2015, p. 41). The INEE has authority in matters of educational evaluation, and is responsible for

Notes, cont.

"evaluating the quality, performance, and results of the national education system" (ibid). Therefore, its duty is to issue the guidelines that federal and local educational authorities must follow when performing their evaluation functions. Meanwhile, federal and local authorities are obligated to abide by these guidelines and utilize them to develop education policies based on results from these evaluations. In other words, INEE does not have a relationship of subordination with respect to the Secretariat of Public Education but rather one of "coordination and collaboration" (Bracho and Zorrilla 2014).

13. During the last decade, unions have become less active in terms of opposition to student learning assessments, probably due to the fact that these kinds of assessments have proven not to be designed as punitive instruments against teachers. However, they have become more active, although not necessarily successful, in preventing the implementation of teacher evaluation programs designed for providing economic incentives or to regulate access to new teaching positions in the educational system.

14. Even low-stakes tests tend to be more regular in terms of administration dates than a decade ago. However, high-stakes tests, because of their nature (i.e. consequences for individuals), need to be more rigorous in this regard.

15. Both assessments, EXCALE by INEE and Enlace by the SEP have been discontinued until a new evaluation program which involves both institutions is in place. The program will be INEE's responsibility, as part of its new legal mandate to coordinate a national evaluation system, and will be named PLANEA (Plan Nacional de Evaluación de Aprendizajes).

16. By format we refer to the way of presenting results in accordance with the methodology used for measuring learning achievement.

17. "Associated factors", the authors' translation of "factores asociados", is the expression used in Spanish to define or describe the analysis of context data used to statistically explain learning results. Analyses typically include school context data (e.g. teacher qualifications, number of books in library, access to internet), and also students and their family's characteristics (e.g. number of books at home, parents' education level, income, access to internet and cable TV). Each of these data items is called "factor". In more general terms, factors can be understood as background or context data used to explain learning results.

18. IDEB does not control for the school's socioeconomic level or size.

19. In recent years Guatemala has shown an increased interest in making more efficient use of evaluation results. In 2013, through panel discussions, interviews with key actors in the field of education and consultations with experts, the Ministry of Education (MINEDUC) discussed the need to use evaluations as criteria for such things as student grading and teacher pay increases, among others (MINEDUC, 2013, p. 33). MINEDUC acknowledged the deficiencies in communication and dissemination strategies, and developed the "EVALUAR" reports for teachers, summarizing evaluation results and offering recommendations to make better use of such results (MINEDUC 2015). It also provided school directors and teachers with Math and Language handbooks with strategies to improve the quality of teaching.

20. Many of these reports include learning assessment data disaggregated by gender, ethnic or linguistic background, among others. The number and variety of reports available on line, however, would require more time for analysis in order to provide a thorough account of all the variables used during data collection, and then for reporting learning results and their "associated factors". Peru offers a useful example. Since 2007, the Ministry of Education administers a census-based evaluation of student performance (ECE) in its schools. ECE assesses second grade students in math and reading comprehension. It also assesses fourth grade students whose first language is an indigenous language and who attend schools classified as Intercultural Bilingual Education (EIB), in reading comprehension in Spanish as a second language. Additionally, in 2007, 2008, 2010, and 2012, the evaluation also tested students' reading comprehension in one of four native languages—Awajún, Quechua Cusco-Collao, Aimara y Shipibo-conibo—depending on the region where it was administered. In 2014, the test was also offered for the first time in Quechua Chanka. In spite of showing an improvement with respect to 2012, in the 2014 exam students demonstrated poor reading comprehension in both languages (MINEDU 2014 and MINEDU 2012).

Notes, cont.

21. By contrast, one could think of tablet or smartphone applications where navigational options tend to be more limited but also clearer and easier to use.

22. An interesting case is observed in Uruguay, where the National Administration of Public Education administers evaluations online using computers provided by Plan Ceibal. Teachers can monitor students from their own computers as they take the exam and access results and view each student's scores as well as their individual answers. Test results come with information to help teachers interpret them. For instance, a detailed item "profile" to better understand the question, analyze persistent conceptual errors, and plan corrective interventions (Luaces, 2014). Teachers also have access to "didactic guides," developed by specialists to help teachers improve their teaching methodologies.

23. It is worth mentioning that, in 2003, PREAL published the findings of a regional study (Ferrer and Arregui, 2003) with the financial support of the Peruvian Consortium for Economic and Social Research (CIES). This study can also be taken as a reference in time regarding the changes and continuities in Latin America's participation in international assessments over a decade.

24. The interviewees did not offer any explanations for this lack of interest. It is possible that education authorities in the region feel that the cost of participating in so many international tests is simply too high.

25. Isomorphism in this field has not a few detractors and could possibly be one more reason to oppose international assessments as a new way of cultural colonialism. However, other voices argue that pedagogy and curriculum are plagued with erratic, arbitrary definitions and terminology, and therefore it should be good news that some conceptual and semantic agreement is been reached. The debate, nevertheless, is still open and will probably be for many years to come.

26. For a recent example from Mexico, see Hoyos, Garcia-Moreno, and Patrinos (2015).

References

Bracho, Teresa and Margarita Zorrilla (2014). El INEE en la reforma educativa: De su naturaleza jurídica y sus atribuciones.

CIEN and Inter-American Dialogue (2015). "El estado de las políticas públicas docentes: Guatemala." Ciudad de Guatemala.

Clarke, Marguerite (2012): What Matters Most for Student Assessment Systems: A Framework Paper. Washington, DC: The World Bank. (Available at: http://documents.worldbank.org/curated/en/2012/04/16238771/matters-most-student-assessment-systems-framework-paper)

de Hoyos, Rafael, Vicente A. García-Moreno and Harry A. Patrinos (2015): The impact of low-stakes accountability intervention with diagnostic feedback: Evidence from México, (mimeo). Washington, DC: The World Bank.

Ferrer, Guillermo and Patricia Arregui (2003): Las pruebas internacionales de aprendizaje en América latina y su impacto en la calidad de la educación: Criterios para guiar futuras aplicaciones, Document No. 26. Santiago de Chile: PREAL.

ESCALE (2015). "Magnitudes." [Database]. Lima: Ministerio de Educación del Perú. (Available at: http://escale.minedu.gob.pe/magnitudes)

Ferrer, Guillermo (2006): Educational Assessment Systems in Latin America: Current Practices and Future Challenges. Washington, DC: PREAL.

Fortín, Álvaro (2013). Evaluación Educativa Estandarizada en Guatemala: Un camino recorrido, un camino por recorrer. Ciudad de Guatemala: Ministerio de Educación. (Available at: https://www.mineduc.gob.gt/PORTAL/contenido/enlaces/documents/Documento_Evaluacion.pdf)

INEE (2015): Reforma Educativa: Marco Normativo. Ciudad de México: INEE.

Luaces, Oscar (2014). Evaluaciones en línea: Una innovación educativa en el marco del programa de una computadora por alumno. Montevideo: Administración Nacional de Educación Pública. (Available at: http://www.anep.edu.uy/sea/wp-content/uploads/2014/12/Evaluaciones-en-l%C3%ADnea-_-para-web.pdf)

Martinez Carballo, Nurit (2014, Feb 4). "SEP: En ENLACE hubo corrupción." El Universal. (Available at: http://archivo.eluniversal.com.mx/nacion-mexico/2014/impreso/sep-en-enlace-hubo-corrupcion-212806.html)

MINEDU (2012). Resultados de la Evaluación Censal de Estudiantes 2012 (ECE 2012). Lima: Ministerio de Educación del Perú. (Available at: http://www2.minedu.gob.pe/umc/ece2012/informes_ECE2012/Difusion/ECE%202012_Web_UMC.pdf)

MINEDU (2014). Resultados de la Evaluación Censal de Estudiantes 2014 (ECE 2014). Lima: Ministerio de Educación del Perú. (Available at: http://umc.minedu.gob.pe/wp-content/uploads/2015/02/ECE-2014-Web-270215-27febv2.pdf)

MINEDUC (2015). "Material Pedagógico." Ciudad de Guatemala: Ministerio de Educación de Guatemela-DIGEDUCA. Web. [Accessed 14 July 2015]. (Available at: http://www.mineduc.gob.gt/digeduca/)

OECD (2014). "Education at a Glance 2014: Brazil Country Note." Paris: OECD. (Available at: http://www.oecd.org/brazil/EAG2014-Country-Note-Brazil.pdf)

OECD (2012). Revisiones de la OCDE sobre la evaluación en la educación: México 2012. Paris: OECD Publishing

OECD (2011). Establecimiento de un marco para la evaluación e incentivos docentes: Consideraciones para México. Paris: OECD Publishing.

OEI (2012): Miradas sobre la Educación en Iberoamerica. Madrid: OEI/CEPAL/SEGIB. (Available at: http://www.oei.es/miradas2012.pdf)

SEP (2015). "Descarga serie histórica y pronósticos de la estadística del sistema educativo nacional." Ciudad de México: Secretaría de Educación Pública. (Available at: http://www.snie.sep.gob.mx/estadisticas_educativas.html)

UNESCO Institute for Statistics – UIS (2015). "Teachers by teaching level of education." Web, Database. [Accessed 3 August 2015. Web].

United Nations (2014): Open Working Group Proposal for Sustainable Development Goals, August 2014. (Available at: www.undocs.org/A/68/970)

References, cont.

Wolff, Laurence (2007): The Costs of Student Assessment in Latin America. Washington, DC: PREAL.

Zúniga Molina, Leonel, and José Luis Gaviria (2010), Challenges and Opportunities for the Further Development of the ENLACE Assessment for Evaluation and Teacher Incentives in Mexico, expert paper commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.

WEBSITES REVIEWED

BRAZIL
http://www.inep.gov.br/

COLOMBIA
http://www.icfes.gov.co/
http://www.mineducacion.gov.co/

GUATEMALA
http://www.mineduc.gob.gt/

MEXICO
http://www.inee.edu.mx/
http://www.sep.gob.mx/

PERU
http://www.minedu.gob.pe/umc

URUGUAY
http://www.anep.edu.uy/sea/

Appendix 1: List of People Interviewed

The following people were interviewed for this paper. Their generous contribution is hereby acknowledged. The paper presents the findings of those interviews as interpreted by the authors.

BRAZIL

- María Helena Guimaraes de Castro. Former President of Instituto Nacional de Estudos e Pesquisas (INEP).
- Ruben Klein. Consultant for Fundação Cesgranrio and President of the Brasilian Educational Evaluation Association (Associação Brasileira de Avaliação Educacional, ABAVE)

COLOMBIA

- Margarita Peña. Former President of Instituto Colombiano para el Fomento de la Educación Superior (ICFES).
- María Figueroa. Dean, School of Education at Universidad Externado de Colombia.

GUATEMALA

- Álvaro Fortín. Former Director of the Dirección General de Evaluación e Investigación Educativa (DIGEDUCA).
- Verónica Spross. Research Associate at Centro de Investigaciones Económicas Nacionales (CIEN)

MEXICO

- Margarita Zorrilla Fierro. Former Director General of Instituto Nacional para la Evaluación de la Educación (INEE).
- Marco Antonio Fernández. Director of Reasearch at México Evalúa.

PERU

- José Rodríguez. Former Director of Oficina de Medición de la Calidad de los Aprendizajes (UMC).
- Patricia Arregui. Principal Researcher at Grupo de Análisis para el Desarrollo (GRADE).

INTER-AMERICAN DEVELOPMENT BANK

- Emiliana Vegas. Chief of the Education Division.
- Alejandro Morduchowicz. Education Lead Specialist. Guatemala.

Appendix 2: Protocol for Interviews (Summary of Topics)

A. INSTITUTIONAL FRAMEWORKS

- Stability
- Financing and administrative autonomy
- Human resources
- Autonomy and capacity to disseminate results
- Transparency

B. REPORTING, DISSEMINATION AND USE OF RESULTS

- Coherence between types of reports and expected uses (curriculum development, pedagogy, targeting support, teacher training, selection of students, etc.)
- Adequacy of reports for different audiences (clarity, guides to interpretation, sensitization, etc.)
- Information on associated factors and value-added models
- Delivery: time frames; scope; regularity
- Impact: school use; policymaking/program design; political accountability

C. INTERNATIONAL ASSESSMENTS

- Reasons for participation
- Development of technical capacities
- Dissemination of results and impact on public opinion
- Specific uses of results

Appendix 3: 2009-2014 Evaluation Programs

BRAZIL

COUNTRY AND EVALUATING AGENCY	YEARS ADMINISTERED	GRADES/ YEARS [1]	DISCIPLINARY FIELDS [2]	TYPE OF STUDY [3]	INTERNATIONAL EVALUATIONS
Brazil's Ministry of Education (MEC) INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira)	2009, 2011 (Administered biannually)	5, 9 S, 3 USE SAEB-ANEB	L,M	S	PISA 2009 2012 LLECE-TERCE 2013
	2009, 2011 (Administered biannually)	5, 9S SAEB-Prova Brasil	L,M	C	
	2013 (Administered biannually)	9S, 3 USE SAEB-ANEB	L,M,Ss	S	
	2013 (Administered biannually)	9S SAEB-Prova Brasil	L,M,Ss	C	
	2011, 2012, 2013, 2014	2 S (In two stages: beginning and end of 2nd year) Provinha Brasil	L	V	
	2009, 2010	2S (In two stages: beginning and end of 2nd year) Provinha Brasil	L,M	V	
	2013, 2014 (Administered annually)	3 SAEB-ANA	L,M	C	
	2009-2014	The end of Upper Secondary Education ENEM*	Four tests and one essay, covering multiple fields of knowledge: L, M, LE, A, IT, P, Ch, B, G, H, Ph, S	V	
	2009-2014	Youth and adults Two grade levels: Ensino Fundamental (lower secondary) y Ensino Médio (upper secondary) ENCCEJA	L, M, Ss, H, G	V	

COLOMBIA

COUNTRY AND EVALUATING AGENCY	YEARS ADMINISTERED	GRADES/ YEARS [1]	DISCIPLINARY FIELDS [2]	TYPE OF STUDY [3]	INTERNATIONAL EVALUATIONS
Colombia's Ministry of Education ICFES (Colombian Institute for the Evaluation of Education)	2009	5,9 (SABER)	M,L,NS	C and S [7]	PISA 2009 IEA-ICCS 2009
	2009-2014	11[3] (SABER 11: An evaluation tof the state of Secondary Education)	M, L, Ss, NS, CitE, LE, QR	C	
	2009-2014	Validation of the High School degree [4]	L,M,B,P,Ch ,Ss Ph,LE	V	
	2009-2014	SABER PRO: [5] An evaluation of the state of post-secondary education	General competencies that are necessary for good professional performance Specific competencies common to different types of programs	C (Taking the SABER PRO is a mandatory requisite at this grade level.)	
	2010-2014	Pre SABER 11[6]	M, L, Ss, NS, CitE, LE, QR	V	
	2011	5,9 (SABER)	L,M,NS,Ss	S	PIRLS 2011
	2012	3,5,9 (SABER)	L,M,NS	C	PISA 2012
		5,9 (SABER)	CitE	C	
	2013	3,5,9 (SABER)	L,M	n/a	LLECE-TERCE 2013
		5,9 (SABER)	NS	n/a	
	2014	3 (SABER)	L,M	n/a	
		5 (SABER)	L,M,NS	n/a	
		9 (SABER)	L,M,NS,FE	n/a	

GUATEMALA

COUNTRY AND EVALUATING AGENCY	YEARS ADMINISTERED	GRADES/ YEARS [1]	DISCIPLINARY FIELDS [2]	TYPE OF STUDY [3]	INTERNATIONAL EVALUATIONS
Guatemala's Ministry of Education DIGEDUCA (General Directorate of Educational Evaluation and Investigation)	2009	1,3,6 (National Evaluation of Primary School)	L,M	S	IEA-ICCS 2009
		1 (Sample-Based Evaluation of the Basic Cycle)	L, M	S	
		3M (Evaluation of Third Year of Basic Cycle)	L,M	C	
		3 (Evaluation of Candidates of Scholarships for Excellence)	L,M	V	
		Last year of upper secondary level ("GRADUANDOS" Exam)	L,M	C	
	2010	1,3,6 National Evaluation of Primary School	L,M	S	
		1, 2 (Sample-Based Evaluation of the Basic Cycle)	L, M	S	
		Last year of upper secondary level ("GRADUANDOS" Exam)	L,M	C	
	2011*	1 (PAMI- Evaluation of Early Mathematics Sills)	M	S	
		1 (LEE- Evaluation of Early Reading and Writing Ability)	L	S	
		2 (ELGI- Reading Evaluation for Early Grade Levels)	L	S	
		1,2,3 (Sample-Based Evaluation of the Basic Cycle)	L, M	S	
		Last year of upper secondary level ("GRADUANDOS" Exam)	L, M	C	
	2012	1,3,6 (National Evaluation of Primary School)	L, M	S	
		1 (PAMI- Evaluation of Early Mathematics Sills)	M	S	
		1 (LEE- (LEE- Evaluation of Early Reading and Writing Ability)	L	S	
		1,2,3 (Sample-Based Evaluation of the Basic Cycle)	L, M	S	
		3M (Evaluation of Reading at Third Year of Basic Cycle)	W	S	
		Last year of upper secondary level ("GRADUANDOS" Exam)	L, M	C	
		Last year of upper secondary level (Writing Exam for GRADUANDOS)	W	S	
	2013	1,3,6 (National Evaluation of Primary School)	L, M	S	LLECE-TERCE 2013
		3M(Evaluation of Third Year of Basic Cycle)	L, M	C	
		Last year of upper secondary level ("GRADUANDOS" Exam)	L, M	C	
	2014	3,6 (National Evaluation of Primary School)	L , M	S	
		3M (Evaluation of Third Year of Basic Cycle)	L,M	C	
		Last year of upper secondary level ("GRADUANDOS" Exam)	L, M	C	

MEXICO

COUNTRY AND EVALUATING AGENCY	YEARS ADMINISTERED	GRADES/ YEARS [1]	DISCIPLINARY FIELDS [2]	TYPE OF STUDY [3]	INTERNATIONAL EVALUATIONS
Mexico's Secretariat of Public Education (SEP) ENLACE Exam (National Evaluation of Academic Achievement at Educational Centers)	2009	3, 4, 5, 6 2S, 3S (Enlace Basic Education)	L, M, CEE	C	
		1S (Enlace Basic Education)	L, M	C	
		Last Year of Upper Secondary Education (UPS) (Enlace Upper Secondary Education)	L, M	C	
	2010	3,4,5,6 2S, 3S (Enlace Basic Education)	L, M, H	C	
		1S (Enlace Basic Education)	L,M	C	
		Last grade of Upper Secondary Education (USE) (Enlace Upper Secondary Education)	L,M	C	
	2011	3,4,5,6 1S (Enlace Basic Education)	L,M ,G	C	
		2S, 3S (Enlace Basic Education)	L,M	C	
		Last grade of Upper Secondary Education (USE) (Enlace Upper Secondary Education)	L,M	C	
	2012	3,4,5,6 1S, 2S, 3S (Enlace Basic Education)	L,M, Ss	C	
		Last grade of Upper Secondary Education (USE) (Enlace Upper Secondary Education)	L, M	C	
	2013	3,4,5,6 2S, 3S (Enlace Basic Education)	L, M, CEE	C	
		1S (Enlace Basic Education)	L,M	C	
		Last grade of Upper Secondary Education (USE) (Enlace Upper Secondary Education)	L,M	C	
	2014	Last grade of Upper Secondary Education (USE) (Enlace Upper Secondary Education)	L,M	C	
	2009 (2008-2009 Cycle)	6 (EXCALE 06)	L, M, CivE, NS	S	IEA-ICCS 2009 PISA 2009
	2010 (2009-2010 cycle)	3 (EXCALE 03)	L, M, CivE, NS	S	
		Last grade of Upper Secondary Education (USE) (EXCALE 12)	E, CitE	S	
	2011 (2010-2011 cycle)	3 (Prescolar) (EXCALE 00)	L,M	S	
	2012 (2011-2012 cycle)	3S (EXCALE 09)	L,M, CEE, Ss	S	PISA 2012
	2013 (2012-2013 cycle)	6 (EXCALE 06)	L, M, CEE, NS	S	LLECE-TERCE 2013

PERU

COUNTRY AND EVALUATING AGENCY	YEARS ADMINISTERED	GRADES/ YEARS [1]	DISCIPLINARY FIELDS [2]	TYPE OF STUDY [3]	INTERNATIONAL EVALUATIONS
Peru UMC (Office for the Measurement of the Quality of Learning)	2009	2 (ECE 2009)	L,M	C	PISA 2009
		4* (ECE 2009)	RC1	C	
	2010	2 (ECE 2010)	L,M	C	
		4* (ECE 2010)	RC1, RC2	C	
	2011	2 (ECE 2011)	L,M	C	
		4* (ECE 2011)	RC1	C	
	2012	2 (ECE 2012)	L,M	C	PISA 2012
		4* (ECE2012)	RC1,RC2	C	
	2013	2 (ECE 2013)	L,M	C	
		4* (ECE 2013)	L2	C	
	2014	2 (ECE 2014)	L,M	C	
		4* (ECE 2014)	RC1, RC2	C	

LEGEND:

[1]

M: Middle School
S: Secondary Education
USE: Upper Secondary Education

[2]

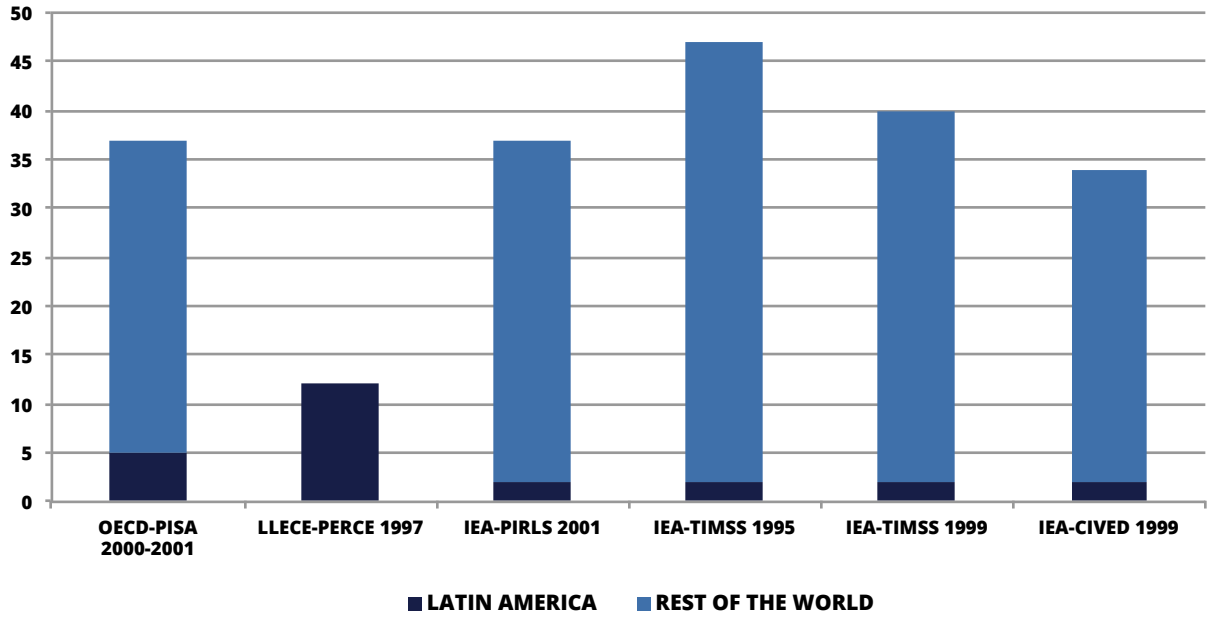
A: Art
B: Biology
CC: Citizen Competencies
CEE: Civics and Ethics Education
Ch: Chemistry
CitE: Citizenship Education
CivE: Civic Education
H: History
FE: Financial Education
FL: Foreign Language
G: Geography
IT: Information Technology
L: Language (Spanish)
M: Mathematics
NS: Natural Sciences
P: Physics
Ph: Philosophy
QR: Quantitative Reasoning
RC1: Language Comprehension in Spanish as a Second Language
RC2: Language Comprehension in Four Original Languages (Aimara, Awajún, Quechua Cusco-Collao, Shipibo and Conibo)
S: Sociology
Ss: Sciences
W: Writing

[3]

C: Census-Based
S: Sample
V: Voluntary

Appendix 4: Participation in International Tests

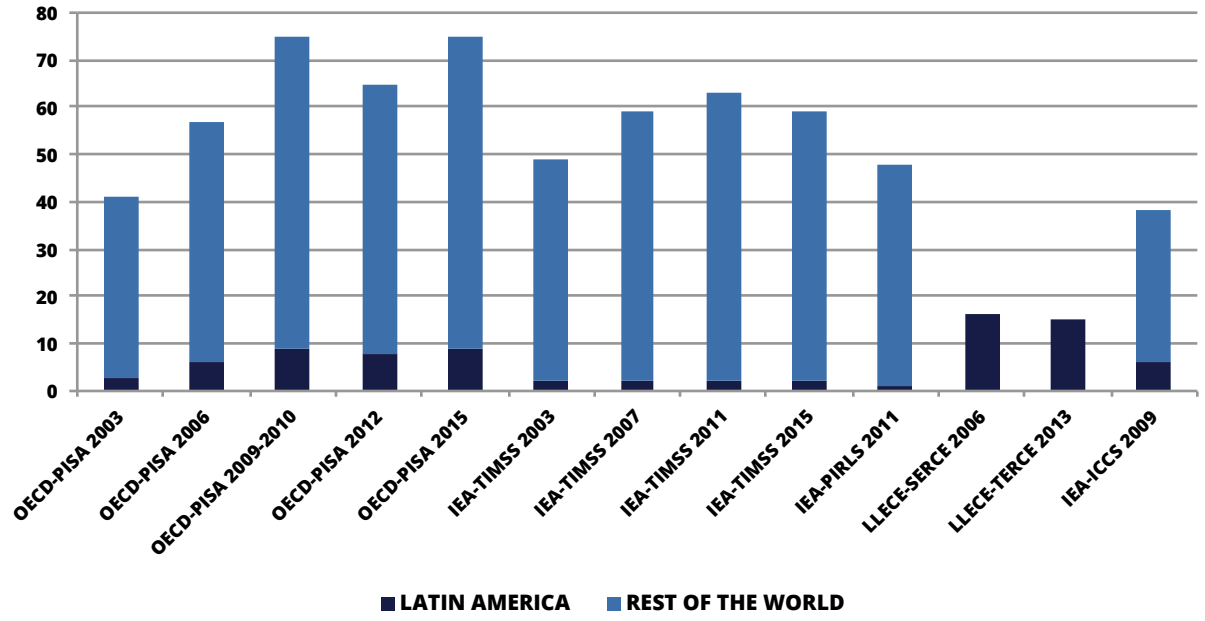
NUMBER OF COUNTRIES PARTICIPATING IN ARREGUI AND FERRER (2003)



ASSESSMENTS	LATIN AMERICA	REST OF THE WORLD
OECD-PISA 2000-2001	5	32*
LLECE-PERCE 1997	12	0
IEA-PIRLS 2001	2	35
IEA-TIMSS 1995	2	45
IEA-TIMSS 1999	2	38
IEA-CIVED 1999	2	32
Total of assessments in the Region: 6		

*Another 11 countries completed the same assessment in 2002.

NUMBER OF COUNTRIES PARTICIPATING IN LATEST APPLICATIONS (2003-2015)



ASSESSMENTS	LATIN AMERICA	REST OF THE WORLD
OECD-PISA 2003	3	38
OECD-PISA 2006	6	51
OECD-PISA 2009-2010	9	66
OECD-PISA 2012	8	57
OECD-PISA 2015	9	66
IEA-TIMSS 2003	2	47
IEA-TIMSS 2007	2	57
IEA-TIMSS 2011	2	61
IEA-TIMSS 2015	2	57
IEA-PIRLS 2011	1	47
LLECE-SERCE 2006	16	0
LLECE-TERCE 2013	15	0
IEA-ICCS 2009	6	32
Total of assessments in the region: 13		

TIMSS and PIRLS: Benchmarking participants are not included.

Inter-American Dialogue

1211 Connecticut Ave. NW, Suite 510, Washington, DC 20036
Tel: 202-822-9002 / Fax: 202-822-9553
Email: education@thedialogue.org
Web: thedialogue.org/agenda/programs/education