



Establishing a Framework for Evaluation and Teacher Incentives

CONSIDERATIONS FOR MEXICO



Establishing a Framework for Evaluation and Teacher Incentives

CONSIDERATIONS FOR MEXICO



This work is published on the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Organisation or of the governments of its member countries.

Please cite this publication as:

OECD (2011), *Establishing a Framework for Evaluation and Teacher Incentives: Considerations for Mexico*, OECD Publishing.
<http://dx.doi.org/10.1787/9789264094406-en>

ISBN 978-92-64-09439-0 (print)

ISBN 978-92-64-09440-6 (PDF)

This publication is a product of the co-operation agreement established between the government of Mexico and the OECD, which aims to improve the quality of education in Mexico.

Photo credit: ©UNESCO/José Gabriel Ruiz Lembo.

Corrigenda to OECD publications may be found on line at: www.oecd.org/publishing/corrigenda.

© OECD 2011

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at info@copyright.com or the Centre français d'exploitation du droit de copie (CFC) at contact@cfcopies.com.

Foreword

Education is the basis for a successful future of our societies. Equally, teachers are the building blocks of the success of a country's education system. Indeed, a well developed system combines many different elements, including national curricula and standards, the management and performance of schools, the quality, motivation and perspectives of teachers, and an effective education evaluation system. But teachers are key and therefore many governments are putting more emphasis on their role.

This report, *Establishing a Framework for Evaluation and Teacher Incentives: Considerations for Mexico*, examines both the design and gradual implementation of effective policies on teacher evaluation and incentives. Together with its sister publications, *Improving Schools: Strategies for Action in Mexico* and *Evaluating and Rewarding the Quality of Teachers: International Practices*, it presents a comprehensive policy strategy for Mexico's educational reform project.

Mexico has seen impressive progress by raising student enrolment, and more recently by building a solid institutional framework for the evaluation of learning outcomes. But more needs to be done.

First, Mexico has to further improve its teachers' workforce. In order to move from *adequate* to *good* and then from *good* to *great* in its education performance, Mexico will need to put teachers on par with other professions in terms of status. A comprehensive reform package to attract the best graduates to become teachers and to develop them into effective instructors will require improving pedagogical practices through better training and recruiting practices, reforming the reward and pay system, and putting in place the proper and differentiated incentives.

Second, Mexico has to focus on "three E's" and install a more Effective Education Evaluation system. The success of educational efforts needs to be measured by the learning outcomes of students. The use of further assessment and evaluation tools, while strengthening the existing system, will be crucial. These tools need to be increasingly performance-based, link information between teaching and learning outcomes better, and be part of a comprehensive and well-aligned instructional learning system.

We encourage the Mexican government to undertake the necessary reforms by: *a)* further developing its assessment system focused on student learning outcomes; *b)* strengthening its teacher policy, including taking the necessary steps towards teacher evaluation; and *c)* ensuring all actors are committed and motivated to improve performance. The OECD stands ready to help with the implementation of Mexico's comprehensive reform agenda.

This report was prepared under the guidance of the OECD Steering Group on Evaluation and Teacher Incentive Policies in Mexico, a group of internationally renowned experts and OECD analysts, chaired by Carlos Mancera. The following members of the Steering Group and invited experts contributed to the report: José Luis Gaviria, Jorge Juárez Barba, Enrique Roca Cobo, Halsey Rogers, Lucrecia Santibáñez, Susan Sclafani, Margarita Zorrilla, Leonel Zúñiga, Mathew Springer and Eva Baker. The report was produced under the auspices of the Indicators and Analysis Division of the OECD Directorate for Education under the responsibility of Alejandro Gomez Palma, Marlene Gras, Andreas Schleicher, Michael Davidson, William Thorn, Elisabeth Villoutreix, Isabelle Moulherat, Niccolina Clements and Marika Boiron.


Angel Gurría
 OECD Secretary-General

Special Acknowledgements

Since the start of the Co-operation Agreement in 2008 to improve the quality of education, the OECD and the government of Mexico have co-operated closely to support current and future education reform efforts. The work has focused on effective policy design and implementation, evaluation and assessment, teacher incentives, teacher policy and school performance. This report presents the findings and main considerations for Mexico to support reform in these areas. As part of this work, the OECD has mobilised its existing stock of knowledge and expertise, reviewed relevant international practice on these issues, and participated in consultations with national and international experts, as well as with Mexican stakeholders through technical meetings, international workshops and review visits.

The OECD Steering Group on Evaluation and Teacher Incentive Policies in Mexico is deeply grateful to Minister Lujambio and the many departments of the Mexican Ministry of Education for their readiness to share information, their willingness to share their knowledge, and their interest in the outcomes of our joint efforts. In particular, we would like to thank Francisco Ciscomani, Head of the Unit of Planning and Evaluation of Education Policies (UPEPE) and his team, particularly Bernardo Rojas, Lourdes Saavedra, Silvia Ojeda and Janina Cuevas, whose professionalism and human qualities made this challenging effort an even more enjoyable task. We would also like to take this opportunity to thank those high-ranking officials who supported our work: Fernando González, Deputy Director of Basic Education and Rodolfo Tuirán, Deputy Minister of Tertiary Education. Many General Directors of the Ministry also contributed to the success of our joint efforts: Ana María Aceves, Rafael Freyre Martínez, Marcela Santillán, Leticia Martínez and Juan Martín Martínez. For their time, support and thoughtful contributions we are truly thankful. Silvia Ortega, Dean of the National Pedagogical University, provided thoughtful insights throughout the discussions. We would also like to extend our gratitude to Josefina Vázquez-Mota, former Minister of Education, Jorge Santibañez, former Head of UPEPE, and their respective teams for their interest and support, particularly in the initial stages of the Co-operation Agreement.

We are indebted to all of the education authorities from the Mexican states that supported our efforts and made valuable contributions, especially those of Aguascalientes, Chiapas, Nuevo León and Veracruz. The work and the resulting recommendations have benefited from their generous support and insights. The interest and engagement of different national and international experts, civil society organisations, legislators and representatives of the National Union of Education Workers (SNTE) are also greatly appreciated. The success of reform efforts in Mexico, as elsewhere, will rest largely on the plurality and common vision that can be shared among stakeholders.

We would also like to thank all of the students, school principals, teachers, supervisors and technical staff (*Asesores Técnico Pedagógicos*), who shared their experience, challenges, concerns and hopes with us. Their input provides a salutary reminder that the ultimate success of reform efforts will be measured in their classrooms and schools.

We would like to extend our gratitude to Ambassador Agustín García López, Permanent Delegate of Mexico to the OECD, and also to Luisa Solchaga, Education Counsellor, and their team, for their support and contributions. The Steering Group on Evaluation and Teacher Incentive Policies would like to express special gratitude to the Chair, Carlos Mancera, for his committed engagement and valuable guidance throughout the Project.

Information on the events and relevant reports produced under the OECD-Mexico Agreement to Improve the Quality of Education in Schools in Mexico can be found on the website: www.oecd.org/edu/calidadeducativa.

Table of Contents

EXECUTIVE SUMMARY	9
CHAPTER 1 MEXICO RESPONDS TO EDUCATION CHALLENGES	19
CHAPTER 2 THE PUBLIC POLICY FRAMEWORK FOR IMPLEMENTING EDUCATION REFORMS	27
2.1 Education policy reforms in an international context.....	28
2.2 Mobilising OECD research, international practices and national knowledge.....	30
2.3 Policy dimensions of basic education reform: Asking the right questions.....	32
2.4 Considerations for Mexico.....	34
CHAPTER 3 ACCOUNTABILITY AS A POLICY DRIVER FOR IMPROVING STUDENT LEARNING OUTCOMES	39
3.1 Types and features of educational accountability systems.....	40
3.2 Teacher performance: Towards a fuller understanding of accountability.....	42
3.3 Considerations for Mexico.....	44
CHAPTER 4 USING STUDENT LEARNING OUTCOMES TO MEASURE IMPROVEMENT	49
4.1 Student learning outcomes: Assessment instruments and measures.....	50
4.2 The ENLACE assessment system in Mexico.....	55
4.3 Challenges and opportunities for further development of the ENLACE assessment system.....	60
4.4 Summary recommendations for Mexico.....	62
CHAPTER 5 ASSESSING THE VALUE-ADDED OF SCHOOLS: ENHANCING FAIRNESS AND EQUITY	67
5.1 Value-added models with the school as the unit of accountability.....	68
5.2 The importance of quality data and information.....	73
5.3 Consequences linked with fair and credible assessment of schools and teachers.....	73
5.4 Considerations for Mexico.....	75
CHAPTER 6 IN-SERVICE TEACHER EVALUATION: POLICY AND IMPLEMENTATION ISSUES	79
6.1 International practices.....	80
6.2 Four key questions evaluation systems must address.....	82
6.3 Considerations for Mexico.....	83
CHAPTER 7 INCENTIVES FOR IN-SERVICE TEACHERS	91
7.1 Types of teacher incentives.....	93
7.2 National guidelines and local implementation: Finding the right balance.....	103
7.3 Piloting, monitoring and evaluating incentives.....	103
7.4 Considerations for Mexico.....	105
Appendix 7A Piloting, Monitoring, and Evaluating Teacher Incentive Programmes: Recommended Practices	107
CONCLUSION	121

Boxes

Box 4.1	<i>Prova Brasil</i> for accountability and improvement.....	51
Box 4.2	Mixed systems of student assessments.....	54
Box 5.1	Jurisdictions in Mexico.....	69
Box 5.2	Example of a linear regression value-added model.....	71
Box 7A.1	Important definitions.....	107
Box 7A.2	Importance of stakeholder engagement.....	114
Box 7A.3	The Data Quality Campaign.....	115

Figures

Figure 2.1	Schematic representation of effective Knowledge Mobilisation, Analysis and Application (Knowledge MAP): Country-specific heuristics.....	32
Figure 2.2	Schematic representation of a country-specific heuristics model: The Public Policy Framework for Education Reform.....	33
Figure 3.1	Schematic representation of the capacity-threshold concept of teacher accountability.....	43
Figure 5.1	Schematic representation of a simple value-added model.....	70
Figure 7.1	Model of programme piloting, evaluation and monitoring.....	108

Tables

Table 1.1	Areas covered by the Alliance for the Quality of Education and the OECD-Mexico Agreement.....	20
Table 1.2	OECD methodology to support policy implementation and deliverables.....	20
Table 1.3	Dimensions of the basic education system, school year 2008/09 (formal).....	22
Table 1.4	Basic education by type of school, school year 2008/09.....	22
Table 3.1	Common types of assessments used in accountability systems.....	41
Table 4.1	Instruments and sources of evidence to assess student learning.....	52
Table 4.2	Reliability of ENLACE.....	58
Table 4.3	Correlation between subscales of problem solving, reading and science, PISA 2003.....	59
Table 4.4	Percentages of probable test cheating cases detected for ENLACE 2006 to 2009.....	60
Table 5.1	Benefits and policy implications of value-added methods for accountability and school improvement.....	74
Table 6.1	General overview of teacher evaluation practices in Mexico.....	84
Table 6.2	Implementation steps in sequential order.....	84
Table 6.3	Summary of specific recommendations for Mexico regarding teacher evaluation.....	85
Table 7.1	Summary of benefits of conducting pilots.....	104
Table 7A.1	Focus and purpose of evaluation questions in the context of incentive pay.....	109
Table 7A.2	Evaluation designs to investigate the impact of programme and policy interventions.....	110
Table 7A.3	Strategies to monitor quality of incentive pay design and implementation.....	114

Executive Summary

Mexico, as the world's 14th largest economy (2009), faces important challenges in education. Despite the significant progress of the past decades in terms of access to education, improvements in completion rates for lower education levels and development of learning assessments, considerable improvement is still needed. Mexico already invests a high percentage of the public budget in education (at nearly 22%, it is the highest among OECD countries). Results from the 2009 round of the Programme for International Student Assessment (PISA) have shown that although improvement is possible in a relatively short period of time, important challenges remain. In addition to improving the quality of educational services, increasing attainment levels and reducing drop-out rates are also priorities. It is equally important, however, for Mexico to ensure that all students, including those from disadvantaged socio-economic backgrounds and indigenous families, have equal educational opportunities.

To address these issues, in 2008, the Mexican government and the OECD established the Co-operation Agreement "Improving Education in Mexican Schools". The purpose of the Agreement was to determine not only *what* policy changes to consider in Mexico, but also *how* to design and implement policy reforms effectively, given local conditions, constraints and opportunities. One of the strands of this Agreement has focused on developing appropriate policies to evaluate the quality of schools and teachers, particularly assessments, and to link learning outcomes to incentives for continuous improvement. This part of the work has been led by a group of international experts forming the OECD Steering Group on Evaluation and Teacher Incentive Policies in Mexico.

This summary report presents the main findings and policy recommendations developed by the Steering Group and the OECD Secretariat over the course of the Co-operation Agreement.¹ It draws on the results of international workshops and technical meetings with stakeholders in the Mexican education system, field visits, thematic reports from invited experts, and the stock of OECD research and knowledge. Since no single model of education reform can serve to guide all of the reform efforts in Mexico, the recommendations draw on experiences from over 20 countries.

OPPORTUNITY FOR EDUCATION REFORM IN MEXICO

The Mexican government established policy priorities for education reforms in its Education Sector Programme 2007-12. To monitor progress towards achieving its objectives, the Mexican Ministry of Education (SEP) established improvement indicators for student achievement as measured by the national ENLACE assessment and PISA. Other key indicators relate to teachers' professional development, school empowerment, equity in educational opportunities, and reforms relating to content and curriculum. To facilitate policy reforms, in 2008, the Mexican government established the Alliance for the Quality of Education with the national teachers' union (SNTE), which helped define the thematic focus of the Co-operation Agreement with the OECD. In this context, the following recommendations and considerations aim to provide SEP and relevant stakeholders with guidance on the policy priorities for a lasting and effective reform process.

1	The public policy framework for implementing education reforms: For countries and education systems to adapt and implement policy reforms tailored from best practices and international examples, local conditions, constraints and opportunities must be adequately addressed. When combined with international practices and comparable evidence, local knowledge mobilisation can provide a vital link in adapting best practices for effective education reforms suited to national priorities and contexts. The purpose of the <i>public policy framework for education reforms</i> presented in this report is to provide relevant stakeholders in Mexico with guidelines for continued local knowledge mobilisation to inform current and future reform processes.
1.1	In combination with international practices and available research evidence, <i>country-specific knowledge mobilisation</i> on particular policy issues is a vital element to effectively design, plan and implement educational reforms that are viable and sustainable given the conditions, constraints and opportunities in Mexico. Experience clearly suggests that reliable and up-to-date knowledge about particular policy topics is crucial in the process of adopting best practices and policy recommendations.
1.2	Current and future education reform efforts in Mexico would benefit from a methodical consideration of each of the following <i>six dimensions of the public policy framework</i> for lasting and successful education reforms: i) Data, information and indicators: This implies a consideration of the quality and quantity of relevant data and information available (on students and teachers, schools, performance, and linkages between them), for target-setting and to identify deficient areas to be addressed. ii) Social relevance and stakeholder engagement: This includes considering strategy options for communication, engagement and consultation with primary stakeholders, including the general public, teachers, principals and local educational authorities. It is important to identify how the proposed reform can be translated into a socially relevant and meaningful message for the average family, teachers and principals. iii) Public funding: It is important to consider amounts and consistency of public funding for development and implementation of the policy reform (e.g. whether it is annual or fixed in the budget), including potential cost-benefit analysis, cost projections and economies that can be obtained by re-channelling existing budget items or programmes. iv) Institutional arrangement: This includes a consideration of public institutions (central and state educational authorities), to identify specific bodies that should contribute to developing standards, evaluations, and proposing modifications. v) Legal and regulatory framework: It is important to foresee potential conflicts and possible modifications that may be necessary in education laws and related areas (e.g. labour laws) to carry out education reforms. vi) Decentralisation and devolution process: This includes looking at <i>formal</i> and <i>de facto</i> levels of devolution across the main federal bodies and state educational authorities responsible for providing educational services (including resources, capacity, information management, evaluation and supervision).
2	Public accountability: All stakeholders should feel responsible and be held publicly accountable for student learning and overall educational results.
2.1	<i>Performance, equity and value for investment in education are challenges for Mexico</i>, as in many other countries undertaking important educational reforms. This is illustrated by Mexico's performance in international comparisons, in the great diversity that exists between and within Mexican regions and states, and in the importance that education spending continues to have in terms of share of the public budget, despite modest per-pupil spending compared to other OECD countries. Holding all actors involved in Mexico's education system accountable for increasing the performance of all students, in all schools, provides a clear message and a way to align efforts and resources.

2.2	<i>Actors should be held accountable for student learning and growth, and provided with the necessary assistance and capacity building.</i> A clearly defined accountability system focused on the results of student learning and growth can provide the necessary coherence, given the size, complexity and multiple interests of the participants in Mexico's education sector. The use of student learning as a key criterion against which state education authorities, schools, principals and teachers will be held accountable, reflects a focus on outcomes rather than input-focused policy reforms. International practice regarding performance-based teacher incentives, for example, reflects this change. This does not imply, however, that issues of infrastructure or social inclusion are no longer important for the Mexican educational system. Rather it implies that learning and development for all students – fostered, cultivated, assessed and evaluated through various means – should be the ultimate goal of policy action and reforms. Support to students, schools, principals and teachers, as well as professional development, are vital complements to increased accountability.²
2.3	<i>Accountability focused on student learning and growth implies establishing clear standards.</i> The development of standards as a key component of the accountability system focused on student learning should address at least three priorities: <i>i)</i> appropriate development of standards for content, student performance and teacher performance; <i>ii)</i> alignment and coherence between standards, assessment, evaluation and professional development; and <i>iii)</i> alignment of standards to international best practice and internationally competitive benchmarks of student knowledge and skills. Within a standards-based accountability framework, actors should have incentives to meet or exceed the expectations that are reflected in standards.
2.4	<i>Accountability measures should include complementary criteria of effort as well as performance.</i> A standards-based accountability system for students, schools and teachers in Mexico should consider using measures of student learning and growth (from standardised assessments and other reliable methods, where possible), as well as complementary criteria regarding individual, group and school performance. This is important in Mexico as student and teacher attendance, punctuality and time-on-task remain important issues. An accountability system in Mexico should take into account the fact that some principals and teachers may not be performing to their current capacities. Incentives are needed, therefore, to increase basic effort and performance, as well as supporting capacity building and professional development. Reduction of student drop-out rates, for example, can also be considered as an indicator. Accountability also implies that some teachers who receive adequate technical assistance and opportunities for professional development, but who do not improve performance, would be counselled out of the profession.
2.5	<i>The focus should be on students, schools and teachers for continuous improvement with the school as a basic unit of accountability.</i>³ Although different levels and actors in the education system should be held accountable, the school can serve as the basic unit of accountability, with individual data, information and monitoring for students and teachers. Student and teacher data and information at the school level can be used to support improvement efforts, teacher incentives and stimuli, education interventions for low performers, and the identification of good practices for modelling and to inform the development of teaching standards, for example.
2.6	<i>It is important to define a gradual process to develop complementary approaches of assessment using multiple sources of evidence.</i> Developing a robust standards-based accountability system is a gradual process, with clear stages, and with complementary approaches to assessment and evaluation. Both summative and formative assessments of student learning and growth, as well as school and teacher performance, should form part of the accountability system in Mexico. The development of such a system, however, should be delineated in stages with a thorough consideration of current and projected capacities, methods and costs.

3	Importance of student learning outcomes: Student learning and growth over time should be a key criterion to gauge the performance of schools, teachers, parent participation bodies, state and federal educational institutions, and the system as a whole. Results from standardised assessments are important, but other reliable and valid measures of student learning should be employed for a fuller picture of student achievement.
3.1	<i>Student learning and growth as the basis of accountability and standards require multiple, cross-referenced, valid and reliable measures.</i> All of the current measures and instruments of student learning and growth (teacher assessments, portfolios of student work, classroom observation, and standardised tests, among others) present potential sources of error and bias. A complementary approach that uses valid evidence from multiple sources should be gradually developed to take into account current instruments in Mexico, estimate costs, and determine the capacity building and instrument development that are required. With clear content and performance standards of what students are expected to know and know how to do, measures and procedures to assess the learning and improvement expected from students can be further developed.
3.2	<i>The use of student performance data should be accompanied, when possible, with complementary and reliable measures of student learning, as these are developed, tested and validated.</i> The relative importance of student data and school-based or teacher assessments can be redefined over time. The state of Victoria in Australia, Hong Kong-China, and Canada are examples of better performing systems that combine standardised assessments with school-based assessments (e.g. locally graded but externally moderated), student projects and extended papers.
3.3	<i>Student performance data, such as those from the annual ENLACE assessment in Mexico, can play an important role in accountability and school improvement efforts.</i> Current efforts by SEP and state educational authorities regarding the presentation and use of ENLACE demonstrate the high degree of social acceptance and potential of ENLACE. Student performance data aggregated at the group, school, zone or state levels can be employed in static, improvement or growth models.
3.4	<i>The ENLACE assessment in Mexico has shown to be a valid and reliable measure of student achievement.</i> This provides Mexico with a valuable opportunity to exploit the potential of the student performance data provided annually by the assessment.
3.5	<i>A specific development programme should be established for the ENLACE assessment, considering issues of cognitive demand, curricular alignment and coherence.</i> The best-available evidence on student learning progression and standards should be considered. The development of ENLACE should also set clear stages and goals that address technical (e.g. vertical equating), administrative (e.g. unique student, teacher and school identifiers and linkages) and logistical (e.g. improved test supervision) considerations. ⁴ With expanded use of the ENLACE assessment in the future, enhanced supervision and security of test administration, for example, should be addressed. The programme should also have a long-term vision that takes internationally benchmarked content and performance standards into account. As content and performance standards are established in Mexico, student performance data can be used, in conjunction with analytical models (e.g. growth) for specific policy objectives and programmes. Throughout the process, consideration should be given to the alignment and coherence between standards, assessment and professional development for teachers. A clear vision of the evaluation framework in Mexico should allow for the distinct but complementary purposes of different assessments (i.e. ENLACE, EXCALE or possible school-based assessments), and how they should continue to develop in the future within a common national framework.
3.6	<i>With student performance data and appropriate growth models, low performers, high performers and cases needing follow-up observation can be identified.</i> As the assessment and evaluation process becomes more established, consequences such as incentives, further observation, and assistance to schools and teachers can be linked to the results. This implies the possibility of having multi-stage consequences and responses to the results. Schools determined to be repeatedly underperforming or performing near the top, for example, could be subject to on-site visits and reviews to identify potential causes and determine appropriate responses relating to improvement, technical support and the channelling of additional resources to under-performers.

4	Fair assessment of the value-added of schools: All students, regardless of socio-economic, ethnic or linguistic background, should have the same opportunities to learn and achieve at higher levels. Although student performance has been shown to be highly correlated with family background, results from assessments and evaluations should reflect the true contribution to students' learning and not the socio-economic context of the school or its students.
4.1	<i>Given the large diversity of educational contexts in Mexico, value-added models can offer a fair and more accurate measure of student growth and school performance.⁵ Current efforts by SEP and state educational authorities regarding the presentation and use of ENLACE results are a good starting point and could be built upon with value-added results for schools. The challenges involved in designing, planning and implementing an assessment system for accountability and school improvement that uses value-added modelling should be addressed rigorously throughout all stages, including the initial knowledge mobilisation, analysis and application phase of education reforms in Mexico.</i>
4.2	<i>Value-added models can offer a better option than raw test-scores to accurately and fairly identify the contribution of schools to student learning, by taking into account the context and background of the students. The technical challenges involved with developing an assessment and evaluation framework based on value-added modelling should be considered and addressed from the initial phases of design and planning. The robustness and frequency of the ENLACE assessment in Mexico, however, provides an invaluable opportunity.</i>
4.3	<i>Given the current conditions of the educational system in Mexico, value-added models can be based primarily on the school as the unit of accountability, although school zones,⁶ student groups, municipalities and states can all be used for analysis and action. Vertical equating should be among the first of the technical issues to be reviewed in the further development of ENLACE. The quality and availability of information that could be used for contextualised value-added models should also be assessed.</i>
4.4	<i>The first phases of the development of value-added modelling in Mexico should use actual student data to identify the weaknesses and strengths of different value-added models. Even before applying value-added methods to student performance data, however, schools could be grouped based on socio-economic contexts, and contextualised attainment models could be used as possible precursors of full-fledged value-added analysis. Therefore, the process of establishing value-added modelling can have different phases:</i> <i>i) Stratification of similar schools (based on type and socio-economic or other relevant information) for within-group comparisons of average results of raw scores. Issues regarding quality and completeness of test data and contextual information should also be identified and addressed.</i> <i>ii) Internal value-added modelling exercises conducted by education authorities to select models and address technical issues with data. A three-year-moving average is suggested for the modelling. In addition, education authorities could use VAM analysis to monitor and conduct evaluation trials of specific policies, programmes and jurisdictions, such as <i>Programa Escuelas de Calidad</i>, for example, with particular emphasis on differences within and between municipalities, school zones, states and ethnic groups, among others.</i> <i>iii) Public information, awareness and engagement with stakeholders on the merits, challenges and opportunities of value-added modelling, which could be linked to a re-launching of the ENLACE assessment, for example, with a clear plan for its further development.</i> <i>iv) Attributing consequences (low-stakes at first for under-performing schools (further exploration, observation and assistance), as well as for high performers. The same value-added analyses could be used by SEP and state education authorities to identify schools that may have teachers and practices worthy of replication and modelling. Logistical issues relating to test administration should also be addressed.</i>

5	Evaluation of teachers for accountability and improvement: Teachers are vital to student learning. It is difficult to improve, however, what is not measured. A fair and reliable in-service teacher evaluation process should provide incentives for teachers at all levels of the performance spectrum to improve, to be recognised and to contribute to overall educational results.
5.1	<i>Teacher standards should be developed to provide teachers with clear guidance as to what is considered good teaching practice.</i> Teaching standards could also be used in designing opportunities for professional development and improvement (training, modelling, observation, technical support, etc.). International examples and models of standards provide a useful starting point to further adapt and develop appropriate teaching standards in Mexico.⁷ The standards developed by Mexico should meet the following criteria: i) cover all of the defined teaching domains; ii) establish different levels of competency for each specific aspect that defines the domains of teacher and school work; iii) reflect a core group of performance traits that should be observable in all teachers and all schools; iv) define and operationalise intended goals and outcomes of good teaching; and v) be dynamic to ensure proper scaling. The standards should also cover at least the following domains: use of instructional time (attendance, punctuality, time on task), planning and preparation (the design of instructional activities and evaluation procedures for all students), classroom environment (making the classroom a safe place for risk taking), instruction (adapted to different students, engaging and challenging), and professional responsibilities. Special care should be placed on the ability of the teacher to strive for equity: to attend to the needs of the diversity of students in order to achieve learning outcomes for all.
5.2	<i>Establishing consensus among stakeholders on the importance of developing a comprehensive, transparent and fair in-service teacher evaluation framework is vital.</i> Given the importance of teachers to student learning, an in-service teacher evaluation framework should be designed and planned for the short, medium and long term. The gradual building of capacities at different levels to ensure fairness and objectivity of the evaluation framework, including a cadre of well-trained external evaluators, should be considered. To facilitate acceptance and sustained reform, it is important to set up a <i>communication, engagement and consultation strategy</i> with primary stakeholders (including the general public, opinion leaders and teachers). Other important policy dimensions such as legal, regulatory and financial considerations⁸ can ensure that solid foundations for reform are established in a transparent and participatory manner, even if the complete teacher evaluation system is not implemented during the mandate of a single government administration.
5.3	<i>In the context of increased accountability and along with opportunities for capacity building and professional development for teachers, it is important to ensure that all teachers meet minimum levels of professional performance and results.</i> Growth in student learning should be at the heart of the evaluation process. In addition, however, issues relating to basic teacher effort, such as attendance, punctuality and time-on-task, can be included in the earlier stages of the teacher evaluation framework as a way of getting all teachers to perform at capacity. Basic criteria such as these can produce considerable and timely gains for the teacher evaluation system in a cost-efficient manner (<i>i.e.</i> ensuring that all of the “low-lying fruit” is collected first).

6	Incentives and stimuli for in-service teachers: Although performance rewards have been used effectively in other fields of employment, their recent use in the education sector, particularly for teachers, is still being explored, monitored and evaluated. Thus, SEP, state educational authorities and stakeholders will need to determine the specific combination of monetary and non-monetary incentives and stimuli that will be most effective in Mexico. Regardless of the rewards or consequences that are linked to results, however, for teachers to be considered effective, their students should demonstrate satisfactory levels of achievement growth, while no teacher should be rated as ineffective if students show satisfactory levels of achievement growth.
6.1	<i>For an effective and sustained in-service teacher incentives policy, the following five principles should guide its development:</i> <i>i) Incentives should reflect the quality of teaching. The criterion for success of the incentive programme should not simply be better pay for better performing teachers, but the contribution of teachers to improved student learning outcomes.</i> <i>ii) The incentive system should recognise and support the individual teacher, the team of teachers at the school and the profession as a whole. Incentives should be embedded in a system that supports the continuous improvements of students, teachers, schools and the education system. In the longer-term, incentives based on tests should be complemented by a sound human resource management capacity in schools and at local levels that can accurately assess the quality of work, with robust external validation and corroboration methods jointly owned by government and the teaching profession.</i> <i>iii) The incentive system should build on a sound understanding of what motivates teachers and should embrace multiple dimensions of motivation, with the aim to foster an attractive work environment, create and facilitate advancement along a career path, provide access to professional development, and identify and promote effective teaching practices. Incentives and stimuli should therefore consider financial and non-pecuniary incentives, such as working conditions, material inputs for schools and classrooms, social recognition, enhanced training and professional development opportunities, or a combination thereof.</i> <i>iv) The incentive system should provide good feedback mechanisms and access to professional development, to ensure that teachers who do not receive the incentive understand what they can do to improve performance and have incentives to change behaviour. It should foster a culture based on evidence and data.</i> <i>v) The incentive system should reward both good performance and relative improvement, and consider the value added by teachers and schools, net of socio-economic factors. While value-added analytical models are being developed, however, simpler methods can be employed to ensure that students, schools and teachers are compared with those in similar contexts (e.g. socio-economic stratification and/or contextualised attainment models).</i>
6.2	<i>It is important to clearly distinguish an in-service teacher-incentive policy from other teacher-related programmes that may appear to be similar, but that do not fundamentally provide incentives to teachers to improve performance.</i> Incentive policies should be communicated clearly to teaching professionals in advance of the assessments and measures that will be used for the awards. In addition, each eligible teacher should have a probability of being rewarded for outstanding performance that is greater than zero, which is currently not the case for similar programmes in Mexico. Of particular importance will be finding a balance between national guidelines for the incentives (and at least partial funding), and state-level flexibility and co-participation in resources (financial or otherwise) for incentives and stimuli to teachers. Finally, a pilot of possible incentive programmes is highly recommended to ensure viability and cost-effectiveness of policy design. Pilot exercises should be rigorously monitored and evaluated in order to be most useful and worthwhile, with a base line and as much control as possible (e.g. randomised or quasi-experimental trials, if conditions allow).

-
- 6.3** *In-service teacher incentives in Mexico should motivate individual teachers to improve performance, but use the school as the basic unit of accountability, given the current state and prospects of data systems and the quality of information available. With the data and systems currently available, local education authorities can develop measures to confirm and validate the eligibility of teachers for incentive awards (e.g. with on-site inspections and data validation of student, school and teacher information). As a robust and credible individual teacher evaluation system is developed, incentive policies could be modified to ensure that teachers are able to receive incentives individually in the future.⁹ For school incentives, schools should be made publicly accountable for the additional resources received. If schools have discretion over the allocation of resources provided by the incentives, mechanisms should ensure transparency and the progressive involvement of relevant stakeholders, including parents and local school councils.*
-
- 6.4** *Financial and non-financial incentives and stimuli to teachers should be based on a fair and adequate assessment and evaluation process. Given the diversity of the Mexican educational system, a valid and reliable assessment process to identify eligible teachers for incentives needs to be developed. The success of incentives is directly linked to the credibility and fairness of the assessment and evaluation process upon which they are based. Models that take into consideration the socio-economic diversity of Mexican students, as well as other factors that can largely influence student performance, such as Spanish as a second language and ethnicity, for example, should be used when making comparisons among schools and their teachers. Special-education schools and programmes, as well as pre-primary schools, could be evaluated on the basis of appropriate measures of teacher performance and student learning, where possible. Given the diversity between and within states, the incentives policy should also consider a relative premium for disadvantaged rural schools, as opposed to non-disadvantaged urban or rural schools. Incentives should also support continued improvement of schools and teachers across the entire performance spectrum.*
-

NOTES

1. The other publications are *Improving Schools: Strategies for Action in Mexico*, stemming from the work of the OECD Steering Group on School Management and Teacher Policy in Mexico, and *Evaluating and Rewarding the Quality of Teachers: International Practices*, edited by Susan Sclafani for the OECD. In addition, a Spanish edition of the OECD report on evaluating school contributions to student learning *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas* has been updated and produced through the Co-operation Agreement (OECD, 2010). There are also numerous working papers from invited experts and OECD staff that have contributed to the work of both Steering Groups.
2. Recommendations regarding support, capacity building and professional development for teachers, for example, are provided in the sister OECD publication *Improving Schools: Strategies for Action in Mexico*.
3. The unit of accountability refers to the level at which the effort, capacities and performance of students, teachers and principals are monitored, assessed and evaluated. Although students are assessed individually, for example, and teachers should be motivated at the individual level to improve their performance, results are grouped so that individuals are held collectively accountable at the school level.
4. The specific technical, administrative and logistical recommendations on further development of the ENLACE assessment are presented in Chapter 5.
5. A detailed discussion of the benefits, characteristics and design issues of value-added modelling is presented in the updated OECD 2010 publication available in Spanish: *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas*.
6. In Mexico, the school zone is an administrative designation of a group of schools for the purposes of supervision and administrative monitoring. Similarly, municipalities are one of the three basic jurisdictional units of government and can contribute significantly to infrastructure and material conditions of schools.
7. For example, C. Danielson's *Framework for Teaching*, Perrenoud (2004); Rewards and Incentives Group (2009); Ontario Ministry of Education (2009); Khim Ong (2008); and Singapore Ministry of Education (2006). Current development efforts in Mexico regarding standards should be considered and evaluated based on recommended criteria.
8. As suggested by the public policy framework presented in Chapter 2.
9. Regarding the appropriate amounts for incentives, a review of international programmes shows that individual teacher incentives can range from less than 1% to more than 360% of monthly salary (OECD, 2009), although experts suggest that between 4% and 8% of annual salary can be adequate for incentives to be meaningful but not cause unwanted behaviour.

CHAPTER 1

Mexico Responds to Education Challenges

To address persisting challenges in education, the Mexican government established clear policy priorities for education reforms in its *Education Sector Programme 2007-12* (SEP, 2007). To monitor progress towards achieving objectives, the Programme established improvement indicators for student achievement as measured by the national ENLACE assessment and PISA (SEP, 2007).¹ Other key indicators relate to the professional development of teachers, school empowerment, equity in educational opportunities, and reforms relating to content and curriculum. As an indication of its commitment to reform processes, the Mexican government established in 2008 the *Alliance for the Quality of Education* (*Alianza por la Calidad de la Educación*) with the national teachers' union, which helped define the focus of the Co-operation Agreement with the OECD (SEP, 2008). The purpose of the Agreement was to determine not only what policy changes were required in Mexico, but also how to design and implement policy reforms effectively, given local conditions, constraints and opportunities.

One of the elements of the Agreement with the OECD was developing appropriate policies and practices to evaluate the quality of schools and teachers, and to link learning outcomes to incentives for continuous improvement. This work was led by the OECD Steering Group on Evaluation and Teacher Incentive Policies, consisting of invited experts. The Steering Group focused first on teacher incentives, as this was one of the main priorities for the Mexican Ministry of Education (*Secretaría de Educación Pública*, SEP).

Table 1.1

Areas covered by the Alliance for the Quality of Education and the OECD-Mexico Agreement

Alliance	OECD-Mexico Agreement
Modernisation of schools	School management and social participation
Professionalisation of teachers and education authorities	Teacher selection
Students' well-being and personal development	Career paths
Students' preparation for life and work	Teacher incentives and stimuli
Evaluation to improve the quality of education	Evaluation

Note: Blue text indicates those strands of the Co-operation Agreement focused on by the OECD Steering Group on Evaluation and Teacher Incentive Policies.

Source: SEP, 2008.

Table 1.2

OECD methodology to support policy implementation and deliverables

Methodology	Outcomes
Comparative analysis	Field visits to the Mexican states of Nuevo León, Chiapas, Aguascalientes and Veracruz. Expert papers. Reports and publications on teacher incentives from a comparative perspective, the ENLACE assessment, policy frameworks, accountability, student learning outcomes to measure improvement, in-service teacher evaluation, teacher incentives, piloting, monitoring and evaluation. A state review on teacher incentive practices in Mexico. ²
Recommendations	Specific recommendations for Mexico in the above-mentioned areas.
Communication	International workshops, technical meetings, presentations and dissemination events.

In order to promote effective policy design and implementation, the OECD assigned several staff members to work on the Co-operation Agreement, including one person located in Mexico City to enhance communication and interaction with stakeholders. The Steering Group consisted of national and international experts in the field of education who developed a series of policy recommendations in consultation with the Secretariat. Throughout the process, there was continued engagement with stakeholders.

This report presents the main findings and policy recommendations developed by the OECD Steering Group on Evaluation and Teacher Incentive Policies and the OECD Secretariat over the course of the Co-operation Agreement. The report draws upon the results of international workshops and technical meetings with stakeholders in the Mexican education system, field visits, thematic reports from invited experts, and the stock of OECD research and knowledge. The report and related publications form part of the larger body of work that includes findings and recommendations from the OECD Steering Group on Teacher Policy and School Management. The other publications are:

- *Improving Schools: Strategies for Action in Mexico*, stemming from the work of the OECD Steering Group on School Management and Teacher Policy;
- *Evaluating and Rewarding the Quality of Teachers: International Practices*, edited by Susan Sclafani for the Steering Group on Evaluation and Teacher Incentive Policies; and
- *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas*, updated and edited specifically for the Co-operation Agreement by the Steering Group on Evaluation and Teacher Incentive Policies.

In addition to presenting the main findings and recommendations, this report aims to encourage the Mexican government to continue the policy processes already underway within the framework of reform.³ The OECD gathered extensive information during the course of the Agreement and consulted repeatedly with Mexican stakeholders. For the OECD, this Agreement has meant an enriching joint effort which provided an opportunity to co-operate with a member country such as Mexico to support reform efforts, as well as to engage in in-depth analysis directly with stakeholders.

The report presents findings and recommendations on how an appropriate framework for evaluation and teacher incentives can be established. Chapter 2 provides an overview of the knowledge available to decision makers regarding education reform efforts. The chapter presents a public policy framework to assist policy makers in Mexico and other countries when designing, planning, implementing and evaluating specific policy initiatives. Chapter 3 explores accountability as a policy driver for improving student learning outcomes and proposes specific elements that should be considered by educational authorities and stakeholders. Chapter 4 provides an overview of the most common measures of student learning outcomes and suggests that a complementary approach combining multiple sources of evidence can be anchored by quantitative measures. To illustrate this point, the chapter provides an overview of the main strengths, challenges and opportunities of the ENLACE assessment in Mexico, and concludes with specific recommendations for its further development. Chapter 5 provides an overview of the benefits, challenges and implementation issues relating to the use of value-added models for school improvement efforts and enhanced accountability focused on student learning and growth. Chapter 6 gives an overview of common issues and implementation challenges surrounding in-service teacher evaluation, and provides specific recommendations for Mexico. Drawing on international examples, Chapter 7 presents a broad survey of the main issues of teacher incentives, and provides recommended practices for the monitoring and evaluation of these types of policies. Each chapter begins with a general discussion that may be relevant for policy makers charged with improving student achievement in varied contexts, and concludes with specific considerations and recommendations for Mexico.

BACKGROUND

Mexico's education system has experienced accelerated growth since 1950, making coverage the main priority during most of the 20th century. The size of Mexico and its geographical characteristics make for a complex and challenging task in providing education. One quarter of the total population is in the normative age group for attending basic education. The basic education system⁴ is very large and is divided by type of school according to the size and location of the community, as well as ethnic background in the case of indigenous communities (INEE, 2010b) (Table 1.3 and Table 1.4).

Inequality is a reality for many children in Mexico. For example: nation-wide, 92.4% of children enter school at the appropriate age, but this varies between states, dropping to 82.7% in Chiapas, compared with 96.6% in Aguascalientes.⁵ Similarly, in Chiapas for every 1 000 students who entered basic education at the appropriate age for the academic year 2000/01, only 476 exited on time six years later (school year 2007/08), compared with 747 in Aguascalientes (INEE, 2010b). There are also clear differences between types of schools. Out of 1 000 students, 198 students drop out of general lower secondary schools after five years, while the figure is 241 for technical secondary schools⁶ (INEE, 2010b).

Location, ethnic background and poverty are also factors associated with inequality in school enrolment in Mexico. In general primary schools, children starting school over-aged⁷ make up 4.6% of the school population, rising sharply to 15.1% in indigenous schools. The differences between schools in urban and rural areas, as well as with different socio-economic contexts, are clear: 2.8% of students are over-aged in schools in urban locations with low poverty, 6.9% in schools located in high poverty urban areas, 5.6% in low poverty rural areas and 11.5% in high poverty rural areas.⁸

Table 1.3

Dimensions of the basic education system, school year 2008/09 (formal)

	Students 25.6 million (76% of total)	Teachers 1.1 million (66% of total)	Schools 222 000 (89.5% of total)
Pre-school	18%	19%	40%
Primary school	58%	49%	44%
Lower secondary	24%	32%	16%

Source: INEE, 2010b.

Table 1.4

Basic education by type of school, school year 2008/09

Pre-school	Primary school	Lower secondary
General 89% Indigenous 8% Community 3%	General 94% Indigenous 6% Community 1%	General 51% Technical 28% Telesecondary 20% Community 0.3%

Source: INEE, 2010b.

Challenges also exist in terms of the quality of education, both among schools and within schools. Results from PISA have placed Mexico among the lowest performers across OECD countries. Although there have been considerable policy initiatives since the 1990s, these have had mixed results (OECD, 2010b). At the beginning of the 1990s, Mexico began a rapid process of governance and public sector reforms. Issues of transparency, an informed society and increased accountability of the public sector began permeating the public agenda.

This suggests that Mexico has an important opportunity to establish the basis for sustained education reform efforts. Increased accountability of schools, teachers, and educational authorities, clearly defined content and performance standards, school management, an effective evaluation framework, and a comprehensive teacher policy could contribute to improving student achievement in Mexico.

As yet, evaluation is a fairly recent practice in Mexico (Zorrilla, 2003). It was only in 1993 that a type of teacher evaluation was introduced with the *Carrera Magisterial Programme*, first to serve as a supplement to the salaries for teachers, and then to provide the possibility of progression in the teaching profession. It is only since 2006 that census-based standardised testing has been applied to students. The analysis reflected in this report refers back to these important milestones and then considers the opportunities that Mexico has to improve the quality of education in the country, particularly in the areas of student assessment⁹ and teacher evaluation.¹⁰ The creation of the National Institute for Educational Evaluation (INEE) in 2002 responded to the increasing social demand for an independent body to carry out reliable evaluations of the education system (INEE, 2006).

In addition to the benefits to individuals and societies from increased educational achievement (OECD, 2010a), one of the underlying core issues that needs to be addressed in Mexico is the right to a quality education for all students. In this context, accountability measures and policies based on increased transparency and evidence are also important. There is still much to be done, however, for Mexico to ensure the right to quality education for all students. And many of the issues that need to be addressed have to do with socio-economic factors (INEE, 2010a):

- Populations attending different educational services show strong social, economic and cultural differences.
- Access to education is still an unresolved issue in the basic education system in Mexico:
 - One of the problems is that Mexico has no direct measurement of student attendance.
 - In 2008, of every 100 children between 6 and 11 years of age, 98 attended school; for those aged 12 to 14, only 92 in every 100; and for those aged 15 to 17, only 65 in every 100.
 - Among children aged 4 to 6 and 15 to 17, those coming from poor families have a greater risk of not attending school.
 - Child labour for more than 20 hours per week is still a cause for not attending school.
 - Non-attendance is more common in rural than in urban areas.
 - Poverty is still an obstacle to achieving universal coverage. The greater the poverty, the greater the number of children who do not attend school.
 - The smaller the locality, the larger the percentage of children who do not attend school.
 - Ethnic background also affects attendance at school.
 - For ages 12, 13, 14 and 15, progress decreases and is regarded as unsatisfactory. Although less marked, this phenomenon is also true for primary education.
- The quality and relevance of learning is important to meet the requirements of a rapidly changing world:
 - By the end of lower secondary education, 7 out of every 10 students have not achieved the educational objectives established by the national curriculum (as measured by EXCALE).
 - Achievement of educational objectives is closely related to the nature of the school the student attends.
- Material and human resources also vary according to the location and type of school:
 - In primary education, the proportion of schools with at least one computer for educational use is 0.5% in community schools, 23.1% in indigenous schools and 55.7% in general schools.
 - The proportion of primary schools with electricity is 99.5% in urban schools, 81.1% in indigenous schools and 51% in community schools.
 - Schools with the largest percentages of teachers who have only middle-school studies or less are community and indigenous primary schools.

The challenge is not simple and solutions at different levels of the service delivery chain will need to be found. This requires a long-term vision that can guide a gradual process, but should start with immediate steps taken in the right direction.

At the beginning of the Co-operation Agreement, SEP requested assistance from the OECD in determining how to best reward teachers based on performance. This required a thorough review of international experience and evidence to identify a set of best practices that could be applied in Mexico. The results are presented in *Evaluating and Rewarding the Quality of Teachers: International Practices* (OECD, 2009), available in English and Spanish as part of the Co-operation Agreement. In addition, it was necessary to review elements of the teaching profession as a whole, including the selection of candidates through the initial training to retirement, with a special emphasis on in-service teachers. From international evidence it is clear that evaluation serves as an instrument for improvement, and there is a growing awareness among stakeholders in Mexico that the country needs to move in that direction. In order to investigate this further, expert papers were commissioned focusing on in-service teacher evaluation, summative and formative forms of assessments, accountability mechanisms and standards in basic education.

With the creation of INEE, Mexico has been systematically gathering test results along with other indicators of quality education such as coverage, efficiency, drop-out rates and repetition rates. This increased volume of data and knowledge on evaluation represents a significant basis for accountability. Mexico, however, still needs to improve the quality of the indicators it uses in relation to education, and also in the social and economic fields. The quality of the standardised tests needs important improvements to ensure comparability and to allow deeper analysis. Mexico also needs to create new instruments to measure educational outcomes and growth, in addition to standardised tests.

A great deal of the analysis undertaken by the Steering Group and invited experts also focused on the ENLACE system and on the opportunities and challenges that exist for its further development. ENLACE is arguably the largest and most important evaluation exercise in the country, due to the number of groups it covers and the number of years for which it has already been applied. The rich data sets that are produced by the assessment have the potential to support further policy developments. Detailed and rigorous technical analysis was conducted by invited experts on the characteristics of the ENLACE assessment. The results show that ENLACE is a robust instrument for measuring student achievement. As Mexico continues its reform efforts, this and other opportunities suggested by the report should be seized.

NOTES

1. It is important to note that the first indicator mentioned in the government's planning document is to improve Mexico's mean score in PISA from 392 (average of the two country mean scores in mathematics and reading from the 2006 assessment), to 435 by 2012. The second indicator relates to the percentage of students achieving a satisfactory level on the national ENLACE assessment (SEP, 2007, p. 15).
2. Analysis and specific recommendations on teacher policy and school management can be found in the sister OECD publication: *Improving Schools: Strategies for Action in Mexico*. This publication also includes a description of the main characteristics of the education system in Mexico.
3. Mexican authorities have undertaken several initiatives, such as *Programa de Transparencia y Rendición de Cuentas*, *Programa de Estímulos a la Calidad Docente*, *Escuelas de Calidad* and *Sistema Nacional de Información de las Escuelas*. Mexico is also currently working with the OECD on reviewing methodologies for the construction of indicators and their international comparability.
4. Basic education comprises pre-primary (three years), primary education (six years) and lower secondary education (three years).
5. Cohort of 1994, matriculated in 2000.
6. School years 2003/04-2007/08.
7. Two years or more above the normative age by grade and education level.
8. School year 2008/09.
9. ENLACE and EXCALE are the main standardised student tests applied in Mexico. The objectives and coverage of these tests are different. ENLACE is broadly applied as a census exam while EXCALE is sample-based.
10. *Carrera Magisterial* is a system of horizontal promotion based on certain attributes of teachers. It has proven ineffective in improving student learning outcomes. Teacher participation in *Carrera Magisterial* is voluntary.

References

INEE (2006), *Policies and Systems for the Assessment of Education: Achievements and Challenges*, Mexico.

INEE (2010a), *El Derecho a la Educación en México, Informe 2009*, Mexico.

INEE (2010b), *Panorama Educativo 2009*, Mexico.

Organisation for Economic Co-operation Development (OECD) (2010a), *The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving PISA Outcomes*, OECD Publishing, Paris.

OECD (2010b), *Improving Schools: Strategies for Action in Mexico*, OECD Publishing, Paris.

SEP (2007), *Programa Sectorial de Educación – Secretaría de Educación Pública 2007-2012*, Government of Mexico, accessed 24 April 2010 at www.sep.gob.mx/wb/sep1/programa_sectorial.

SEP (2008), *Alianza por la Calidad de la Educación*, accessed 24 April 2010 at <http://alianza.sep.gob.mx>.

Zorrilla, M. Coordinadora (2003), *La Evaluación de la Educación Básica en México: Una Mirada a Contraluz*, Universidad Autónoma de Aguascalientes, Mexico.

CHAPTER 2

The Public Policy Framework for Implementing Education Reforms

2.1 Education policy reforms in an international context.....	28
2.2 Mobilising OECD research, international practices and national knowledge.....	30
2.3 Policy dimensions of basic education reform: Asking the right questions.....	32
2.4 Considerations for Mexico.....	34

In this era of information and knowledge-based economies, education systems all over the world are facing the challenge of providing quality education to all citizens as a means of opening opportunities for current and future generations alike. With the global economic crisis, however, governments now have to face this challenge using policy options that allow sustainable public finances while enhancing long-term economic growth and development. Education systems are therefore under increasing pressure to deliver performance in terms of student learning, equity in educational opportunities, and value for the public investment in education (OECD, 2010a).

In this context and in order to contribute to the current debate regarding effective and efficient education policy reforms, this chapter proposes a framework which countries and jurisdictions can use to calibrate and design the important initial stages of their reform processes. It outlines an evidence-based process of knowledge mobilisation, analysis and application to bridge the gap between international practices and evidence on the one hand, and effective policy design and implementation on the other (OECD, 2003, 2007a). The chapter begins with a review of some of the main trends and findings relating to education reforms in an international context. It then identifies some of the common challenges facing education systems and acknowledges the growing recognition that international benchmarking should be used to monitor and evaluate educational results. Drawing on international practices and evidence provided by the OECD and other research, the chapter presents common policy priorities that have provided insights for educational policy reforms, which are treated in subsequent chapters.

OECD and non-OECD countries, however, are increasingly searching not only for *what* topics policy changes should focus on, but also *how* to effectively design and implement education reform, given local conditions, constraints and opportunities. The chapter identifies a commonly overlooked but crucial element for countries to adapt and implement reforms more efficiently and effectively: reliable country-specific knowledge. When combined with international practices and internationally comparable evidence, local knowledge mobilisation can provide a vital link in adapting best practices for effective education reforms that are suited to national priorities and contexts. The proposed *Public Policy Framework for Education Reform* provides key policy areas to be considered for country-specific knowledge mobilisation and analysis once policy priorities have been identified by governments and stakeholders.

2.1 EDUCATION POLICY REFORMS IN AN INTERNATIONAL CONTEXT

For the current and future well-being of their citizens, countries must meet the demands of a rapidly changing, globally competitive international context that is driven by information and knowledge. Developed and developing economies therefore require education systems that provide students with the skills and knowledge they need to be effective and innovative, not only in local or national contexts, but also in globalised economies. Although quality and equity in education require considerable effort and investment, the improvement achieved in educational results can dramatically increase the benefits to public finances, societies and individuals (OECD, 2010a, 2010b; Hanushek and Woessmann, 2009a).

Recent economic modelling that relates cognitive skills to economic growth shows that small improvements in lower levels of education can yield very large benefits. With a modest improvement of 25 points in PISA scores over the next 20 years, for example, even a top-performing country such as Finland would benefit by an increase in gross domestic product (GDP), in present value terms, of USD 553 billion, while low-performing countries such as Mexico would increase GDP by nearly USD 5 trillion (OECD, 2010b, p. 23). Equity in educational opportunities is equally vital for future economic growth. If countries could ensure that all of their students perform at the minimum levels of PISA (400 point score) – including those from disadvantaged socio-economic backgrounds and different ethnic groups – the benefits would be even greater: nearly USD 15 trillion for Turkey, USD 72 trillion for the United States and USD 26 trillion for Mexico (OECD, 2010b, p. 26).

The benefits are also clear at higher levels of education. The average net public return from a tertiary education (USD 86 404 for a male) across OECD countries, for example, is nearly three times the public investment (OECD, 2010a, Indicator A8). In addition to other social benefits relating to health, employment, civic engagement and societal cohesion, higher education also brings lifelong benefits to the individual. Having a tertiary education in OECD countries, for example, represents a net financial return over an individual's working life of more than USD 145 000 for men and USD 92 000 for women (OECD, 2010a, Indicator A8).

Conversely, the costs of low educational performance and the corresponding missed opportunities for social well-being and economic development are also considerable. This is true for developing as well as wealthy economies. In Latin America, for example, one of the lowest-performing regions in international educational assessments, the lack of educational quality and equity can largely account for the limited economic and social development in the region (Hanushek and Woessmann, 2009b). For countries at the other end of the GDP spectrum, estimates show that if the achievement of students in the United States in 1998 had been the same as those of countries such as Finland and Korea, its GDP would have been up to USD 2.3 trillion higher in 2008 (McKinsey & Company, 2009).

Thus, in addition to national and local assessments, countries are increasingly focusing on the performance and characteristics of their education systems in comparison to other countries, which is made possible by international instruments such as PISA since 2000, the Teaching and Learning International Survey (TALIS) from 2009, and others.¹ The most recent studies on education reform have stressed the need for countries and local education systems to compare effectiveness and efficiency against international standards, through benchmarking to those of the highest-performing countries (OECD, 2010c; ECS, 2008; McKinsey & Company, 2007). International benchmarking in education implies that country-specific issues such as standards, teaching practices, professional development, assessment and incentives are aligned to international best practices (OECD, 2009a; ECS, 2008).²

Education reforms are vital, but challenging. Past and current education reforms in several developed and developing economies within and outside of the OECD (e.g. Japan, Mexico, the United States, Brazil and China) have met with resistance and controversy. Commonly cited shortcomings of reform efforts, in addition to their sheer lack of results, are that reforms reflect short-term priorities in education rather than longer-term improvements, that they are implemented too hastily and without sufficient piloting and local adaptation, without effective assessment and evaluation systems, and that they exclude important stakeholders such as teachers and parents (Canadian Council on Learning, 2009; OECD, 2009a). In addition, some of the factors that contribute to the particular complexity of education reforms are the following:³

- Large investment: Public spending on education institutions across OECD countries was 6.2% of GDP, with public and private investment in all levels of education increasing between 1995 and 2007 by at least 8% in real terms (OECD, 2010a, Indicator B2).
- Size of sector: Teachers are a particularly large occupational group, representing approximately 4% of employment in OECD countries in 2007 (OECD, 2010c). Not surprisingly, the largest trade union in Latin America, for example, is a national teachers' union (Santibañez and R. Jarillo, 2007).
- Nearly universal: Education is experienced by almost everyone, to varying degrees, and the sheer magnitude of services rendered daily (OECD, 2010a, Indicator C1) makes it one of the most socially visible sectors.

These factors, in addition to the local conditions, interest and constraints that may apply in a given country, imply that attempts to introduce even modest reforms in the education sector can prove challenging.

International experience, however, also shows that improvement is possible. Countries such as Finland, Japan, Korea and Poland show that high and equitable learning outcomes are possible, as well as rapid improvement (OECD, 2001, 2004b, 2007d; McKinsey & Company, 2007). Examples of successful education

reforms also reflect international trend changes away from merely increasing coverage of services to ensuring quality (e.g. in Brazil and Mexico), from focusing on educational inputs (Hanushek, 2003), towards a focus on the results of investments in education (e.g. the United States). The international focus on student learning outcomes, as one of the key criteria to measure success, has also highlighted another issue: increasing public spending does not necessarily translate directly into increased learning outcomes. Although the United States has tripled average real per-pupil spending since 1960 (Hanushek, 2003), for example, educational achievement has been stagnant on both PISA and the national assessment (OECD, 2001, 2004b; NCES, 2003).

Education reforms are currently subject to strained public budgets, with deficit burdens for even wealthy economies oscillating around 7.8% of nominal GDP in 2010 (OECD, 2010d). In the midst of the global economic crisis, efficiency of education systems, as well as effectiveness, will continue to be a priority. For countries with increasing deficits and limited public budgets, it will simply not be possible in the near future to increase salaries for all teachers, for example, to make them commensurate with the status and prestige they enjoy in countries with higher performing education systems (e.g. Korea, Ireland and Singapore), that also boast some of the best paid teaching professionals (OECD, 2010a, Indicator D3; Sclafani, 2008; Murphy and Coolahan, 2003). Increasing consideration is being given to implementing other forms of incentives to motivate teachers to improve teaching practices and achieve higher levels of learning gains for all their students, as well as attracting and retaining the best and brightest into the teaching profession (OECD, 2009a).

If countries and education systems are to adapt and implement policy reforms tailored from international best practices, however, local conditions, constraints and opportunities must be adequately addressed. Studies of reform processes undertaken in multiple regions, countries and contexts have reached similar conclusions (OECD, 2010c, 2009a, 2007a; The World Bank, 2008a, 2008b; IDB, 2006, 2008, 2010). The process of achieving this, however, is still largely a matter of trial and error for governments and decision makers.

2.2 MOBILISING OECD RESEARCH, INTERNATIONAL PRACTICES AND NATIONAL KNOWLEDGE

*We have a comparative advantage on spotting the “what” of reform.
[...] But now we have been asked to help countries also with the “how” of reforms...*

Angel Gurría,
OECD Secretary-General
September 2008, Mexico⁴

Countries in the midst of education reforms focused on student performance, equity in educational opportunities,⁵ and increased value for investment, therefore, must not only determine what policy priorities to adopt, but also how to implement them. Policy makers and stakeholders look to international practices and research evidence to assist them in the design, planning and implementation of policy reforms. International organisations such as the OECD are increasingly being asked by member countries and partners to provide an analysis of state-of-the-art education policies and reform processes.

As the OECD and other international expert bodies attempt to respond effectively to such requests for guidance, they must sift through and distil a significant stock of recent information, data and knowledge. Through original research and project-specific work, the knowledge readily available on education can be classified as: indicators-based national profiles and international comparisons (e.g. the yearly editions of *Education at a Glance*); international assessments and surveys (e.g. PISA, TALIS, the Programme for the International Assessment of Adult Competencies or PIAAC); topic-specific research (e.g. on teacher policy, vocational training and on the costs of low educational performance); country-specific reviews, thematic reviews and notes (e.g. migrant education in Denmark); and reviews of international practices (e.g. *Evaluating and Rewarding the Quality of Teachers* and best practices for school value-added models).⁶

Despite the considerable accumulation of knowledge, OECD member countries express the need for a dynamic, real-time, country-specific tool that can direct policy makers and practitioners to the most up-to-date reforms, evidence and implementation strategies. This need arises from the difficulty of clearly linking policies and reforms to outcomes, particularly given the relatively long time horizon required to assess the impact of changes in the education sector (OECD, 2010c). Few conclusive research findings on specific policy topics exist, which puts further pressure on the OECD and other international bodies to leverage the cumulative stock of knowledge to highlight areas where further research is needed (OECD, 2007a). The most recent efforts within the OECD in this area reflect not only the demands of rapidly changing globalised economies (*i.e. OECD Skills Strategy*), but also the knowledge demands of countries undertaking reforms (*i.e. Mexico Co-operation Agreement and Leveraging Knowledge for Better Education Policy – GPS Project*).

International practices, research and evidence, as well as a range of economic and political shocks (OECD, 2010c), have combined to shape an education reform agenda for basic education. Although there is still much debate on specifics and further research is needed, there appears to be a convergence on some of the core elements related to quality, equity and value for investment in education. Depending on the national context and local priorities, the emerging core reform elements cluster around:

- Accountability – focusing on different levels of the education service delivery chain (system, institutions, schools and teachers, and students).
- Standards, assessment and curriculum – including measures of student learning and growth, content and performance standards for students and teachers, formative and summative assessments, alignment and coherence.
- Teacher policies – including evaluation, professional development, incentives, education and recruitment.
- School leadership evaluation and improvement – including assessments of net contributions to student learning, issues of autonomy, management, parental participation, competition and school choice.
- Incentives and stimuli – including those for jurisdictions (*e.g.* state bodies and federal funding), schools and teachers.

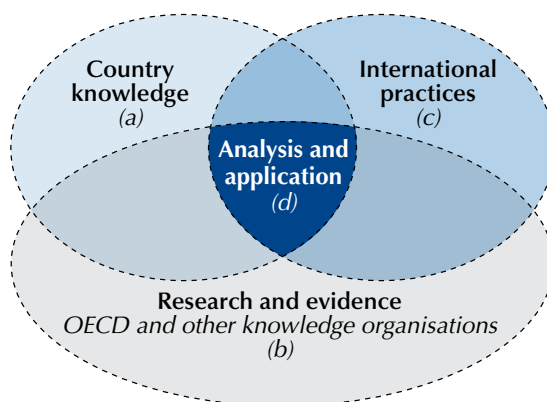
Nevertheless, a list of education priorities and levers – the *what* of potential interventions – is not enough to design, plan and implement effective policy reforms in any one or a combination of these areas. Even when best practices, research and evidence are available from the OECD and other sources, country-specific conditions, capacities and costs can severely constrain their usefulness (OECD, 2010c, 2009a, 2007a; IDB, 2010, 2008, 2006; Spiller, Stein and Tommasi, 2003; The World Bank, 2008a, 2008b). Country-specific factors are particularly important to any education reform because there is no universal model or specific set of policy instruments that can generate a consistent outcome across the board, even for particular topics such as fiscal instruments (OECD, 2010c), accountability mechanisms,⁷ or teacher incentives and stimuli (OECD, 2009a). Moreover, policy design, consultation, planning, piloting, implementation and rigorous evaluation are frequently at odds with the fixed-term mandates and limited budgets of most governments. OECD and non-OECD countries have explicitly recognised the importance of local research and knowledge in designing evidence-informed policy (OECD, 2007a).⁸ Thus, the *how* of reforms requires a robust heuristic model of policy reform processes that considers country-specific knowledge analysed in conjunction with research, evidence and international experience. A schematic representation of this “Knowledge MAP” (Mobilisation, Analysis, and Application) is presented in Figure 2.1. The combined knowledge from different sources (*a*, *b* and *c* in Figure 2.1) on specific policy topics can then be analysed, processed and ultimately applied for effective implementation and sustainable outcomes (*d* in Figure 2.1).

In this heuristic model of knowledge mobilisation, the challenge is to marshal relevant and specific knowledge in the most timely and cost-effective manner in order to adapt, design and implement education policies

to fit a country context. Rigorous analysis of this data can help identify the most suitable interventions and implementation methods to adapt and incorporate into policy reforms at the country level. This occurs in area (d) in Figure 2.1, where evidence-based knowledge from different sources can be analysed and applied in light of local realities and timing considerations (*i.e.* administrative mandates).

Figure 2.1

**Schematic representation of effective Knowledge Mobilisation,
Analysis and Application (Knowledge MAP):
Country-specific heuristics**



2.3 POLICY DIMENSIONS OF BASIC EDUCATION REFORM: ASKING THE RIGHT QUESTIONS

In spite of the long time horizon necessary to introduce policy reforms, government administrations generally work within shorter timeframes and often have to make policy decisions relatively quickly, with limited evidence and resources in challenging – if not outright contentious – contexts (OECD, 2010c). In addition, unless previously planned, public finances seldom allow large outlays for *ex ante* research studies on specific topics to reach better-informed policy design decisions. Although most countries have fairly robust and effective tracking systems for public accounts across the different branches of government (*e.g.* national accounts, legislative records, and statistical and evaluation entities), the plethora of registries, data banks and statistics available within the education sector and throughout related areas (*e.g.* across ministries, horizontally and vertically), may also make it more difficult to identify the most relevant and usable data and information for timely decisions on policy design.

The OECD has therefore drawn up a *Public Policy Framework for Education Reform* to help member countries not only to pinpoint the relevant national data on any topic that is vital for policy design, but also to identify those areas that may need further methods of study and systematisation. The Framework addresses six policy areas that shed light on the conditions, constraints, opportunities and costs in a particular country or jurisdiction. The Framework addresses the specific policy priority or particular topic that requires informed decisions for design, planning and implementation. Basic indicators need to be developed to ensure the information gathered on any given topic is relevant and applicable to the policy issue being analysed. When possible, these indicators should be linked to internationally comparative data available on other high-performing systems, and for benchmarking. The Framework provides local governments and stakeholders with a method to mobilise local knowledge and analysis on particular issues. The six policy dimensions reflect

the key components that have been shown to either facilitate or hamper reform efforts (OECD, 2010c, 2009a, 2007a; IDB, 2010, 2008, 2006; Spiller, Stein and Tommasi, 2003; The World Bank, 2008a, 2008b):

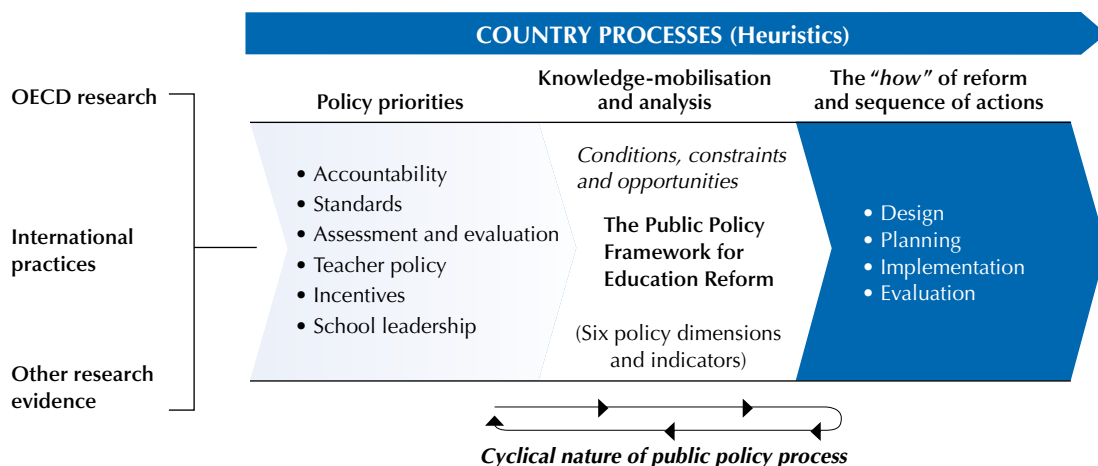
- Quality and quantity of relevant *information and data available* (on students and teachers, performance, possible linkages and for baseline and target setting).
- *Communication, engagement and consultation strategy* with stakeholders (including the general public, opinion leaders, professional organisations and teachers). Reform efforts that are not seen as being relevant to the average family or to society as a whole, for example, may be more susceptible to failure during the implementation process and as difficulties arise (OECD, 2009a, 2008).
- Amount and consistency of *public funding* for development and implementation of the policy reform (e.g. annual or fixed), including potential cost-benefit analysis, cost projections, and possible economies or reallocation of funds from existing programmes.
- *Legal, regulatory and administrative framework*, potential conflicts and possible modifications (including relevant labour and education contract laws, parental participation, information and privacy laws).
- *Institutional arrangement* of mandated public institutions (for evaluation, curriculum, research, etc.).
- *De jure and de facto decentralisation process* across entities responsible for educational services (resources, services and capacities).

Although countries may have varying capacities regarding knowledge generation, mobilisation and management, it is often basic and readily available information (“the lower hanging fruit”) that may prove most useful and least costly to collect. Academic, government or internationally supported bodies can provide assistance to any country or jurisdiction that requires support to carry out the necessary knowledge mobilisation, particularly in quantitative methods (OECD, 2003). With updated, timely and reliable knowledge on specific policy topics, any country or jurisdiction can carry out a robust process of policy adaptation, design and planning. In combination with other sources of evidence and policy insights (international practices, OECD and other research), policy design that takes into account best practices can address local conditions, constraints, costs and opportunities (Figure 2.2).

Figure 2.2

**Schematic representation of a country-specific heuristics model:
The Public Policy Framework for Education Reform**

Performance, equity and value for investment in education



The purpose of the Policy Framework in this model is to provide governments and stakeholders with a specific method for knowledge mobilisation at the local level, focused on the particular policy priorities being considered or established. It is intended, therefore, to ensure that policy makers and stakeholders can base decisions on international practices, evidence and advice, but also on the most updated and relevant local knowledge available.⁹ The six dimensions of the policy process can also be used as a checklist to ensure that policy and programme proposals that offer “answers” to policy reformers are based on the right set of questions (i.e. from the six policy dimensions). In policy making, as in scientific research, the relevance of the results are directly linked to starting off with the right questions.

2.4 CONSIDERATIONS FOR MEXICO

The Mexican government established clear policy priorities for education reforms in its *Education Sector Programme 2007-12* (SEP, 2007). To monitor progress towards achieving objectives, the Programme established improvement indicators for student achievement as measured by the national ENLACE assessment and PISA. Other key indicators relate to teacher professional development, school empowerment, equity in educational opportunities, and reforms relating to content and curriculum. As an indication of its commitment to reform processes, the Mexican government established in 2008 the *Alliance for the Quality of Education* with the national teachers’ union, which helped define the focus of the Co-operation Agreement with the OECD. It is important to note that SEP has undertaken several reform initiatives that are currently in progress, as mentioned in the preceding chapter, but that will need to continue developing. In this context, this chapter aims to provide SEP and relevant stakeholders in Mexico with recommendations on how policy priorities can be further informed and strengthened to ensure a lasting reform process.

- In addition to international practices and available research evidence, **country-specific knowledge mobilisation** on particular policy issues is a vital element to effectively design, plan and implement educational reforms that are viable and sustainable given the conditions, constraints and opportunities in Mexico. OECD work in this area, and evidence from different regions and across multiple contexts, clearly suggest that reliable and up-to-date knowledge about particular policy topics is crucial in the process of adopting best practices and policy recommendations. Knowledge mobilisation focused on relevant policy dimensions can occur fairly quickly (e.g. 6-12 months) at the country level and would greatly contribute to the analysis, design and planning of policy initiatives gleaned from international practice and research evidence.
- Although there are a number of policy issues surrounding reforms aimed at enhancing student learning and school performance (e.g. standards, assessment, evaluation, incentives and professional development), **there are common policy dimensions that can guide the process of local knowledge mobilisation**. The policy dimensions provide a framework through which SEP and stakeholders can continue the process of policy analysis, design, planning, implementation and evaluation. Current and future education reform efforts in Mexico would benefit from updated local knowledge regarding each of the following six policy dimensions:
 - i) Quality and quantity of relevant data and information available (on students and teachers, performance, and linkages between them), for target setting and to identify deficient areas that need to be addressed.
 - ii) Strategy options for communication, engagement and consultation with primary stakeholders, including the general public, teachers, principals and local education authorities (includes identifying how the proposed reform can be translated into a socially relevant and meaningful message for the average family, teachers and principals).
 - iii) Amount and consistency of public funding for development and implementation of the policy reform (e.g. annual or fixed), including potential cost-benefit analysis, cost projections and economies that can be obtained by re-channelling existing budget items or programmes.

- iv) Institutional arrangement of mandated public institutions (SEP, INEE and state education authorities, for example, to identify specific bodies that should contribute to developing standards, evaluations and proposing modifications).
- v) Legal, regulatory and administrative framework, potential conflicts and possible modifications that may be required in related areas (e.g. labour laws).
- vi) *De jure* and *de facto* devolution process across the main federal and state bodies responsible for providing educational services (resources, information management, evaluation, supervision and the provision of educational services).
- To inform and advise education authorities and stakeholders, the designation of an ***inclusive, objective and credible entity charged with knowledge management and analysis*** can facilitate the policy development process. The areas responsible for policy analysis and design within SEP and related bodies should be strengthened to contribute to this entity that can be considered a type of brokerage agency at the national level. It is not necessary to create a new formal entity, but to designate a body to address challenges often met in education reforms:
 - The need for timely and informed review, analysis and opinion, drawn from its own experts, international practice and available research and evidence.
 - The need for credible, objective and evidence-based expert opinion on specific issues at the centre of assessments, evaluation and teacher incentives, for example, including oversight and consultation regarding monitoring and impact evaluations of key programmes.
 - The differences in decision-making times between government administrations with fixed terms (federal, state and municipal) and the requirements of long-term policy planning, piloting, evaluation and educational research. Entities can take a number of forms (e.g. consortium, clearinghouse, council, synthesis programme) and those established in Brazil, Canada, Denmark, Singapore and the United Kingdom may offer useful examples for Mexico.

NOTES

1. Others include the Trends in International Mathematics and Science Study (TIMSS), and regional efforts such as the UNESCO/LLECE assessments (1997 and 2006) in 16 countries in Latin America and the Caribbean.
2. As discussed in the following chapters, standards can provide an effective method to ensure that content, performance and teaching practices in an education system are internationally competitive.
3. Adapted from “Making Reform Happen in Education” in OECD, 2010c, p. 161.
4. From the opening remarks of OECD Secretary-General, Angel Gurría, delivered on 25 September 2008, in Mexico City, Mexico, entitled *The Art of Making Reform Happen: Learning from Each Other*.
5. The importance of equity for reform efforts is highlighted by research showing that although accountability policies can lead to higher results in student achievement, for example, even in contested contexts such as the United States, the improvement in learning gains is not equal among all student groups (Hanushek and Raymond, 2006).
6. Another important example of this kind of comparative approach is the 2007 McKinsey & Company report that underscored the importance of international benchmarking of student performance.
7. Evidence shows that policy approaches and design are important in achieving effective accountability policies to increase student learning (Hanushek and Raymond, 2006; Hanushek, Raymond and Rivkin, 2004). It is precisely local constraints, contexts and opportunities of a given country or educational system that largely determine the specifics of policy design.
8. It is important to distinguish between research commissioned to inform action (for evidence-informed policies) and purely scientific research that, along with different standards and burdens of proof of causality (OECD, 2007a), may also have much longer timeframes for results.
9. This would also highlight contradictory information and evidence that may be available locally. In these cases, further consideration would be a necessary and valuable step for education authorities and stakeholders.

References

- Canadian Council on Learning** (2009), *“Changing Our Schools: Successful Educational Reform”*, Ottawa, ON.
- Education Commission of the States (ECS)** (2008), *From Competing to Leading: An International Benchmark Blueprint*, Denver, CO.
- Hanushek, Eric A.** (2002), “The Long Run Importance of School Quality”, NBER Working Paper No. W9071, July.
- Hanushek, Eric A.** (2003), “The Failure of Input-Based Schooling Policies”, *Economic Journal*, Vol. 113, No. 485, February, pp. F64-F98.
- Hanushek, Eric A., Margaret E. Raymond and Steven G. Rivkin** (2004), *Does It Matter How We Judge School Quality?*, paper presented at the American Education Finance Association Annual Meetings, Salt Lake City, Utah, 11-13 March.
- Hanushek, Eric A. and L. Woessmann** (2009a), *Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation*, National Bureau of Economic Research, Working Paper 14633, Cambridge, MA.
- Hanushek, Eric A. and L. Woessmann** (2009b), *Schooling, Cognitive Skills, and the Latin American Growth Puzzle*, National Bureau of Economic Research, Working Paper 15066, Cambridge, MA.
- Hanushek, Eric A. and Margaret E. Raymond** (2006), “School Accountability and Student Performance”, *Regional Economic Development*, Vol. 2, No. 1, pp. 51-61.
- Inter-American Development Bank (IDB)** (2006), *The Politics of Policies: Economic and Social Progress in Latin America*, Harvard University Press, Cambridge.
- IDB** (2008), *Political Institutions, State Capabilities, and Public Policy: International Evidence*, Inter-American Development Bank, Working Paper Series IDB-WP-661.
- IDB** (2010), *Political Institutions, Policymaking, and Economic Policy in Latin America*, Inter-American Development Bank, Working Paper Series IDB-WP-158.
- King, E.M. and S. Cordeiro Guerra** (2005), “Education Reform in East Asia: Policy, Process, and Impact”, in *The World Bank, East Asia Decentralizes: Making Local Government Work*, pp. 179-207.
- McKinsey & Company** (2007), *How the World’s Best-Performing School Systems Come Out on Top*, Identifying Teacher Quality Project, Washington, DC.
- McKinsey & Company** (2009), *The Economic Impact of the Achievement Gap in American Schools*, McKinsey Social Sector Office, NY.
- Murphy, I. and J. Coolahan** (2003), “Attracting, Developing and Retaining Effective Teachers: Country Background Report for Ireland”, by permission on the OECD Home Page, OECD Publishing, Paris.
- National Centre for Education Statistics (NCES)** (2003), “NAEP Trends in Academic Progress Through 1999”, in *NCES Digest of Education Statistics 2003*, NCES, Washington, DC.
- Organisation for Economic Co-operation and Development (OECD)** (2001), *PISA 2000: Knowledge and Skills for Life*, OECD Publishing, Paris.
- OECD** (2003), *New Challenges for Educational Research*, OECD Publishing, Paris.
- OECD** (2004a), *Innovation in the Knowledge Economy: Implications for Education and Learning*, OECD Publishing, Paris.
- OECD** (2004b), *PISA 2003: Learning for Tomorrow’s World*, OECD Publishing, Paris.

- OECD** (2007a), *Evidence in Education: Linking Research and Policy*, OECD Publishing, Paris.
- OECD** (2007b), *No More Failures: Ten Steps to Equity in Education*, OECD Publishing, Paris.
- OECD** (2007c), *Getting it Right: OECD Perspectives on Policy Challenges in Mexico*, OECD Publishing, Paris.
- OECD** (2007d), *PISA 2006: Science Competencies for Tomorrow's World*, OECD Publishing, Paris.
- OECD** (2008), *Measuring Improvements in Learning Outcomes: Best Practices to Assess the Value-Added of Schools*, OECD Publishing, Paris.
- OECD** (2009a), *Evaluating and Rewarding the Quality of Teachers: International Practices*, OECD Publishing, Paris.
- OECD** (2009b), *Education Today: The OECD Perspective*, OECD Publishing, Paris.
- OECD** (2010a), *Education at a Glance 2010: OECD Indicators*, OECD Publishing, Paris.
- OECD** (2010b), *The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving PISA Outcomes*, OECD Publishing, Paris.
- OECD** (2010c), *Making Reform Happen*, OECD Publishing, Paris.
- OECD** (2010d), *Economic Outlook No. 87*, OECD Publishing, Paris.
- Santibañez, L. and B. Jarillo** (2007), "Conflict and Power: The Teachers' Union and Education Quality in Mexico", *Well-Being and Social Policy*, Vol. 3, No. 2, pp. 21-40.
- Sclafani, S.** (2008), "Rethinking Human Capital in Education: Singapore as a Model for Teacher Development", paper prepared for The Aspen Institute Education and Society Program, Aspen Institute, Washington, DC.
- Spiller, Pablo T., E. Stein and M. Tommasi** (2003), "Political Institutions, Policymaking Processes, and Policy Outcomes: An Intertemporal Transactions Framework", Latin American Research Network (IDB).
- The World Bank** (2008a), *Raising Student Learning in Latin America: The Challenge for the 21st Century*, Washington, DC.
- The World Bank** (2008b), *The Road Not Travelled: Education Reform in the Middle East and North Africa*, Washington, DC.
- Woessmann, L. and T. Fuchs** (2004), *What Accounts for International Differences in Student Performance? A Re-Examination Using PISA Data*, IZA Discussion Paper No. 1287, CESifo Working Paper Series, No. 1235.

CHAPTER 3

Accountability as a Policy Driver for Improving Student Learning Outcomes

3.1 Types and features of educational accountability systems	40
3.2 Teacher performance: Towards a fuller understanding of accountability	42
3.3 Considerations for Mexico	44

Governments have increasingly focused on accountability systems as part of larger reforms in public sector management (OECD, 2009a). In the education sector, this has been reflected in a move away from a focus on inputs and processes to one which focuses on outcomes (Hanushek and Raymond, 2006; Hanushek, 2002). As the ultimate purpose of education systems is student learning, this naturally implies that actors should be held accountable to students' achievement. Paradoxically, increased accountability also implies increased autonomy of actors in terms of how resources provided to them are deployed to realise the objectives to which they are held accountable. Accountability in this sense implies that actors should publicly demonstrate that they are effectively pursuing established goals, at the same time that they are given the capacity, support, and sufficient autonomy to do so.

The purpose of this chapter is to provide an overview of the main features of accountability systems in education and how these can serve to support efforts to improve student achievement. The first section begins by presenting the main types and features of accountability systems. Based on a consideration of the common characteristics of effective educational accountability systems, it is clear that both content and performance standards are important.

Drawing from illustrative international examples such as Australia, Canada, Finland and Hong Kong-China, this chapter argues that a broader concept of accountability is needed upon which to build improvement efforts. Effective education systems have gone beyond a narrow notion of "test-based" accountability associated with standards-based reform (Darling-Hammond, 2004). A broader notion of accountability has direct applications with regards to teacher performance and effort. The chapter concludes with considerations for education authorities and stakeholders in Mexico interested in increasing the performance of Mexican students, schools and teachers.

3.1 TYPES AND FEATURES OF EDUCATIONAL ACCOUNTABILITY SYSTEMS

The nature and characteristics of educational accountability depend on the established goals. As such, there are at least five broad categories of accountability that can be considered (Darling-Hammond, 2004; Ladd, 2007; NAE, 2009):

- **Political accountability:** Elected officials and education authorities named by officials are accountable to the local constituencies that elected them to office (e.g. recent controversy in New York City over assessment).
- **Legal accountability:** Schools, teachers, and local bodies such as school councils are obliged to follow the established legal and legislative framework.
- **Administrative accountability:** Different jurisdictions responsible for oversight, supervision, or service delivery of educational services will establish regulatory and administrative procedures to ensure that services and schools comply with expected performance (e.g. recent reform efforts in the United States).
- **Professional accountability:** Teachers, principals, and other school staff must meet basic requirements with regards to specialised knowledge, entry into the profession, and standards of expected teaching practices and performance. Teachers feel "laterally" accountable to their fellow teachers and school principals, as well as to parents.
- **Market accountability:** This is often tied with the degree of school-choice and the funding and other consequences for the school linked to family and student preferences.

Depending on the stated objectives and policy priorities, certain types of educational accountability will be focused on by educational systems. Practices from several countries suggest that reform efforts in recent years have started to diversify to include not only legal and bureaucratic accountability (Darling-Hammond, 2004), but also professional and, increasingly, market-based approaches (OECD, 2005, 2007a; Baker, 2010).

Evidence and international experience do suggest, however, that accountability systems focused on student learning should (OECD, 2009a, 2007a; NAE, 2009; ECS, 2008):

- a) be built on clear expectations of the desired outcomes;
- b) use adequate measures, both formative and summative, to assess progress in meeting targets;
- c) be supported by adequate data and information systems;
- d) include reporting mechanisms for public accountability and transparency; and
- e) be linked to actions as a result of performance.

Clear expectations of education systems should be reflected in clear content and performance standards for students and teachers (OECD, 2009a, NAE, 2009). In addition, issues of alignment and coherence between standards, assessments, and professional development for teachers are important to ensure that all of the elements for accountability and improvement are working together (Baker, 2010; NAE, 2009). This implies that a complementary approach to assessment that combines both summative and formative measures are important, as evidence by some of the best performing systems such as Australia, Finland and Alberta, Canada (Darling-Hammond and McCloskey, 2008; Government of Alberta (Canada) – Education, 2009). Even in cases where content standards are developed after assessments are in place, as was the case in Australia with the annual National Assessment Program in Literacy and Numeracy and the Australia National Curriculum, for example, alignment should be sought between curriculum (content), performance standards, teaching practices, and assessments (VCAA, 2010). In an interesting example of co-operation between educational systems with regards to best-practices in assessment, Hong Kong-China and Queensland, Australia have similarly focused approaches (Darling-Hammond and McCloskey, 2008). In Finland, the national curriculum provides teachers guidelines so that teachers can develop school-based assessments, without external standardised assessments to rank students or schools. In this context, Table 3.1 therefore presents a summary of common types of assessments and their uses in these and other accountability systems.

Table 3.1
Common types of assessments used in accountability systems

Type of Assessment	Purposes	Measures	Strengths	Weakness or Cautions
Classroom-based formative assessment	To improve student learning based on diagnostic information given as needed.	Teacher-made, curriculum embedded with diagnostic grain size. Teacher marked.	Relevant to instruction; uses teacher knowledge; potentially supportive; permits the use of extended student constructive response of greater validity.	Depends upon teachers' capacity to design assessments, to interpret findings and to apply appropriate remedies. Adaptive to students, but probably not comparable; creates classroom management difficulties.
Externally provided formative assessment	To improve students' learning and to give diagnostic information.	Externally provided; teacher or externally marked.	Relieves teacher of assessment design requirements; well aligned with standards; may allow extended responses.	May be inappropriately timed for many students; may be mismatched with curricula as practiced. May not be well understood by teachers.
Summative assessment	To report on status and growth of students.	Externally provided, usually marked by different teachers from those teaching students.	Low cost, if selected responses. Reliable; may be used in growth and value added models.	Poor transfer; inflated by teaching the test. Poor alignment; low validity.
Combined classroom and summative assessment for accountability	To report on status.	Weighted elements from formative assessments.	Includes more valid assessments; aligned; provides for transfer.	High cost to implement. Conflict of interest for teachers for teacher effectiveness use. Proportion of weighting problematic.

Source: Baker, 2010.

External assessments within accountability systems, however, are important. Multilevel analysis of results from PISA 2006 show that there is a positive correlation between student achievement and the publication of school results, as well as with the existence of standards-based external examinations (OECD, 2007b). With regards to adequate data and information systems, reporting mechanisms, and actions linked to performance, Chapters 4, 5, 6 and 7 of this report provide an overview of recommended practices and international examples. Despite the benefits of increased public accountability centred on student performance, however, narrow notions of accountability that reduce it to a one-dimensional “test-based” policy have limited its potential (Ladd, 2007). The following section, therefore suggests the need for a broader notion of accountability.

3.2 TEACHER PERFORMANCE: TOWARDS A FULLER UNDERSTANDING OF ACCOUNTABILITY

Based on the previous discussion, accountability can be understood as comprising three different aspects of the same concept. First, accountability in a given educational system can refer to a shared and collective sense in a society of responsibility for the learning and growth of students. This aspect of accountability relates to the importance and value that is ascribed to education in a given country, and the degree of priority it enjoys socially and in the public agenda (Baker, 2010). Second, accountability can also refer to “high takes” systems in which significant consequences to individuals, schools, or jurisdictions can result as part of the accountability system, (OECD, 2009a). Reforms in the United States, for example, have become synonymous with the debate regarding accountability across jurisdictions and actors (Ladd, 2007). There exists, however, a third aspect of accountability which can be considered “low-stakes” or basic accountability. This type of accountability is associated to the degree of effort shown by teachers and school staff in order to comply with basic, minimum standards of professional behaviour, such as showing up to school, arriving on time, and focusing on the tasks at hand. Given the importance of teachers to student achievement (OECD, 2009a), it is useful for systems to not only look at school-based accountability, but teacher accountability to professional standards.

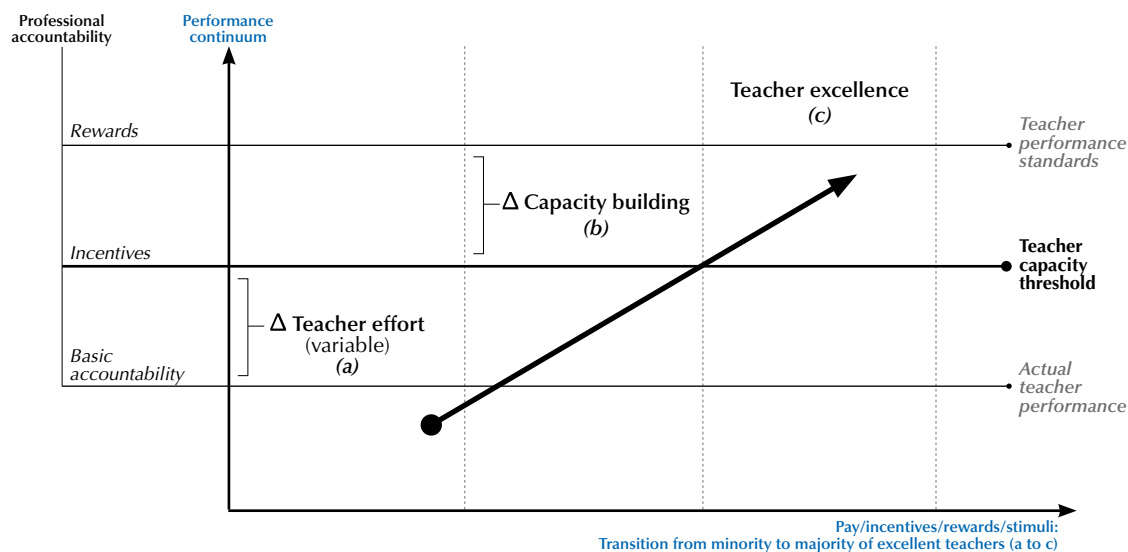
Within an educational system that has established performance standards for teachers, the common approach is to provide capacity-development opportunities to all teachers to assist them in reaching the performance standard. Different forms of incentives can be provided to teachers so that they are motivated to seek out opportunities for capacity building to be in a better position to achieve standards (Lavy, 2009). For those teachers that exceed the performance standards, larger rewards can be offered so that even high-performing teachers have incentives to continue improving their practice (OECD, 2009a).

In a conventional system of low-accountability, contract theory provides the basis for the common problem of information asymmetry between teachers and households or policy makers, leading to moral hazard (Ferreira and Liang, 2010). One of the common implicit assumptions of teacher evaluation systems focusing on formative assessment, capacity building and professional development is that actual teacher performance is equal to or very nearly equal to teacher capacity. The reality facing most educational systems, however, is that certain teachers will not exert sufficient effort, thereby performing considerably under their true capacity. Although these under-performing teachers (from lack of effort) may take advantage of and benefit from capacity-building opportunities and professional development, the most efficient scenario is one in which the difference between teacher effort and teacher capacity (“ Δ Teacher effort” in Figure 3.1) is minimised.

Teachers performing below their natural capacity would receive, in theory, just the basic mandated salary (area “a”). For those teachers that are performing at their natural capacity, incentives and capacity-building options will allow them, in theory, to reach the performance standards. Inefficiencies arise, however, when teachers that could naturally, through effort and natural capacity, achieve or surpass the performance standards, even without further capacity building, fail to do so. In this scenario, basic accountability would have to be strong enough so that the teacher is motivated to work harder (*i.e.* exert more effort), and also seek out capacity-building opportunities.

Figure 3.1

Schematic representation of the capacity-threshold concept of teacher accountability



Source: Developed from OECD (2009a) and Murphy (2001).

The optimal scenario in terms of resources and achievement is to have as many teachers as possible perform at their capacity threshold. This would automatically, without capacity building or further inputs from the system, increase average teacher performance and student performance. From this basis (*i.e.* capacity threshold concept), teachers will continue to be motivated to improve performance through capacity building. In this simplified model, the distance between the average performance of teachers and performance standards will vary depending on the overall quality of the teaching force. The usefulness of this model is significant for at least three reasons:

- i) As discussed in Chapter 6, one of the differences between educational systems with low-accountability (*e.g.* tenured teachers that cannot be separated from the profession), such as those that are found in Latin America, and other systems, is that in addition to attracting and keeping the best and brightest teachers in the profession (*e.g.* Singapore and Sweden), it is very important to improve and support in-service teachers.
- ii) With increased public and systemic commitment to student achievement, teacher performance becomes important. In a context of increased accountability, teachers will be expected to exert more effort as the norm to meet minimum professional behaviour.
- iii) Teacher behaviours that are more related to effort than capacity, such as attendance, punctuality, and time-on-task, can have important effects on student learning. Studies in the United States and developing countries have shown that teacher absences can have surprisingly large effects on student learning, particularly for students in the poorest and most remote communities. Das *et al.* (2007) report that for primary schools in Zambia, for example, a 5% increase in teacher absences reduced a year's learning of a typical student by 4 to 8%. Evidence also suggests that identifying discrepancies and the margin of error of official records is important. Evidence from Ecuador and India, for example, suggests that official data can differ from field-observations of teacher attendance/absence by between 7% and 25%. This is particularly relevant for certain states and regions of Mexico.¹ Results from the OECD TALIS survey suggest that teacher absenteeism and time-on task are important issues in Mexico (OECD, 2009b).

The capacity-threshold model thus allows clear linkages between accountability and teacher incentives, along a performance continuum. Moreover, there are substantial potential gains in average teacher performance and student learning by increasing accountability in a system so that teachers are working at their capacity. This has direct policy implications for the design of teacher assessment criteria as well as teacher and school performance measures. The ultimate goal of an accountability system that includes performance incentives, for example, would therefore be to have as many teachers as possible transition from *a* to *b* and eventually to *c* in Figure 3.1. This would also be reflected in the performance of systems as they transition from *poor* to *good* and from *good* to *great* in terms of their students' achievement.

3.3 CONSIDERATIONS FOR MEXICO

Based on the above considerations and on the specific conditions, constraints and opportunities in Mexico regarding public accountability centred on student achievement (Salieri *et al.*, 2010), the following are considerations for education authorities and stakeholders.

The importance of public accountability: All stakeholders should feel responsible and be held publicly accountable for student learning and overall educational results.

Actors should be held accountable for student learning and growth, but provided with the necessary assistance and capacity building. A clearly defined accountability system focused on the results of student learning and growth in achievement can provide the necessary coherence, given the size, complexity and multiple interests of the participants in Mexico's education sector.

The use of student learning as a key criterion against which state education authorities, schools, principals and teachers will be held accountable reflects a focus on outcomes rather than input-focused policy reforms. International practice regarding performance-based teacher incentives, for example, reflects this change. This does not imply, however, that issues of infrastructure or social inclusion are no longer important for the Mexican educational system. Rather it implies that learning and development for all students – fostered, cultivated, assessed and evaluated through various means – should be the ultimate goal of policy action and reforms. Support to students, schools, principals and teachers, as well as professional development, are vital complements to increased accountability.²

Accountability focused on student learning and growth implies establishing clear standards.

The development of standards as a key component of the accountability system focused on student learning should address at least three priorities: *i*) appropriate development of standards for content, student performance and teacher performance; *ii*) alignment and coherence between standards, assessment, evaluation and professional development; and *iii*) alignment of standards to international best practice and internationally competitive benchmarks of student knowledge and skills. Within a standards-based accountability framework, actors should have incentives to meet or exceed the expectations that are reflected in standards.

Accountability measures should include complementary criteria of effort as well as performance.

A standards-based accountability system for students, schools and teachers in Mexico should consider using measures of student learning and growth (from standardised assessments and other reliable methods, where possible), as well as complementary criteria regarding individual, group and school performance. This is important in Mexico as student and teacher attendance, punctuality and time-on-task remain important issues. An accountability system in Mexico should take into account the fact that some principals and teachers may not be performing to their fullest capacities. Incentives are needed, therefore, to increase basic effort and

performance, as well as supporting capacity building and professional development. Reduction of student drop-out rates, for example, can also be considered as an indicator. Accountability also implies that some teachers who receive adequate technical assistance and opportunities for professional development, but who do not improve performance, would be counselled to leave the profession.

The focus should be on students, schools and teachers for continuous improvement with the school as a basic unit of accountability.³

Although different levels and actors in the education system should be held accountable, the school can serve as the basic unit of accountability, with individual data, information and monitoring for students and teachers. Student and teacher data and information at the school level can be used to support improvement efforts, teacher incentives and stimuli, education interventions for low performers, and the identification of good practices for modelling and to inform the development of teaching standards, for example.

It is important to define a gradual process to develop complementary approaches of assessment using multiple sources of evidence.

Developing a robust standards-based accountability system is a gradual process, with clear stages, and with complementary approaches to assessment and evaluation. Both summative and formative assessments of student learning and growth, as well as school and teacher performance, should form part of the accountability system in Mexico. The development of such a system, however, should be delineated in stages with a thorough consideration of current and projected capacities, methods and costs.

NOTES

1. Alcázar *et al.* report that teacher absence rates in Peru increase from an average of 11% to 16% and 21% when considering poor and remote schools. Evidence from primary schools in Bangladesh, Ecuador, India, Indonesia, Papua New Guinea, Peru, Uganda and Zambia show that teacher absences can vary from 11% in Peru to 27% in Uganda (cited in Rogers and Vegas, 2009).
2. Recommendations regarding support, capacity building and professional development for teachers, for example, are provided in the sister OECD publication *Improving Schools: Strategies for Action in Mexico*.
3. The unit of accountability refers to the level at which the effort, capacities and performance of students, teachers and principals are monitored, assessed and evaluated. Although students are assessed individually, for example, and teachers should be motivated at the individual level to improve their performance, results are grouped so that individuals are held collectively accountable at the school level.

References

- Alcázar, L., H. Rogers, N. Chaudhury, J. Hammer, M. Kremer and K. Muralidharan** (2006), "Why are Teachers Absent? Probing Service Delivery in Peruvian Primary Schools?", *International Journal of Educational Research*, Vol. 45, pp. 117-136.
- Baker, E.L.** (2010), "Summative and Formative Evaluation in Educational Accountability", Working Paper prepared for the work of the Steering Group on Evaluation and Teacher Incentive Policies, UCLA, CA.
- Chaudhury, N., J. Hammer, M. Kremer, K. Muralidharan and H. Rogers** (2006), "Missing in Action: Teacher and Health Worker Absence in Developing Countries", *Journal of Economic Perspectives*, Vol. 20, No. 1, pp. 91-116, retrieved 12 April 2009, from www.atypon-link.com/AEAP/doi/pdf/10.1257/089533006776526058?cookieSet=1.
- Darling-Hammond, L.** (2004), "Standards, Accountability, and School Reform".
- Darling-Hammond, L. and L. McCloskey** (2008), "What Would it Mean to Be 'Internationally Competitive'?", Stanford University, accessed 10 May 2010 from http://edpolicy.stanford.edu/pages/events/kerner/materials/intnl_assessment_pdk.pdf.
- Das, J., S. Dercon, J. Habyarimana and P. Krishnan** (2007), "Teacher Shocks and Student Learning – Evidence from Zambia", *The Journal of Human Resources*, Vol. 42, No. 4, pp. 820-862.
- Education Commission of the States (ECS)** (2008), *From Competing to Leading: An International Benchmark Blueprint*, Denver, CO.
- Ferreira, M.M. and P.J. Liang** (2010), "Information Asymmetry and Equilibrium Monitoring in Education", Carnegie Mellon University.
- Glewwe, P., N. Ilias and M. Kremer** (2008), "Teacher Incentives in Developing Countries: Recent Experimental Evidence from Kenya", Working Paper 2008-09, National Center on Performance Incentives, retrieved 15 January 2010, from www.performanceincentives.org/data/files/directory/ConferencePapersNews/Glewwe_et_al_2008.pdf.
- Government of Alberta (Canada) – Education** (2009), "The Alberta Student Assessment Study: Final Report", Edmonton, Alberta.
- Hanushek, Eric A.** (2002), "The Long Run Importance of School Quality", NBER Working Paper No. W9071, July.
- Hanushek, Eric A. and Margaret E. Raymond** (2006), "School Accountability and Student Performance", *Regional Economic Development*, Vol. 2, No. 1, pp. 51-61.
- Hanushek, E.A. and M.E. Raymond** (2006), "School Accountability and Student Performance", Federal Reserve Bank of St. Louis, *Regional Economic Development*, Vol. 2, No. 1, pp. 51-61.
- Hong Kong Examinations and Assessment Authority** (2010), Press Release: 2010 Hong Kong Advanced Level Examination Results Released, accessed 20 June 2010 from www.hkeaa.edu.hk/DocLibrary/Media/PR/20100629_HKALE_Results_ENG.pdf.
- Ladd, H.F.** (2007), "Holding Schools Accountable Revisited", Association for Public Policy Analysis and Management, Washington, DC.
- Murphy, K.** (2001), "Performance Standards in Incentive Contracts", *Journal of Accounting and Economics*, Vol. 30, pp. 245-278.
- Muralidharan, K. and V. Sundararaman** (2009), "Teacher Performance Pay: Experimental Evidence from India", Working Paper 15323, National Bureau of Economic Research, retrieved 15 December 2009, from www.nber.org/papers/w15323.pdf.
- National Academy of Education (NAE)** (2009), "Standards, Assessments, and Accountability – Education Policy White Paper", Washington, DC.
- Organisation for Economic Co-operation and Development (OECD)** (2005), *Teachers Matter: Attracting, Developing and Retaining Effective Teachers*, OECD Publishing, Paris.

OECD (2007a), *Education at a Glance 2007: OECD Indicators*, OECD Publishing, Paris.

OECD (2007b), *PISA 2006: Science Competencies for Tomorrow's World*, OECD Publishing, Paris.

OECD (2009a), *Evaluating and Rewarding the Quality of Teachers* (S. Sclafani, ed.). OECD Publishing, Paris, www.sourceoecd.org/education/9789264061989.

OECD (2009b), *Creating Effective Teaching and Learning Environments: First Results from TALIS*, OECD Publishing, Paris.

Parandekar, S., E. Amorim and A. Welsh (2008, March), "Prova Brasil: Building a Framework to Promote Measurable Progress in Learning Outcomes", *En breve*, 121, pp. 14.

Rogers, F.H. and E. Vegas (2009), *No More Cutting Class? Reducing Teacher Absence and Providing Incentives for Performance*, World Bank Policy Research Working Paper 4847, World Bank, Washington, DC.

Salieri, G., L. Santibañez and B. Naranjo (2010), *State-Level Teacher Evaluation and Incentive Practices in Mexico: Diagnostic Study*, study commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.

The State of Queensland: Department of Education and the Arts (2005), *Queensland Curriculum, Assessment and Reporting Framework*, Brisbane: Strategic Policy and Education Futures, DEA, retrieved 23 August 2010 from http://education.qld.gov.au/qcar/pdfs/qcar_white_paper.pdf.

Victorian Curriculum and Assessment Authority (2010), website: www.vcaa.vic.edu.au/.

CHAPTER 4

Using Student Learning Outcomes to Measure Improvement

4.1 Student learning outcomes: Assessment instruments and measures	50
4.2 The ENLACE assessment system in Mexico.....	55
4.3 Challenges and opportunities for further development of the ENLACE assessment system	60
4.4 Summary recommendations for Mexico.....	62

As discussed previously, student learning and growth over time are key criteria against which educational systems, local education authorities, schools and teachers are to be held accountable. An important challenge, therefore, is to properly assess student learning and growth. A single type of assessment cannot fully reflect student learning. All forms of assessments, from standardised tests to portfolios of students' work have issue of validity, reliability and objectivity (Baker, 2010). It is important to develop a system that uses different measures of student achievement and multiple sources, in which assessment data can serve as a quantitative anchor (OECD, 2010).

The first section of this chapter provides an overview of the options for assessing student learning, highlighting the strengths and weaknesses of the different methods, with relevant international examples. Based on the discussion of the formative and summative assessment options, the section suggests that educational systems such as Australia, Alberta (Canada) and Hong Kong-China benefit from having different sources of information regarding student performance to ensure the highest level of completeness and accuracy.¹ Having a battery of valid, reliable and varied measures of student learning and growth, however, also has clear implications in terms of costs and capacities required, particularly at local levels. Not surprisingly, the majority of OECD and partner countries apply student assessments based on tests of student achievement (OECD, 2008).² Assessments of student learning are used in different countries for a range of purposes, including for gauging the performance of the system as a whole, for diagnostic purposes tied to improvement efforts (e.g. Mexico), for accountability (e.g. the United States and United Kingdom), for incentives for teachers and schools (e.g. Chile), and for combinations thereof (OECD, 2009a). As an integral part of an educational system, assessments themselves are reviewed, evaluated and modified in OECD and partner countries to better reflect policy priorities, education reforms in related areas such as curriculum, and the demands of a rapidly changing world (e.g. Australia, Brazil, Norway, the United Kingdom and the United States).

Although there is no single model of assessment that can be gleaned from international practice, some of the technical, logistical and political challenges are common across education systems. Some of these common issues and some of the recommended practices in terms of linking assessments to standards and curriculum, are presented in this chapter. Based on a country's policy priorities and the conditions, constraints and opportunities of a particular education system, the challenge will be to find the right combination of different assessments, their relative weighting, and their uses and consequences (OECD, 2009b).

Developing these different complementary methods of student assessments takes time and resources. Education authorities can establish a gradual process that takes advantage of the immediately available sources for school improvement and accountability initiatives, while also having a longer-term vision of assessment. Because of the cost-effectiveness of standardised student assessments, as well as the relative comparability of results across diverse national contexts, externally administered standardised assessments are used in several OECD and non-OECD countries for both accountability and improvement purposes. The ENLACE assessment in Mexico, begun in 2006, and the *Prova Brasil* that the Brazilian federal government implemented for the first time in 2005 (Box 4.1), offer good examples of dynamic development (Zúñiga Molina and Gaviria, 2010; Parandekar *et al.*, 2008). Other examples across OECD countries show that it is possible for education systems to implement external assessments, while allowing education practitioners to innovate in their practice (OECD, 2009). As a specific example of the opportunities for standardised assessment to allow innovative practices, the chapter provides an overview of the main characteristics of the ENLACE assessment in Mexico, and concludes with specific considerations and recommendations for educational authorities for its further development.

4.1 STUDENT LEARNING OUTCOMES: ASSESSMENT INSTRUMENTS AND MEASURES

Student results, whether actual scores or marks, or the proportion of students attaining specific and pre-determined performance levels, may be based on one or more measures of student learning. These may involve student essays, extended projects, portfolios of student work, multiple choice or short answer tests, among others, and may be used for formative or summative assessment (OECD, 2009; Baker, 2010).

Box 4.1 Prova Brasil for accountability and improvement

Brazil's first census-level student assessment *Prova Brasil* was administered by the Ministry of Education for the first time to test proficiency in mathematics and Portuguese in 2005. The assessment is administered every three years to primary and secondary students in Grades 4 and 8, and represents one of the government's main efforts to establish a results-oriented accountability framework focused on student achievement. A recent study conducted by the Ministry of Education used student achievement data from the assessment and regression analysis to identify municipalities with superior performance, even after considering students' family and socio-economic background. Using qualitative methods, such as classroom observation and interviews, the study further attempted to identify good policies and good practices at the local level that may be contributing to superior performance.

Further information is available (in Portuguese) at <http://provabrasil.inep.gov.br/>.

Source: Moriconi, 2009; Parandekar *et al.*, 2008.

It is common to suggest that assessments should best represent the cognitive demands or thinking skills that are considered most important by the education system. Constructed student responses are often considered preferable because students have to reach into their repertoires, search and then apply their learning (Baker, 2010). Likewise, for measures addressing skills either easily memorised or easily developed outside of school (e.g. via the Internet), education systems may choose more efficient and cost-effective testing processes, saving extended and more expensive assessments for learning that requires difficult understanding, applications and communication of rich or complex content (Baker, 2010).

Technically, however, there is not a bi-univocal relation between the cognitive level of the skill to be assessed and the type of items to be used (Zúniga Molina and Gaviria, 2010). An important consideration is the amount of information provided by a specific item. A single constructed response item can contain a higher amount of information, if carefully stated, than the corresponding multiple choice question. A series of multiple choice items, however, designed to the same cognitive demand, could provide, if carefully designed, the same amount of information. A trade-off exists, therefore, between the information provided by each particular item and the facility to automatically mark the responses given by the test takers.³ Often decisions regarding which instruments and measures to employ are made on the basis of cost and feasibility rather than on optimal assessment for particular standards (Baker, 2010). To compare approaches, Table 4.1 presents a summary of some of the main characteristics, strengths, limitations, cost implications and technical issues of different assessment options.

There are a limited number of high-quality approaches to ensure the comparability of performance assessments, particularly if the interest is in providing some degree of feedback regarding the teaching and learning process. If not properly designed, assessments may not shed sufficient light on which aspects are well or poorly learned, thus providing little or no guidance for system improvement. Performance assessments may be developed using relatively clear domain boundaries for content and cognitive demand, in which case comparability may be easier to establish. Because any set of constraints limits the range of student performance that can be assessed, including tasks that do not have a specific focus could determine the degree to which students can transfer their learning to new situations (Baker, 2010). The design of the PISA assessments, for example, follows this logic. The degree of transfer may be relatively small (but nonetheless difficult), for example where students are asked to perform a mathematics procedure presented in a previously unseen format. Transfer may be more difficult when students are given a problem requiring the application of different strategies rather than a common procedure for solution (Baker, 2010).

Table 4.1

Instruments and sources of evidence to assess student learning

Assessment format	Comments/Uses	Strengths	Weaknesses (sources of bias, validity and reliability issues, cost-considerations, and capacity requirements)
Long open-ended responses, projects, essays	Formative, summative or combined use. Curriculum embedded; may require teacher assistance.	Task validity, rich content, high cognitive demands; transfer and application a possibility.	For accountability: cost and comparability; replacement costs; training of teachers for valid and reliable scores; weak alignment with standards.
Portfolios of student work	Classroom use; weak evidence for accountability, except selection to higher programmes; may include required or chosen elements in the same or different content.	Assessment over time, showing progress or flexibility on multiple topics; choice of topics, style or content; transfer and application possible.	Scoring unreliability if student choice is offered; comparability among students; cost of scoring if external to the classroom; conflict for teachers if used in accountability; requires extensive teacher training; if used for accountability, high cost.
Classroom observation	For use in teacher effectiveness or as opportunity to learn explanation for outcomes. May be conducted by peers in school, other teachers, administrators, or pedagogical or content experts.	Real time sense of teacher and student activity; feedback for teacher; may be conducted by peers; value to explain data on student or value-added modelling reporting; feedback for teacher evaluation.	Requires agreement on learning model and relationship to standards; need multiple visits; trained observers; high and low inference rating scheme; if purpose is feedback, training to observe and give feedback validly and reliably; observation biases teacher and student behaviour; not easily scalable unless random samples of video with significant scoring costs.
School- and class-based tests of any format	Strengthens instruction and alignment; with quality control builds repertoire of assessment events and instructional interventions; may be combined with summative assessment.	Fits the curriculum as taught; immediate feedback to students and intervention possible; builds teacher capacity; provides new examples for outcome examinations.	May be closely or loosely related to standards; may be of poor quality (psychometric characteristics); scoring schemes may not be explicit. Training in assessment design, administration and scoring required. If used for formative assessment, strategies for improvement required; if used as part of the accountability system teacher conflict of interest is possible.
Standardised assessments using multiple choice and short answer (externally provided)	Used commonly for broad summative purposes (diagnostic and accountability-focused).	Inexpensive to score; good for vertical equating and growth modelling; reliable. External character decreases sources of bias stemming from local relations.	Validity is a question if not properly tested; carefully designed content and cognitive demands may be shallow; alignment to content and curriculum may be weak. May encourage teaching to the test and other non-desirable behaviours; limited transfer of knowledge and skills to applied settings.

Source: Baker, 2010.

Marking

Open-ended or constructed student responses are commonly marked by teachers, the students' own or by teams of teachers specially trained to mark examinations (Baker, 2010). Training may occur by having markers examine a range of student responses against a set of pre-validated or expertly scored examples. In some cases, a chief examiner may prepare the paper, and marking is based on deviations from the model. In other cases, training involves exposing markers to the variety of ways students may achieve a score level. In all cases, data are captured about the effectiveness of the training; usually by requiring the teacher to mark a set of papers at a level that is considered adequate to qualify. Some marking training sessions emphasise reliability, focusing on the degree to which individual teachers agree with one another. This approach, when used without validation, may lead the markers to define quality in terms of socially shared expectations, and as one group of markers may differ from another, resulting in varying levels of stringency used for different groups of students. This can undermine the trustworthiness of the results. With adequate quality training, and with common scoring dimensions, the reliability of markers and the validity of their marks can be high (Baker, 2010).

Logistics, costs and technology

Training to mark an examination consisting of extended student work may take between four hours and two days (Baker, 2010). Many training sessions are followed by marking sessions, so costs include the cost of the markers (if teachers, their typical daily work time), travel, housing and food. In an effort to contain costs, common training may be conducted in a set of centrally located sites or remotely by computer, with telephone or chat support either mandated or available. The actual marking may be done at home, with no supervision, but with access to or mandated calls by the marking supervisor. Outsourcing examination marking to non-teachers or teachers in other locales has been used in some educational contexts, but it is not common (Baker, 2010). Individuals can be trained either in person or remotely. When marking is done without supervision (e.g. at the teacher or marker's home), the costs of marking can be significantly reduced, since no payments for travel or other on-site expenses are required, and compensation may be on a piecework basis, depending upon the number of papers marked. In some countries and sub-national jurisdictions, marking papers is a routine expectation and included in any collective bargaining conducted by the teachers' union (e.g. Alberta, Canada), while in others, marking of assessments "need not be done by teachers".⁴

More recently, technology has been used to score student essays with a reasonable degree of success. Approaches involve the use of pre-marked essays and a complex regression model that includes linguistic and lexical aspects of the students' work (Burstein, 2003). Some of these approaches require time-consuming training and the rating of papers in advance of the computer marking, a procedure that must be carried out for each and every change in topic. Other approaches work in a manner similar to grammar and spelling checkers in word processing software. Complex natural language understanding systems are also available, but to date these still require intensive work to adapt them to different topics and different levels of student work (Baker, 2010). Computer scoring of problem-solving and other open-ended responses is being developed (Chung and Baker, 2003; Chung *et al.*, 2001), and is most useful if the problem has a specific set of right answers (Baker, 2010). Computer scoring can also map (through neural networks) the paths taken by different students to achieve success. The latter data are useful for formative assessment (Baker, 2010).

Two of the primary limitations, however, for computer marking have been the availability of equipment and computer literacy of students and teachers. There have been significant improvements in optical scanning of student writing and voice recognition software. Within a few years, this may have important ramifications for the current paper-based assessment practice worldwide (Baker, 2010). The structural constraints of computer access, computer literacy and connectivity, however, remain challenges for emerging and developing economies (ITU, 2009).⁵

Quality of assessments

The technical quality of assessments is a major issue when findings are used to make high-stakes decisions for students, teachers, principals, or other individuals.⁶ A key criterion regarding the technical quality of an examination is *validity*. Validity depends upon the purpose of the test and the evidence that the uses of the test are appropriate. If the purpose of the test is to assess accurately students' acquisition of content and skills, then inspection of the tasks or items and the estimate of depth of sampling are important considerations of content validity. Older notions of validity, including "face" or content validity, concurrent validity and predictive validity, have been subsumed in an overall consideration of validity for tests or examinations (Baker, 2010; Linn [ed.], 1993).

If the purpose of the test is to select candidates for higher levels of schooling, then the test used may well be examined with regard to its ability to predict success in further schooling. As accountability systems and the assessments within them have evolved, assessments appear to have multiple purposes, a fact that makes validity estimates more difficult. For example, if an accountability test is designed to place students accurately

in a classification or on a continuum reflecting their level of mastery of subjects, then the assessment needs to be able to differentiate among students in different classes, asking more difficult questions of those expected to have more highly developed skills. If the test is intended to measure the effectiveness of the educational system, then validity evidence should be available to show that the test is sensitive to high-quality instruction. If the test performance does not change as a function of teaching, but rather by maturation or other non-school influences, it is not appropriate for use in accountability systems. In addition, results from the same examination may be expected to be used by teachers to revise their instructional sequences, either in the same school year or across years. There must be evidence, therefore, that the reported results provide sufficient and relevant information for that function. If tests are expected to monitor students' growth over a number of years, then the idea of vertical scaling (difficulty of tests is equivalent in different years) is essential. Such a requirement can also support value-added models that attempt to identify the contributions of schools in improving student outcomes (treated in more detail in Chapter 5).

Box 4.2 Mixed systems of student assessments

Victoria, Australia: Combining school-based and state assessments

In the state of Victoria, Australia, the Victorian Curriculum and Assessment Authority (VCAA) has established the Victoria Essential Learning Standards to provide a yearly description of what all primary and secondary students are expected to learn and achieve. The VCAA also administers the National Assessment Program – Literacy and Numeracy (NAPLAN) and provides school-based, on-demand assessments for schools. Teachers are involved in developing school-based assessments, along with academic support staff, and all prior year assessments are available to the public. At least 50% of the total score for students consists of classroom-based tasks (e.g. lab experiments, investigation on key topics, and extended reports). At the same time, as part of the NAPLAN, approximately 260 000 students in years 3, 5, 7 and 9 undergo standardised assessments throughout Australia. The system thus combines school-based assessments with standardised state assessments.

Further information is available at www.vcaa.vic.edu.au/.

Alberta, Canada: Developing a holistic framework for assessment

In Canada, the Alberta Education Authority commissioned the *Alberta Student Assessment Study*, a review of theory and practices relative to student assessments that provided recommendations on:

- Curricular learning outcomes, performance standards, and the reporting of student achievement.
- How external assessments and classroom-based assessments of student achievement can be used optimally to inform important decisions on student needs, school management and issues at the provincial levels relating to ensuring learning opportunities for all students.

Results from the study were presented in 2009, with specific guidelines and recommendations on how the education system could effectively combine performance standards, classroom-based assessment and provincial assessment, reporting of student achievement and professional development of teachers. Its recommendations are based on sound evidence and provide useful references for other systems looking for ways to establish complementary approaches to student assessment, within an accountability and improvement framework.

Further information is available at <http://education.alberta.ca/departments/ipr.aspx>.

Sources: Victorian Curriculum and Assessment Authority, 2010; Darling-Hammond and McCloskey, 2008; Government of Alberta (Canada) – Education, 2009.

A review of test results should consider whether the examination is relevant to its purposes, and whether there is evidence to document the examination's ability to deliver on its purposes or claims. Getting validity evidence is difficult, especially during a developmental testing period when the prospect of public reporting or sanctions does not exist. Experiments can be conducted to determine whether assessments are sensitive to instruction, whether they properly categorise students who have done well on similar measures (concurrent validity), or whether relevant performance standards have been set at an appropriate level (as opposed to politically defined levels) (Baker, 2010).

The *reliability* of assessments is also a vital issue. This refers to the consistency with which a test measures performance. Reliability may naturally degrade if change (*i.e.* improvement) is desired. Using proper statistical controls may solve this problem partially. Measures used in accountability must also meet the challenge of *fairness*. Fairness does not mean equal outcomes, but that the characteristics of the examination and marking do not advantage any particular group, other than those most well prepared. For education systems that serve heterogeneous student groups (*e.g.* with varying socio-economic backgrounds, from different ethnic groups or with different languages spoken at home), fairness of assessments is an important issue. Linguistic features of students might confound estimates of learning in other subjects, like mathematics or the sciences. Other issues are that tasks may be relevant to only a subset of students (*e.g.* urban or rural) or show differences in performance independent of competence in other similar tasks in the domain. Weaving these technical requirements into a fabric that supports the technical quality of the measures used in accountability requires forethought and discussion with relevant stakeholders to determine which features may pose challenges for wide acceptance (Baker, 2010).

Because all measures of student learning, including assessments, may present potential shortcomings and sources of error and bias that can affect validity and reliability, education systems may opt to have different sources of information on student performance to ensure the highest level of completeness and accuracy (Baker, 2010). Assessments of student learning are used in different countries for different purposes. The challenge is therefore to find the appropriate balance of standardised assessments, school-based assessments, externally and internally marked and referenced, for different purposes and within the capacity, budget and structural constraints of the education system. The following section reviews an important student assessment in Mexico and discusses some of its main characteristics in light of the previous discussion in order to identify challenges and opportunities for its further development.

4.2 THE ENLACE ASSESSMENT SYSTEM IN MEXICO

In 2006, SEP implemented the first round of the annual National Assessment of the Academic Achievement in Schools (*Evaluación Nacional del Logro Académico en Centros Escolares*, ENLACE). ENLACE was designed to provide information to students, parents, teachers, principals and the general public regarding individual student achievement and grouped results at the school level.⁷ In contrast with the EXCALE exams that are administered to samples of students in different grade levels,⁸ the original purpose of ENLACE was to serve as a benchmark to inform improvements in teaching and learning processes at the school and classroom level for primary and secondary students (Zúñiga Molina and Gaviria, 2010).

Mathematics and Spanish have been tested in every round since 2006, with a third subject varying each year: science was included in 2008, civics and ethics in 2009 and history in 2010; geography will be included in 2011. The exam is currently applied to primary students in years three to six and secondary students in years one to three. The test was applied to the first two years of secondary school for the first time in 2009. Overall, in 2009, the ENLACE assessment was taken by more than half (51%) of all students at the pre-primary, primary and secondary levels in Mexico (INEE, 2010, Indicator ED01; Zúñiga Molina and Gaviria, 2010).

Levels of student achievement reflected in ENLACE results since 2006 are in general agreement with PISA 2006 results, that is they are low on average but there has been improvement. Between 2006 and 2009, for example, the percentage of students classified as “unsatisfactory” or “regular” dropped from 78.7% to 67.2%, while the percentage of students classified as “good” or “excellent” rose from 21.3% to 32.8%.

Results from the 2010 ENLACE application, as well as results from PISA 2009, will give further information on trends in student achievement.

Administered by SEP through the General Directorate for Policy Evaluation (*Dirección General de Evaluación de Políticas*, DGEP), the ENLACE assessment has become a socially accepted measure of student performance (Zúniga Molina and Gaviria, 2010; Salieri, Santibañez and Naranjo, 2010). The test is administered in April each year and results are presented publicly in September of the following academic year. DGEP processes test results using both commercial and proprietary software, and produces materials for users to interpret the results that are presented via school information packets and via the Internet (www.dgep.sep.gob.mx). Through this website, students, families and any interested person can obtain information, using a special identification number on the student's answer sheet. The results can also be seen in aggregated form, by school,⁹ by state or at the national level. Media coverage of the results is widespread and although the practice is discouraged by officials, different versions of "school tables" comparing grouped averages of raw scores of students by school is common practice. The importance that SEP and state education authorities have given the public presentation of results of the ENLACE assessment is supported by international comparisons. As discussed in Chapter 3, results from the PISA 2006 results indicate that the strongest impact on student performance across countries was related to the publication of schools' student achievement data (OECD, 2007).

The information that schools are supposed to receive through state educational authorities includes the proportion of students at each achievement level by grade and content subject. Each school should also receive information on the proportions of students at each achievement level compared with the results of the students and schools of the same type, at state and national levels. The information is organised so that it is useful for identifying possible teaching improvement opportunities and for allowing groups to compare their results against those of other schools with similar socio-economic conditions and infrastructure (Zúniga Molina and Gaviria, 2010).¹⁰ Thus, teachers, school principals, and students and their families can assess progress and the difficulties encountered in learning, including identifying parts of the curriculum that have not been appropriately addressed. Teachers are expected to analyse students' results and identify strengths and weaknesses in the subject areas tested.

A recent review of state-level uses of the ENLACE results showed that they have become a national and local reference of students' learning achievement, with most state education authorities conducting some form of follow-up activities (Salieri, Santibañez and Naranjo, 2010). SEP provides all state educational authorities with printed brochures, reports and CDs to be distributed to schools regarding individual and school-grouped performance. Some states such as Jalisco, Nuevo León and Veracruz have developed their own materials that are distributed to supervisory staff (*Supervisores* or *Jefes de Sector*), and offer some form of support and professional development courses to schools identified as under-performing based on collective ENLACE results and needs assessments (Salieri, Santibañez and Naranjo, 2010). This underscores the importance of and opportunities for state educational authorities regarding improvement efforts and accountability mechanisms (treated in Chapter 7 and other chapters of the report).

Design and technical characteristics

Based on a consideration of the basic design elements, characteristics and test results since 2006, the ENLACE assessment instrument presents robust levels of internal consistency, validity and reliability as a measure of student learning (Zúniga Molina and Gaviria, 2010).¹¹ It is important to note, however, that further development of ENLACE may require more in-depth studies, particularly in light of current curricular reforms taking place in Mexico, as well as the uses that the ENLACE assessment may be assigned in the near future. Following is a summary of the elements of the assessment that were reviewed, with preliminary conclusions regarding validity and reliability.

Psychometric model

The tables of specifications for ENLACE were initially established by curriculum experts within SEP who determined the most relevant content to be reflected in each test. The tables were further developed by DGEP staff, with assistance from INEE experts and approved by the Under Secretariat for Basic Education of SEP. For the construction of the assessment, three difficulty levels were established (low, medium and high) to ensure that the tables described the content to be assessed for every grade, difficulty level and subject. The specifications used for construction of test items are publicly available and are open to revisions based on suggestions from teachers, principals and state educational authorities. The tables for mathematics and Spanish, however, have remained unchanged since 2006. The tables allowed for qualitative interpretations of performance differences among students, in order to allow for feedback to improve processes based on results. The scale used to report ENLACE results has a mean of 500 points, with a standard deviation of 100 points, corresponding to 2006 averages as the baseline. ENLACE results have a normal distribution, according to which 99% of students score between 200 and 800 points in each grade and subject. The scale is based on Item Response Theory (IRT) and assigns students to different levels of achievement, with comparability between successive years. Test items are analysed before scoring using a classic model (difficulty index and bi-serial point correlation, as an approximation for item discrimination). The items are then calibrated and the students are classified according to the three-parameter IRT model. A score value is assigned to each student, considering not only the number of correct answers, but also which items were correctly answered. Because a scale is set for each grade and curriculum content-subject, comparisons of scores between different education years or grades is not possible.

Unlike the Rasch model, in the three-parameter model the constructs to be measured are defined before adopting the measurement model. For the development of ENLACE, the existing curriculum guided the construction and selection of items for the test and the parameters were adjusted to the characteristics of the items. In the Rasch model, measurement invariance cannot be obtained simply by using the model on a given set of items; the psychometric model becomes the principle for defining the construct, rather than having the construct guide the development of the test. The three-parameter model used for ENLACE allows the test to reflect the structure of the curriculum without compromising the measurement model. It should be noted that cut scores for achievement levels are not the same in all grades, as they were defined separately for each grade and subject (*i.e.* there is no common scale for all grades).

Dimensionality

The dimensionality of the ENLACE tests is one of the fundamental characteristics that must be analysed in order to determine the structural stability of the results of the assessment over time and hence to allow estimates of actual improvements in learning achieved by students. A recent study of the ENLACE assessment conducted by Lizasoain Hernández and Joaristi Olariaga (2009) concluded that:¹²

- With few exceptions, the ENLACE tests used in the 2008/09 academic year can be considered as essentially one-dimensional or as having weak multidimensionality.
- The results from samples taken from the population and from a control sample do not suggest different dimensional structures.
- The tests can be considered, in general, as having low complexity or a simple structure.
- With regard to possible differences in the dimensionality of the tests, there is some degree of multidimensionality in particular grades. This is probably due to the greater complexity of the curriculum content in these grades (*i.e.* in the third year of secondary school).¹³

These findings suggest that the characteristics of ENLACE, combined with the robustness of the test construction models, ensure the correct scaling of students' responses.

Reliability

Reliability of the ENLACE tests is high, based on the internal consistency coefficient (Cronbach's alpha coefficient) calculated. The calculated values for the ENLACE assessment are within the range that is generally accepted as an indicator of highly reliable scores. For the 2006-08 ENLACE applications, for example, the alpha coefficient values vary between 0.75 and 0.92, with averages for Spanish, mathematics and science well above 0.80 (Table 4.2). Based on a comparison of values calculated for the PISA 2000 assessment (OECD, 2002, p. 152, Table 4.1), the reliability of ENLACE is similar and in some cases exceeds that of the PISA 2000 results (unconditioned unidimensional scaling).

Table 4.2
Reliability of ENLACE

Subject	Year	Reliability of ENLACE		
		2006	2007	2008
Mathematics	3rd Primary	0.896	0.896	0.912
	4th Primary	0.915	0.898	0.922
	5th Primary	0.896	0.874	0.907
	6th Primary	0.872	0.874	0.910
Spanish	3rd Secondary	0.838	0.789	0.865
	3rd Primary	0.876	0.844	0.879
	4th Primary	0.903	0.900	0.906
	5th Primary	0.809	0.837	0.804
	6th Primary	0.880	0.891	0.910
	3rd Secondary	0.835	0.813	0.752
Science	3rd Primary			0.854
	4th Primary			0.853
	5th Primary			0.818
	6th Primary			0.880
	3rd Secondary			0.804

Source: Zúniga Molina and Gaviria, 2010.

Validity

A study was conducted to assess the concurrent validity of ENLACE in relation to other tests such as the PISA assessment. Since sufficient PISA elements were not publicly available, researchers constructed a special test (SEP-ISA) in agreement with the Australian Council for Educational Research, using test items from the item bank used for the construction of PISA, as well as items previously used for PISA tests and later released. For this study, researchers selected a stratified random sample of 11 717 students in the second and third years of secondary school throughout Mexico. These students took both the ENLACE tests for mathematics and Spanish, and the SEP-ISA test (mathematics and reading comprehension).¹⁴ The correlation of the scales from the different tests, corrected for attenuation, were approximately 0.829 for Spanish/reading comprehension and 0.810 for mathematics (Zúniga Molina and Gaviria, 2010). For comparative purposes, the estimated correlations between scales and subscales of mathematics, reading and science from the PISA 2003 assessment are presented in Table 4.3. The correlations obtained between ENLACE and SEP-ISA are of the same magnitude as those estimated for the subscale problem solving with the other dimensions of mathematics in PISA, ranging from 0.79 to 0.83. Comparatively, these results suggest that the levels of validity of ENLACE are quite high.

Table 4.3

Correlation between subscales of problem solving, reading and science, PISA 2003

	Space and shape	Change and relationships	Uncertainty	Quantity
Space and shape		0.89	0.88	0.89
Change and relationships			0.92	0.92
Uncertainty				0.9
Problem solving	0.79	0.83	0.81	0.82
Reading	0.67	0.73	0.73	0.73
Science	0.73	0.77	0.77	0.76

Source: OECD, 2005a, p. 190, Table 13.4.

Quality of equating process

As described earlier, the specific learning content assessed by ENLACE is determined by the tables of specifications for which four criteria were used to identify, prioritise and focus the curriculum content for the tests: *relevance*, *plausibility*, *continuity* and *comprehensiveness*.¹⁵ New versions of the ENLACE tests must be constructed each year given that the test booklets remain in the public domain after application. To ensure that tests are equivalent between consecutive years for the same subject content and grade level, test developers use a variant of the common population design.

The adequacy of the equating process depends on the stability of item parameter estimates and the scores of students: estimates of the parameters should be consistent, regardless of the subset of items used in the estimate. To evaluate the ENLACE assessment in this respect, Zúniga Molina and Gaviria (2010) estimated the item parameters when items were calibrated separately and when these same items were calibrated together with the items from the “pretest form”.¹⁶ Two variables corresponding to the difficulty for the item parameter estimates were obtained for each year and subject area. These variables were found to have a very high correlation in all grade levels and subjects, never dropping below 0.993 (Zúniga Molina and Gaviria, 2010, Appendix I). The authors also compared scores obtained by students in one year level (grade), again using two variables for each grade and subject. These values were also found to have a high correlation, never dropping below 0.985 (Zúniga Molina and Gaviria, 2010, Appendix II).

Equating errors were also calculated for ENLACE 2008 and 2009, and then compared with equating error data for the reading comprehension component of PISA 2003. The values for the equating errors were found to be very similar between the 2008 and 2009 ENLACE results, and in some cases lower than equating error values for PISA 2003 reading comprehension. This suggests that the *horizontal* equating process is reliable for ENLACE, allowing for comparisons between student results for each grade and subject over consecutive years (*i.e.* different students, same grade) (Zúniga Molina and Gaviria, 2010, p. 39).

Vertical equating

Although the ENLACE assessment is not designed for vertical equating, SEP and invited experts have conducted feasibility studies to determine the options for further development of ENLACE in the near future to include a vertical scale. This would allow, for example, comparisons of student results between different grade levels (*i.e.* potentially same students, different years). The preliminary studies focused on the results of 104 487 students in the mathematics component of ENLACE in the sixth year of primary school (4 533 classrooms) in Mexico City (presented in Appendix III of Zúniga Molina and Gaviria, 2010). Results from these trials show that it is possible for ENLACE to include a common scale between fifth and sixth grade mathematics. Furthermore, the drop in results for approximately 36% of students between fifth and sixth grades is commensurate with the results of 15-year-old students assessed by PISA. Vertical equating would also allow testing of the cut-off scores defined

for each grade level of ENLACE. The cut-off points for all of the years of primary school included in ENLACE (third to sixth), for example, would need to be revised based on a revision of the criteria used to establish them, as well as the validity of the vertical scale. Finally, the degree of the vertical equating errors will also need to be studied before incorporating a common vertical scale in ENLACE. The further development of the ENLACE assessment should incorporate these considerations, as well as others that are outlined in Section 4.3.

Copy factor

As ENLACE is a census assessment (*i.e.* all students in the relevant grade are assessed), supervision and control of test conditions and test application are challenges, particularly given the large diversity of school contexts between and within states. To identify the magnitude of probable answer copying in the applications, two different methods are used. Although the average percentages of probable cheating reached a high of 7.0% in 2008 compared to 4.5% in 2006, the trend decreased in 2009 with an average of 6.5% (Table 4.4). Initial results from the 2010 application of ENLACE confirm that answer copying has not continued to increase. Furthermore, a consideration of the general effects of copying conducted by Zúniga Molina and Gaviria (2010) shows that even for 2008, the estimates of validity and reliability for ENLACE remain largely unaffected.

Table 4.4

Percentages of probable test cheating cases detected for ENLACE 2006 to 2009

Year	2006	2007	2008	2009
3rd Primary	6.28	5.50	10.16	6.97
4th Primary	4.90	5.60	7.53	6.65
5th Primary	4.14	3.68	4.78	4.45
6th Primary	4.10	4.86	5.48	4.75
1st Secondary				1.54
2nd Secondary				3.79
3rd Secondary	3.24	3.14	7.08	6.54

Source: OECD, 2005a, p. 190, Table 13.4.

Further monitoring and analysis of the answer copy factor should continue, however, and measures to address this should be considered for every application. Additional resources and supervisory mechanisms should also be included in the planning of the ENLACE assessment.

4.3 CHALLENGES AND OPPORTUNITIES FOR FURTHER DEVELOPMENT OF THE ENLACE ASSESSMENT SYSTEM

In Mexico as in other better-performing educational systems, consensus is emerging on the benefits of clearly defining the progress students are expected to achieve in the acquisition of skills and competencies. In recent years, therefore, SEP has undertaken curricular reforms focused on the development of student skills, with a major emphasis on achieving specific learning outcomes. These reforms have advanced significantly in the pre-school and secondary levels, and are in an experimental phase for primary school. To the extent that these reforms focus on developing competencies and skills for life, ENLACE will need to reflect these changes. With a standards-based framework in Mexico, curriculum-referenced testing will need to evolve to reflect standards of competencies and skills that may be established. Clearly defined content and performance standards for students, developed as part of the curricular reforms in Mexico, could serve as the anchor and reference for

teacher planning, educational materials, teaching practices, capacity-building and professional development, and ultimately assessments. Perhaps most importantly, clear standards for student learning and growth would also provide a coherent view and create shared *expectations* that all educational levels and actors can share and work towards, including state educational authorities, parents and school-level committees and councils. In this context, the following points should be considered:

- The need to preserve the levels of reliability, validity and structural simplicity and stability already achieved by ENLACE while taking into account the curricular transformations being considered and that may be implemented.
- The need to establish a clear development programme for ENLACE, defining policy objectives, targets and timeframes. The further development of ENLACE should include relevant studies to determine the vertical comparability for students and groups, in order to measure progress towards defined learning expectations. Measures of student progress should be net of socio-economic and other relevant factors, to identify the contributions of schools, school zones, regions and states towards student outcomes.
- The need to address administrative, technical and logistical considerations, in order to allow for more reliable measures of student growth over time towards specific learning objectives (e.g. as defined in content standards of student learning). Addressing these elements would not only make the ENLACE assessment more robust but would also contribute to strengthening the evaluation framework in Mexico, for accountability and for improvement efforts (Chapter 3). Furthermore, addressing the following items would permit the development of value-added methods that are presented in more detail in Chapter 5 of this report and in its sister publication.

Administrative conditions

- Completeness and consistency in the identification of students, teachers, schools and principals where appropriate. There must be specific mechanisms to detect and incorporate individuals who might not be in the databases and to correct inconsistencies. Student and teacher mobility within and across school zones and states should be identifiable with proper tracking.
- Unified individual dossiers for students to accompany each student throughout his or her entire school life. Information on the results of assessments, including ENLACE, must be included in order to determine progress in learning. This dossier could also be used for at-risk students and for efforts to reduce drop-out rates.
- Uniform and unique references for cities, towns, municipalities and schools.
- Capacity to link and match the achievement of students on ENLACE with the teachers who have taught them.
- Capacity to match each item in the database of students with their counterparts in the databases of teachers, principals and schools, in order to determine the contribution of different teachers, or the entire school, to the learning gains of each student.

Logistical conditions

- Substantial improvement in the conditions controlling the application of the test. This includes mechanisms to limit the potential for undesired behaviours from teachers and principals as well as addressing the issue of answer copying. The need for security, supervision, and adequate and standardised conditions for the application of ENLACE will only increase and adequate resources should be considered by SEP and state educational authorities.

Technical conditions

- Continued technical robustness of ENLACE assessments. The demonstrated validity, reliability and internal consistency of ENLACE suggest that only substantial changes in the curriculum would warrant a corresponding transformation of ENLACE. In light of the curricular reforms being considered and implemented in Mexico, however, these may offer an opportunity to plan the further development of ENLACE to ensure alignment, coherence, cognitive demand and breadth that are commensurate with policy objectives.
- Implementation of a vertical alignment design for all grades and content subjects that allows the calculation of educational progress for each student. SEP and invited researchers will need to calculate comparability errors and verify the extent to which the errors on the same cohort, in successive grade levels, are additive and whether the magnitude of these cumulative errors prevents the scale going beyond two or three year levels, for example. The cut-off points should be redefined so that they are consistent across consecutive grades.

4.4 SUMMARY RECOMMENDATIONS FOR MEXICO

Based on the considerations presented earlier, the following are the main summary recommendations for Mexico regarding the importance of assessing student learning outcomes, the opportunities afforded by the ENLACE assessment, and the challenges and opportunities for its further development:

- ***Student learning and growth as the basis of accountability and standards requires multiple, cross-referenced, valid and reliable measures.*** Because all of the current measures and instruments of student learning and growth (standardised tests, teacher assessments, portfolios of student work and observation, among others) present potential sources of error and bias, a complementary approach that uses valid evidence from multiple sources should be gradually developed to assess current instruments in Mexico, estimate costs, and determine the capacity-building and instrument development that are required. With clear content and performance standards of what students are expected to know and know how to do, for example, measures that reflect the learning and growth expected from students can be further developed.
- ***The use of student performance data should be accompanied, when possible, with complementary and reliable measures of student learning, as these are developed, tested and validated.*** The relative importance of student data and school-based or teacher assessments can be redefined as needed by the policy objectives and consequences resulting from the assessments. Australia, Alberta (Canada) and Hong Kong-China are examples of better-performing systems that attempt to combine standardised assessments with school-based assessments (e.g. locally graded but externally moderated), student projects, and extended papers.
- ***Student performance data, such as those from the annual ENLACE assessment in Mexico, can play an important role in accountability and school improvement efforts.*** Current efforts by SEP and state educational authorities regarding the presentation and use of ENLACE demonstrate the high degree of social acceptance and potential of ENLACE. Student performance data aggregated at the group, school, zone or state levels can be employed in static, improvement, or growth models, depending on the specific purpose of the policy levers and programmes in Mexico.
- ***A specific development programme should be established for the ENLACE assessment, considering issues of cognitive demand, curricular alignment and coherence.*** The best-available evidence on student learning progression and standards should be considered. The development of ENLACE should set clear stages and goals that address technical (e.g. vertical equating), administrative (e.g. unique student, teacher and school identifiers and linkages) and logistical (e.g. improved test supervision) considerations.¹⁷ With expanded use of the ENLACE assessment in the future, enhanced supervision and security of test administration, for example,

should be addressed. The programme should also have a long-term vision that takes internationally benchmarked content and performance standards into account. As content and performance standards are established in Mexico, student performance data can be used, in conjunction with analytical models (e.g. growth) for specific policy objectives and programmes. Throughout the process, consideration should be given to the alignment and coherence between standards, assessment and professional development for teachers. A clear vision of the evaluation framework in Mexico should allow for the distinct but complementary purposes of different assessments (*i.e.* ENLACE, EXCALE, or possible school-based assessments), and how they should continue to develop in the future within a common national framework.

- ***With student performance data and appropriate growth models, low performers, high performers and cases needing follow-up observation can be identified.*** As the assessment and evaluation process becomes more established, consequences such as incentives, further observation, and assistance to schools and teachers can be linked to the results. This implies both a gradual development of the process and the possibility of having multi-stage consequences and responses to the results.

NOTES

1. For reference, the average performance of students in Alberta in PISA 2006 was significantly above the Canadian average, which was already among the top performers, along with Hong Kong-China (OECD, 2007; Bussiere, Knighton and Pennock, 2007). Australia was among the top-10 performing economies, out of 57 (OECD, 2007).

2. The most commonly tested subjects are mathematics and the national language, with science included in only seven out of 29 countries for which information was available (OECD, 2008).

3. With constructed response items, once the question is stated, the load of work falls on the marker's shoulders, and the correspondence between the score assigned to a particular student and her/his cognitive status depends on the marker's dexterity in correctly detecting the telling signs of that cognitive status. With multiple-choice items, a considerable amount of work for the assessment is conducted previously, when dividing the cognitive task into the relevant steps where the different cognitive levels must be identified through the different combinations of the alternatives.

4. For the government of Alberta, Canada, teachers' professional responsibilities should include the marking of provincial achievement tests (<http://education.alberta.ca/department/ipr/commission/report/reality/governance/bargmodel.aspx>), while marking of student assessments "need not be done by teachers" in a document presented by the largest union in the United Kingdom, the NASUWT (cited in Stevenson, 2004, p. 233).
5. For example, based on the ICT Development Index that combines access, use and skills data from the International Telecommunication Union, Korea ranked 2nd overall, Finland 9th while Mexico placed 75th, below Chile (48th), Turkey (59th) and Brazil (60th) (ITU, 2009).
6. A general resource regarding testing is provided by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 1999).
7. This section is largely based on expert contributions commissioned by the OECD as part of the Co-operation Agreement with the government of Mexico. The two working papers are *Challenges and Opportunities for the Further Development of the ENLACE Assessment for Evaluation and Teacher Incentives in Mexico* (Zúniga Molina and Gaviria, 2010), and *State-Level Teacher Evaluation and Incentive Practices in Mexico: Diagnostic Study* (Salieri et al., 2010).
8. Administered by the National Institute for the Evaluation of Education (*Instituto Nacional para la Evaluación de la Educación*, INEE), these are the Educational Quality and Achievement Exams (*Exámenes de la Calidad y el Logro Educativos*, EXCALE). EXCALE exams are sample-based assessments administered in four-year cycles to students of certain key grade levels at the pre-primary, primary and secondary levels. The assessment in 2011 will be for third-year pre-primary students in Spanish and mathematics, followed by an exam on Spanish, mathematics, natural sciences and social sciences for third-year secondary students in 2012.
9. Schools are classified according to the different basic education programmes in the country. The classifications are used to compare results of schools that have similar characteristics in terms of the student population and the resources available for students and teachers.
10. The application of ENLACE to the control sample used for equating purposes is accompanied by a context questionnaire. Information related to the characteristics of the school is taken from the stratification variables used for the sampling process. Although a few studies relating these contextual variables to ENLACE performance have been undertaken, no further operational use of this information has been made.
11. Considerations of consequential validity are not included in this assessment given the multiple uses and breadth of purpose to which ENLACE is currently subjected.
12. The original study is in Spanish and is included as an annex to the Zúniga Molina and Gaviria (2010) paper prepared for the OECD on which this section is based.
13. Until 2009, the third year ENLACE tests were designed to reflect the cumulative content of the secondary level as a whole (*i.e.* tests included content for the first and second years as well). This may explain the findings of Lizasoain Hernández and Joaristi Olariaga (2009).
14. The correlation values between latent variables of ENLACE and SEP-ISA are included in Zúniga Molina.
15. *Relevance* refers to the relative importance attributed by experts to each topic, and depends on the depth of treatment of the different subjects in textbooks; *plausibility* refers to the feasibility of developing multiple-choice items in relation to the content subjects to be included in each particular learning test; *continuity* refers to the extent to which specific content is part of a teaching sequence that extends beyond a particular year-grade; and *comprehensiveness* refers to the level of inclusion of other content associated with lower degrees of complexity (Zúniga Molina and Gaviria, 2010).
16. As ENLACE is applied annually, every year a parallel test (referred to as the "pre-test") is developed and applied to a control sample of students who also take the normal test given to students that year (this is referred to as the "operational form" of ENLACE). The parallel test items are calibrated in conjunction with the operational form of the test and are then used to form the new test for the following year.
17. The specific technical, administrative and logistical recommendations on further development of the ENLACE assessment are presented in Chapter 5.

References

- American Council on Education** (Linn, R.L., ed., 1993), *Educational Measurement*, Oryx Press, AZ.
- American Educational Research Association, American Psychological Association and National Council on Measurement in Education** (1999), *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, DC.
- Baker, E.** (2003), *Multiple Measures: Toward Tiered Systems*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.
- Baker, E.** (2004), *Aligning Curriculum, Standards, and Assessments: Fulfilling the Promise of School Reform*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.
- Baker, E.** (2010), "Assessment and Accountability", expert paper commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.
- Burstein, J.C.** (2003), "The e-Rater Scoring Engine: Automated Essay Scoring with Natural Language Processing", in M.D. Shermis and J. Burstein (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Erlbaum, Mahwah, NJ, pp. 113-122.
- Chung, G.K.W.K. et al.** (2001), "Knowledge Mapper Authoring System Prototype" (final deliverable to OERI), University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.
- Chung, G.K.W.K. and E.L. Baker** (2003), "Issues in the Reliability and Validity of Automated Scoring of Constructed Responses", in M.D. Shermis and J. Burstein (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum, Mahwah, NJ, pp. 23-40.
- Lizasoain Hernández, L. and L. Joaristi Olariaga** (2009), "Estudio de la dimensionalidad de las pruebas ENLACE (2008) mediante técnicas factoriales clásicas y métodos no paramétricos basados en TRI – Informe Preliminar", included as Technical Annex V to the Zúñiga Molina and Gaviria (2010) expert paper commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.
- Government of Alberta (Canada) – Education** (2009), "The Alberta Student Assessment Study: Final Report", Edmonton, Alberta.
- Instituto Nacional para la Evaluación de la Educación (INEE)** (2010), *Panorama Educativo de México: Indicadores del Sistema Educativo Nacional 2009 Educación Básica*, INEE, Mexico City.
- International Telecommunication Union (ITU)** (2009), "Measuring the Information Society: The ICT Development Index", ITU, Geneva.
- Moriconi, G.M.** (2009), "The Development Index of Basic Education and Teacher Evaluation in Brazil", Presentation given at the OECD/SEP International Workshop "Towards a Teacher Evaluation Framework in Mexico: International Practices, Criteria, and Mechanisms", 1-2 December 2009, Mexico City.
- Organisation for Economic Co-operation and Development (OECD)** (2002), *PISA 2000 Technical Report*, OECD Publishing, Paris.
- OECD** (2005a), *PISA 2003 Technical Report*, OECD Publishing, Paris.
- OECD** (2005b), *Formative Assessment: Improving Learning in Secondary Classrooms*, OECD Publishing, Paris.
- OECD** (2007), *PISA 2006: Science Competencies for Tomorrow's World*, OECD Publishing, Paris.

OECD (2008), *Education at a Glance 2008: OECD Indicators*, OECD Publishing, Paris.

OECD (2009), *Assessment and Innovation in Education*, OECD Working Paper No. 24, OECD Publishing, Paris.

OECD (2010), *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas*, OECD Publishing, Paris.

Parandekar, S.D., E. Amorim and A. Welsh (2008), “Prova Brasil – Building a Framework to Promote Learning Outcomes”, Note No. 121, The World Bank, Washington, DC.

Salieri, G., L. Santibañez and B. Naranjo (2010), *State-Level Teacher Evaluation and Incentive Practices in Mexico: Diagnostic Study*, study commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.

Stevenson, H. (2007), “Restructuring Teachers’ Work and Trade Union Responses in England: Bargaining for Change?”, *American Educational Research Journal*, Vol. 44(2), pp. 224-251.

Zúniga Molina, L. and J.L. 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.

CHAPTER 5

Assessing the Value-Added of Schools: Enhancing Fairness and Equity

5.1 Value-added models with the school as the unit of accountability	68
5.2 The importance of quality data and information	73
5.3 Consequences linked with fair and credible assessment of schools and teachers	73
5.4 Considerations for Mexico	75

The previous chapter presented a summary of the main issues surrounding student assessment options, and considered the challenges and opportunities regarding the ENLACE assessment in Mexico. For education systems that conduct periodic standardised student assessments, however, performance data provide an opportunity not only to present a cross-sectional “snapshot” of where students are performing based on raw test scores, but also to explore student growth over time. Evidence shows that students’ backgrounds can largely influence performance (McCall, Kingsbury and Olsen 2004; OECD, 2008). Assessments that fail to account for this run the risk of having the top and low performers merely reflect the socio-economic conditions of students and families, as well as linguistic or ethnic characteristics.

One of the methods a limited number of OECD countries use to make more accurate and fairer assessments of schools’ contributions to student learning outcomes and growth is value-added modelling (VAM). Not surprisingly, systems that have many years’ experience in standardised student assessments, such as the United Kingdom and the United States, have implemented VAM. For countries that have only recently introduced large-scale student testing, such as Mexico and Brazil, VAM could provide an option to define policy objectives aimed at improving educational results. As discussed in Chapter 4, the Mexican Ministry of Education (SEP) and state educational authorities’ investments in the ENLACE assessment have resulted in its wide social acceptance (Zúniga Molina and Gaviria, 2010; Salieri *et al.*, 2010). Given the diversity of educational contexts in Mexico, value-added methods could be used by federal and state education authorities to strengthen school accountability and improvement efforts. The development of value-added methods, however, requires careful consideration, design and planning to effectively address the challenges involved.

Drawing on OECD work on value-added methods and with examples from several countries (e.g. Norway, Poland, Slovenia, the United Kingdom and the United States), this chapter presents an overview of value-added methods, and considers policy implications, design elements and implementation issues that should be taken into consideration. The chapter also highlights the importance of data quality and information in linking educational interventions and consequences with accurate assessments of schools and teachers. It concludes with specific considerations for Mexico regarding value-added methods to increase accountability and improve school performance.

5.1 VALUE-ADDED MODELS WITH THE SCHOOL AS THE UNIT OF ACCOUNTABILITY

Education systems that serve students from extremely diverse backgrounds face an even greater challenge in making accurate and fair assessments of school performance. Education systems that operate in contexts of high diversity and inequality (*i.e.* income distribution, linguistic or ethnic discrimination, for example), must therefore implement assessments and evaluation procedures that are meaningfully comparable at the national level, but that also take disparities into account. Results from the 2006 Science PISA, for example, show that socio-economic background has an above-average importance to student achievement in Mexico (PISA 2006 Database). In these contexts, value-added methods offer an alternative way to strengthen accountability and improve the performance of schools.¹

Because they include students’ prior performance measures, value-added methods help to address some of the common challenges and problems of other methods:

- Dependence on measures such as average scores on standardised exams or percentage of students progressing to higher educational levels that may not take into account other factors that influence achievement: the innate ability of students, students’ socio-economic background, influence of peers and individuals within and outside the school, external events or shocks that impact learning, and general randomness of student assessments (OECD, 2008).

- Accounting for unobserved factors contributing to the initial measure of student performance, such as student ability or intellectual coefficient, that pose challenges for contextualised attainment modelling (OECD, 2008; Raudenbush, 2004).
- Undesired incentives for school staff to exclude students who are considered under-performers or to retain only higher-performing students (OECD, 2008; Wilson, 2004).

Box 5.1 Jurisdictions in Mexico

In Mexico, the school zone is an administrative designation of a group of schools for the purposes of supervision and administrative monitoring. Similarly, municipalities are one of the three basic jurisdictional units of government and can contribute significantly to infrastructure and material conditions of schools and with social recognition programmes, for example, for schools and teachers.

Source: Salieri *et al.*, 2010.

Value-added modelling addresses these issues because it includes student performance measures as they evolve over time, towards prescribed objectives. Although value-added measures can be calculated for individual students, subject areas, grade levels, schools, and other jurisdictional entities (e.g. school zones, municipalities or states), the general focus here is on the school as the basic unit of accountability.

Based on expert consultations and a review of value-added modelling and related methods in several countries (OECD, 2008, 2010), the following are the key definitions used for this and related OECD reports:

Value-added of schools	The contribution of a school to students' progress towards prescribed education objectives (e.g. cognitive achievement, literacy and numeracy), net of other factors that contribute to the progress of students.
Value-added modelling	Statistical analysis based on models that estimate schools' contribution to student progress in prescribed education objectives (e.g. cognitive achievement, literacy and numeracy), based on measures from at least two points in time.

Despite the challenges involved, an education system can take gradual steps towards implementing a value-added method. These steps may consider statistical and analytical methods that, while attempting to address some of the main factors taken into account in true value-added methods (e.g. students' socio-economic background), should not be considered full-fledged value-added modelling. Education systems can use the following procedures and models as transitional phases in working towards establishing an evaluation framework that uses value-added modelling:

- stratification* of similar schools (based on socio-economic and other relevant information) for within-group comparisons of school average raw scores of students;
- school effects model* based on residuals using multivariate cross-sectional analysis, with school average test scores regressed on aggregated and relevant demographic characteristics of students;
- contextualised attainment models* that estimate the magnitude of contributing factors to student performance or attainment using a specific measure (*i.e.* at a particular point in time). Typically, the model will regress a vector of students' socio-economic backgrounds or other contextual characteristics and a variable of a school-identifier against an achievement measure.

- d) *value-added modelling* using at least two measures of student performance taken at two points in time; and
- e) *contextualised value-added modelling* using quality relevant contextual characteristics of students such as gender, ethnic group and level of education of parents (OECD, 2007). The design and use of models that include student data should address the potential for undesired incentives for school or administrative staff to withhold information.

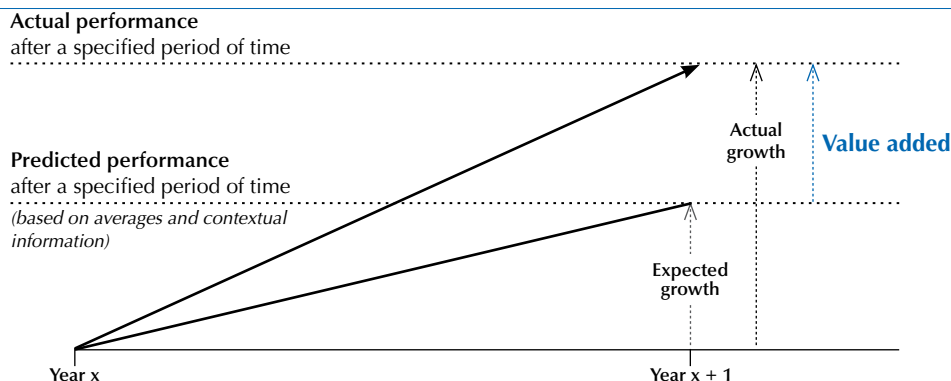
Systems do not necessarily have to progress in this order or through each type of model. Certain methods could be used for public use of analysis, while more complex models could be used internally by education authorities for exploratory purposes or specific objectives, such as programme monitoring and evaluation. The same value-added models can identify not only under-performing schools that may need further inspection and evaluation, or assistance, but also top performers. For education systems that provide services across highly varying contexts and to heterogeneous student populations, value-added modelling can offer “a quantitative assessment of the magnitude of the disadvantage associated with particular characteristics (e.g. ethnicity, income, level of familial education, home language or immigrant status) in relation to student progress, not just in relation to student attainment at a particular point in time” (OECD, 2008, p. 132).

Although it is not yet a common practice in OECD or non-OECD countries, several education systems are already using value-added methods for accountability and school improvement. Two often cited examples are the *Tennessee Value-Added Assessment System (TVAAS)*, in the United States, and the *Achievement and Attainment Tables* in the United Kingdom. Other states and cities in the United States have implemented value-added modelling, as well as other countries such as Poland, Norway and Slovenia. In addition, there are at least eight other countries with student assessments that could be used to explore value-added modelling, including Mexico, as discussed in Chapter 4.² Of these eight countries, only Mexico and the Flemish Community of Belgium produce annual student assessment data (OECD, 2008).

Value-added scores are inherently relative to other schools’ performance. Specifically, the score for an individual school is an estimate of the difference between the school’s contribution to the learning of its students and the average contribution of all the other schools from which data were used in the model. The use of data from another grouping of schools, for example, would yield different value-added scores. Figure 5.1 provides a schematic representation of a simple value-added model. In Figure 5.1, contextual information of students would be used to calibrate the predicted results from the fitted model (*i.e.* average school contribution for the group of schools).

Figure 5.1

Schematic representation of a simple value-added model



Source: Adapted from Martínez-Arias, Gaviria and Castro, 2009, and Goldschmidt *et al.*, 2005.

Several value-added models have been developed and used for different policy objectives within different education systems, and to estimate either annual or cumulative school effects. Most, however, can be classified into five general categories: *i*) linear regression models; *ii*) hierarchical variance component or random effect models; *iii*) fixed-effect models; *iv*) multivariate random effect response models; and *v*) latent growth curve models. A brief description of one common form of these models is provided in Box 5.2. In the initial stages of developing value-added modelling, education authorities and corresponding bodies should explore multiple models and design characteristics to identify the most appropriate model design, which should reflect clear policy objectives, the samples and contextual information used, and the characteristics of the student assessments.

Box 5.2 Example of a linear regression value-added model

$$Y_{ij(2)} = a_0 + a_1 Y_{ij(1)} + b_1 X_{1ij} + \dots + b_p X_{pij} + \varepsilon_{ij}$$

i indexes students within schools j ;

$Y_{ij(2)}$ = final test score;

$Y_{ij(1)}$ = prior test score;

$\{X\}$ denotes a set of student and family characteristics;

$a_0, a_1, b_1, \dots, b_p$ denote a set of regression coefficients; and

ε_{ij} denotes independent and normally distributed deviations with a common variance for all students.

In this model, if students in school j achieve higher final test scores on average, the corresponding residuals will tend to be positive, yielding a positive estimated value-added for the school. For consistent estimates using this model, it is necessary that the included covariates are *uncorrelated* with the error term, which may include a school effect in addition to idiosyncratic errors. Furthermore, it does not take into account the structure of the error term that does happen in other models.

All empirically-based indicators of school performance are subject to variability and bias. As value-added methods use different empirical sources than classroom observations or inspection visits, they can provide a valuable quantitative anchoring of the performance assessment of a school. In designing robust value-added methods, the following statistical, methodological and implementation issues should be addressed:

• Design issues

- Quality of student assessment and test data, including their alignment to curricular goals (including relevant reforms), and validity and reliability of assessment instruments.
- Integrity and coverage of raw test data and contextual data of students, which are to be used for contextualised models.
- Validity and reliability of the assessment that produces the student data that will be used.³
- Clear identification of the dependent variable to be used in value-added modelling.
- School size: in general, results for schools with less than 30 students are more prone to instability, so results should be confirmed using exploratory modelling with country-specific data and information. Additionally, schools with less than 30 students could be grouped to conduct analysis of the aggregates.

- Technical complexity of models, given the availability of data and for communication and understanding of process by school staff and the general public.
- Transparency of objectives, modelling, and communication of results to obtain general acceptance. This requires that decisions regarding results are to be used internally by education authorities, be available to local education authorities and schools, open to the general public, or a combination of the three.
- Differential weighting of score gains (e.g. at the top or low performance end).
- Costs of the collection and validation of data to construct usable data bases usually form part of the existing budget of the assessment that is used. Costs involved with modelling and analysis, by comparison, are relatively modest.

- **Statistical issues**

- Adjustments for student, school or contextual characteristics, if the model used requires adjustments (e.g. not fixed effects).
- Variance of value-added estimates, which can also serve for confidence intervals of school value-added scores.
- Inter-temporal stability of value-added scores is also affected by changes in the assessment instrument being used, changes in contextual data, and the volatility of results for smaller schools with fewer students. In cases where there are a large number of small schools (*i.e.* with less than 30 students), these could be grouped by geographic region or administrative unit to allow inclusion in the analysis (OECD, 2008, p. 196). When possible, a three-year moving average for a school's value-added score is recommended to address issues of instability.
- Bias as a measure of fundamental inaccuracy and robustness to departures from underlying assumptions regarding the nature of the data used (e.g. omission of certain variables), or the structure of the model, or both. This is particularly relevant for education systems with a high diversity of students but with limited quantity and quality of student-level data. In addition, if mobility rates are positively correlated with students from disadvantaged backgrounds or ethnic groups, then this should be taken into consideration.
- Mean squared error, which suggests exploring random effect models, but less variability in these models, is achieved through allowing higher bias.
- Missing data at either the student or school level should be addressed through initial data quality evaluations and corresponding corrective measures.
- Degree of similarity between the value-added values produced by different models. In deciding the model to employ, exploratory modelling should use real student data to identify differences in scores produced by different models and parameters, and the potential implications. Evidence from trials and studies suggests, however, that simpler models should be favoured, particularly in earlier stages of development, unless there are clear advantages, given the purpose of the modelling and country-specific conditions (OECD, 2008). More complex models could also be used internally to monitor the public use of simpler models, or to identify individual schools or jurisdictions showing significant differences. Such discrepancies would require further data collection to identify their source.

- **Implementation issues**

- Establishing policy objectives and school performance measures.
- Choosing an appropriate value-added model or mixed method.
- Developing an adequate database.

- Conducting a rigorous and effective pilot, perhaps on a sub-set of schools, to determine the appropriate contextual characteristics to include, further data requirements that may be needed, and to conduct sensitivity analysis of selected models. Stakeholder engagement, communication and capacity-building issues should also be addressed during the pilot phase.⁴
- Continuous monitoring of the value-added scores produced.
- Costs relating to communication, information and capacity building with stakeholders regarding value-added modelling and results.

5.2 THE IMPORTANCE OF QUALITY DATA AND INFORMATION

As education systems invest in improving the quality and equity of educational services, the efficiency of their investments is becoming increasingly important. This has provided a clear impetus in several countries to improve the quality and coverage of data and information systems in education. Robust data and information systems are both a prerequisite and a continuous process of improvement for value-added methods. As with other forms of analysis, poor data quality can lead to increased variance and bias of the value-added results, which in turn can undermine the credibility of actions and consequences linked to the results.

For the development of value-added models, therefore, the completeness and quality of four main types of data and information should be addressed:

- *Student assessment data*, cross-referenced using student identifiers. This would include composite measures of scores, if applicable, as well as specific measures (*i.e.* minimum literacy requirements) and performance targets.
- *Student-level contextual information*, particularly if there is a need to analyse the performance of specific groups of schools or students (*e.g.* from ethnic groups or different home languages). The issue of student mobility should also be addressed.⁵
- *School level information*, particularly if there are various school types and sizes. Information on programme and policies can also be included, especially if they are to be analysed, monitored and evaluated. For systems with a high diversity of school contexts, this offers the opportunity to ensure that value-added modelling considers comparisons among similar schooling contexts.
- *School evaluation information and reports* that provide possible explanations of value-added scores, and together provide a fuller picture of school performance to which actions and consequences can be linked.

5.3 CONSEQUENCES LINKED WITH FAIR AND CREDIBLE ASSESSMENT OF SCHOOLS AND TEACHERS

Within an accountability framework, fair and credible assessments of school performance usually result in actions and consequences for teachers. Similarly, for school-improvement efforts, assessments should also provide school staff with information on what works and how to improve, as well as the opportunities to do so. Because it provides an accurate and fair measure of students' progress over time, value-added modelling can be used in both cases to evaluate the performance of schools.

The initial phases of establishing an accountability framework that includes value-added modelling should identify opportunities for school improvement efforts. Positive incentives that reinforce and enhance the performance of schools, staff and teachers, could be combined with further evaluations, assistance, and resources for under-performing schools. Chapter 7 presents a discussion of the different types of incentives that could be used to motivate teachers, for example, using the school as the unit of accountability.

As the accountability system progresses, and local capacities are established to provide adequate assistance, support and professional development options to teachers, other measures linked to school performance results could be gradually introduced. In the development of value-added modelling to support accountability and school improvement efforts, the following points are important:

- For incentives or stimuli to be effective, it is vital to ensure the credibility and fairness of performance measures, instruments, and procedures.
- Increased accountability and improvement initiatives can be implemented gradually, within the same general accountability framework. Different types of incentives and consequences could gradually be linked to assessment results.
- Value-added results can be used in conjunction with other measures and evidence of school performance. The complementary nature of different sources of evidence, including value-added results, will be determined by the gradual development of measures and capacities. Measures from different sources will ultimately provide a more complete and accurate picture of school performance, and of the school staff and processes.

Table 5.1 provides a summary of some of the main benefits and policy implications of value-added methods for accountability and school improvement (OECD, 2010).⁶

Table 5.1
Benefits and policy implications of value-added methods for accountability and school improvement

General benefits	Policy implications and possible consequences
– Can provide accurate measures of high- and low-performing aspects of an education system.	– Identification of schools with significant growth in student learning and, indirectly, school staff, that can be eligible for recognition, rewards, professional development opportunities, and modelling of effective practices.
– Can improve the identification and analysis of what is producing results (i.e. good practices).	– Schools and school staff that repeatedly demonstrate significant gains in student learning or schools that are more successful in improving performance of disadvantaged students, can be tapped to provide information, practices, and input towards the development of teaching standards based on proven practices, within a specific educational system.
– Can increase the quality, equity and transparency of accountability systems and school evaluations, thereby helping to provide educational policies for schools to improve.	– Value-added results can complement and serve to assess the quality of subjective evaluations of school and teaching practices, serving as a quantitative anchor of the assessment framework. Certain value-added results could trigger specific school evaluation procedures, for example.
– Can strengthen the development of information systems that allow schools to analyse and evaluate their performance and strengthen school evaluation efforts.	– Compensatory programmes and funding could be channelled to the students, schools, and teaching staff of the lowest-performers. Low value-added scores could trigger further evaluation actions and on-site observation to identify the causes of low performance and address them.
– Can make education funding more effective by allowing resources (human, material, financial) to be directed to where they are most needed.	– Value-added results of specific jurisdictions (states, districts, zones and municipalities), education programmes, or student groups of a specific ethnic background, for example, could be analysed to identify priority areas and actions.
– Can assist in overcoming socio-economic inequalities of a society that might be masked at the school level by indiscriminate and less accurate performance measures.	– Value-added analysis could be used to monitor and evaluate specific public programmes, or as part of their piloting phase.

Source: Developed from OECD, 2010.

5.4 CONSIDERATIONS FOR MEXICO

The main considerations and recommendations for developing value-added methods to evaluate school performance are listed below. They should be considered in light of the conclusions and recommendations provided in other chapters of this report, particularly Chapter 4.

- *Given the wide diversity of educational contexts between and within Mexican states, value-added models can offer a fair and more accurate measure of student growth and school performance.* Current efforts by SEP and state education authorities regarding the presentation and use of ENLACE results are a good starting point and could be built upon with value-added results for schools. The challenges involved in designing, planning and implementing an assessment system for accountability and school improvement that uses value-added modelling should be addressed rigorously throughout all stages, including the initial knowledge mobilisation, analysis and application phase of education reforms in Mexico.
- *The validity and reliability of the ENLACE assessment provides Mexico with a valuable opportunity to exploit the potential of the student performance data it produces annually.* Student data and appropriate modelling can be used for increased accountability of schools and teachers, but also for improvement efforts, educational interventions and the channelling of additional resources to under-performers.
- *Given the current conditions of the education system in Mexico, value-added models can be based primarily on the school as the unit of accountability,* although school zones, student groups, municipalities and states can all be used for analysis and action. Vertical equating should be among the first of the technical issues to be reviewed in the further development of ENLACE. The quality and availability of contextual information that could be used for contextualised value-added models should also be assessed.
- *The first phases of the development of value-added modelling in Mexico can be focused on non-public exercises using actual student data* to identify the weaknesses and strengths of different value-added models. Even before applying value-added methods to student performance data, schools could be grouped into socio-economic contexts, and contextualised attainment models could be used as possible precursors of full-fledged value-added analysis. The process of establishing value-added modelling can have different phases:
 - i) Stratification of similar schools (based on type and socio-economic or other relevant information) for within-group comparisons of average results of raw scores. Issues regarding quality and completeness of test data and contextual information should also be identified and addressed.
 - ii) Internal value-added modelling exercises conducted by education authorities to select models and address technical issues with data. A **three-year moving average** is suggested for the modelling. In addition, education authorities could use VAM analysis to monitor and conduct evaluation trials of specific policies, programmes and jurisdictions, such as *Programa Escuelas de Calidad*, with particular emphasis on differences within and between municipalities, school zones, states and ethnic groups, among others.
 - iii) Public information, awareness and engagement with stakeholders on the merits, challenges and opportunities of value-added modelling, which could be linked to a re-launching of the ENLACE assessment with a clear plan for its further development.
 - iv) Attributing consequences (low-stakes at first) for under-performing schools (further exploration, observation and assistance), as well as for high performers. The same value-added analyses could be used by SEP and state education authorities to identify schools that may have teachers and practices worthy of replication and modelling. Logistical issues relating to test administration should also be addressed.

- Education authorities could use VAM analysis to monitor and conduct evaluation trials of specific policies, programmes and jurisdictions. The monitoring and evaluation of public programmes like *Programa Escuelas de Calidad* could be supported by value-added methods. In addition, evaluations of schools serving particular groups of students, from different ethnic backgrounds for example, could be conducted to identify schools that are the most or least effective in contributing to the progress of these students. Analysis of school performance could also be conducted to identify differences within and between municipalities, school zones and states.

NOTES

1. A detailed discussion of school-level VAM is provided in *Measuring Improvements in Learning Outcomes – Best Practices to Assess the Value-Added of Schools* (OECD, 2008) and in the updated 2010 Spanish edition and sister publication to this report *La medición del aprendizaje de alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas* (OECD, 2010). This chapter draws heavily from both OECD publications.
2. These other countries are Belgium (Flemish Community), the Czech Republic, Denmark, France, Portugal, Spain and Sweden, which also contributed to the OECD report on VAM in 2008.
3. This includes adequate construct representation and limited construct-irrelevant variance. This may be particularly important where test-score scales for different years are vertically linked or where potential incentives for undesired practices have not been adequately addressed.
4. This may include the use of confidence intervals in presenting contextualised value-added scores, but it requires sufficient education and training to increase understanding and use by education stakeholders and the general public.
5. One important aspect regarding the use of student-level data is the existence of privacy laws that may limit the use of this information for value-added modelling, as in Poland, or that may require signed parental consent, as in Slovenia (OECD, 2008).
6. A detailed discussion of the benefits, characteristics and design issues of value-added modelling is presented in the updated OECD 2010 sister publication available in Spanish, *La Medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas*.

References

Goldschmidt, P., P. Roschewski, K. Choi, W. Auty, S. Hebbler, R. Blank and A. Williams (2005), "Policymakers' Guide to Growth Models for School Accountability: How do Accountability Models Differ?", paper commissioned by the Councils of Chief State School Officers, Washington, DC.

Martínez-Arias, R., J.L. Gaviria and M. Castro (2009), "Concepto y evolución de los modelos de valor añadido en educación", in *Revista de Educación*, Vol. 348, pp. 15-34.

McCall, M.S., G.G. Kingsbury and A. Olson (2004), *Individual Growth and School Success*, Lake Oswego, OR: Northwest Evaluation Association.

Organisation for Economic Co-operation and Development (OECD) (2007), *Learning for Tomorrow*, OECD Publishing, Paris.

OECD (2008), *Measuring Improvements in Learning Outcomes – Best Practices to Assess the Value-Added of Schools*, OECD Publishing, Paris.

OECD (2010), *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas*, OECD Publishing, Paris.

Raudenbush, S.W. (2004), "Schooling, Statistics, and Poverty: Can We Measure School Improvement?", Princeton, NJ.: Educational Testing Service.

Salieri, G., L.G. Santibañez and B. Naranjo (2010), *State-Level Teacher Evaluation and Incentive Practices in Mexico: Diagnostic Study*, study commissioned by the OECD for the Co-operation Agreement between the OECD and the government of Mexico.

Wilson, D. (2004), "Which Ranking? The Impact of a 'Value-Added' Measure of Secondary School Performance", *Public Money and Management*, January, pp. 37-45.

Zúniga Molina, L. and J.L. 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.

CHAPTER 6

In-Service Teacher Evaluation: Policy and Implementation Issues

6.1 International practices	80
6.2 Four key questions evaluation systems must address	82
6.3 Considerations for Mexico	83

As discussed in previous chapters, education systems should provide access to basic education for all children and improve student learning. This chapter addresses one of the key factors for improving student learning: the quality of teaching. Research has clearly shown that the quality of teaching, and therefore the performance of every individual teacher, is the factor that has the greatest effect on student achievement (Manzi and Sclafani, 2010; OECD, 2009b). At the same time, there is compelling evidence, as discussed in Chapter 2, that higher educational achievement is strongly related to economic growth, with benefits to society as well as to the individual (OECD, 2010a). Teacher evaluation systems should therefore help to ensure that every classroom has an effective teacher, even in the most challenging environments.¹ The chapter begins by briefly reviewing some of the main elements of teacher evaluation systems based on international practices. It then considers the basic policy dimensions and issues commonly involved in implementing teacher evaluation systems. The chapter concludes with a series of considerations and recommendations for Mexico to support current and future efforts aimed at establishing an effective in-service teacher evaluation system.

6.1 INTERNATIONAL PRACTICES

Building a highly skilled professional educator workforce is central to a country's ability to improve the outcomes of schooling for its young people (Manzi and Sclafani, 2010; OECD, 2005). Continuous improvement and accountability require robust and accurate data and measurement systems that allow not only the tracking of student and school progress, but also intervention in a timely way with appropriate support. It must be stressed that public accountability also implies the responsibility of educational entities outside the school (e.g. districts, state governments). Schools themselves, particularly in very challenging environments, may not be able to adequately address shortcomings and problems relating to systemic or contextual issues. In many countries, even in the world's wealthiest such as the United States, inequities exist among schools in terms not only of the backgrounds and needs of students, but also of the resources and the professional qualifications of school staff that they are able to attract and retain (OECD, 2009b; NAE, 2009).

Thus, teacher evaluation across OECD countries forms part of a broader framework of accountability regarding the effectiveness of educational systems, institutions and actors (OECD, 2007). Within a larger context of public accountability, fair and effective teacher evaluations can provide crucial information for improvement and additional support. As the whole system and all of the actors need to be held accountable for student learning and growth, evaluation initiatives, including teacher evaluation, should be part of this comprehensive mechanism of aligned efforts, resources and objectives. In this context, teacher evaluation also functions as a quality assurance mechanism that provides a diagnostic picture of current performance levels, as well as evidence for decision making (OECD, 2007).

The recent trend in reform in many countries is towards test- and performance-based accountability (Sahlberg, 2009), as a way to ensure that overall reforms are fair and to provide evidence to support this. For evaluation to be effective, however, it is important that policies be based on shared responsibility and trust (Sahlberg, 2009).² This is particularly relevant for Mexico, where a robust in-service evaluation system that uses a wide array of instruments to measure teacher performance could also foster a culture that values the teaching profession. In such a system, every school and every teacher follows good teaching practices and meets expectations, and continuous development is a daily task.

In order for teachers to know what areas to focus on for improvement, as well as what constitutes "good" teaching practice, summative evaluation based on clear expectations and teaching standards can provide important information. Results from summative evaluations can serve as an important source of evidence to hold teachers accountable to expectations and professional performance (OECD, 2007). A clear conception of what is considered "good teaching" and the creation of teaching standards are fundamental to the development of a teacher evaluation system.

Recent and current reform efforts in the United States provide a rich example of the challenges and issues faced by different levels of government and schools in attempting to increase achievement, accountability, and provide fair and accurate teacher evaluation results. Although the United States has not been a top performer in PISA, it provides a valuable example of how local jurisdictions, such as states and school districts, can develop teacher evaluation systems within broader, federal guidelines. The case of the state of Delaware, for example, provides an interesting approach in the use of student learning outcome data in their *Performance Appraisal System*, where teachers cannot be rated effective or better unless their students demonstrate satisfactory levels of learning growth (Delaware Department of Education, 2010).

Another important issue regarding standards and evaluation is coherence. Specifically, coherence between curricular and performance standards, standards of good teaching, assessment and professional development is essential (NAE, 2009). In this sense, the development of teaching standards is an important step for a standards-based approach to improving the performance of the education system through accountability.³ The United Kingdom provides a good example of coherence in aligning the assessment of student outcomes and teacher practices relating to curriculum (Qualifications and Curriculum Development Agency, 2010).

The design of teacher evaluation schemes poses challenges for countries, with a range of issues needing to be addressed (Manzi and Sclafani, 2010):

- What should be the different components of a fair teacher evaluation system?
- How should the formative and summative purposes of evaluation be balanced?
- How should teachers be engaged in the design and implementation of teacher evaluation systems?
- How can reliable standards of teaching practice be developed, implemented and evaluated to form the basis for such evaluations?
- How can student assessment results be used in evaluating teachers?
- What kind of stakes or consequences should be attached to the results of teacher evaluations?

Countries take different approaches to in-service teacher evaluation (OECD, 2007).⁴ Some use it primarily for formative purposes,⁵ focused on identifying weaknesses in the teaching practice of individual teachers and to support improvement. Other countries use evaluation for summative purposes, attaching certain consequences for teachers according to the evaluation results, and some for both.

What is clear, however, is that in-service teacher evaluation can improve teacher performance (Barber and Mourshed, 2007). In Chile, for instance, evidence shows a positive relationship between student averages in the SIMCE (*Sistema de Medición de Calidad de la Educación*) student achievement test and the number of teachers performing well in teacher evaluations (Manzi and Sclafani, 2010). Teacher evaluation improves teacher practice because it identifies effective teachers as well as those who need support. It helps in designing better teacher development programmes and contributes to the retention of good teachers, providing tools for the design and provision of incentives and payment schemes (OECD, 2009b). In this sense, coherence between a teacher evaluation system, continuous options for capacity-building and rewards is a complex but crucial balance countries strive to achieve.

The task is indeed complex. Besides coherence and thorough planning, teacher evaluation requires thoughtful and careful implementation or it is unlikely to have much impact on student performance. As discussed in Chapter 2, effective implementation can prove challenging. The following section presents common issues and policy areas that should be considered.

6.2 FOUR KEY QUESTIONS EVALUATION SYSTEMS MUST ADDRESS

Four areas should be considered when designing a teacher evaluation system (Mancera and Schmelkes, 2010), and are presented below:

WHY evaluate?

The two primary objectives of teacher evaluation are good educational results, which is the ultimate goal of teaching, and assessment of the teaching process. Teacher evaluation aims to ensure that teachers perform at their best to enhance student learning. At the same time, it seeks to improve a teacher's own practice by identifying strengths and weaknesses for further professional development. These two approaches commonly refer to summative and formative evaluation, respectively (OECD, 2007). Educational results depend on many factors, but three main criteria are effective teachers, effective schools and effective school leadership. Systemic reform should see the school as the unit of accountability.

WHAT to evaluate?

Teacher evaluation systems should be able to identify effective teachers and effective teaching practices. Since the ultimate goal of the education system is student learning, student outcomes should be taken into account. A teacher evaluation system therefore needs standards of good teaching and a well-planned comprehensive evaluation framework. Participation of all stakeholders, especially teachers, in the design of the framework is important for success. Evaluation should be accompanied by feedback and support for all teachers to be able to improve their performance. Most importantly, the connection between the evaluation system and professional development needs to be clear. According to Danielson (2007), the following domains must be included in the teacher evaluation framework: planning and preparation, the classroom environment, instruction and professional responsibilities.

International experience also shows that a teacher evaluation system should build upon solid standards of good teaching. Standards should have certain characteristics (Mancera and Schmelkes, 2010): *i)* cover all the teaching domains defined; *ii)* establish different levels of competence for each specific aspect that defines the domains of teacher and school work; *iii)* reflect a nuclear group of performances that should be observable in all teachers and all schools; *iv)* define and operationalise intended goals and outcomes of good teaching; and *v)* be dynamic to allow ongoing revision of the standards, so these remain accurately scaled and take account of all aspects of teaching practices. To ensure the success of an evaluation, teachers need to be involved both in the construction of the standards and in effective training.

HOW to evaluate?

The challenge is to design a system that is fair, transparent, objective and credible to teachers. Hence, it is crucial to build an evaluation system with an array of instruments for measuring teacher performance, together with mechanisms for cross-referencing information that looks at teachers from various angles, allowing teachers' performance to be judged as objectively as possible, covering most aspects of the teaching profession. Some key instruments are student performance, student portfolios, self-evaluation, interviews and knowledge tests. School visits and classroom observation help triangulate and validate results from various instruments. A comparative analysis on teacher evaluation practices by Manzi and Scalfani (2010) highlights some important issues in this regard. For example, they found that international practice differs in the instruments used and that classroom observations serve to overcome the fact that standardised tests do not cover all areas taught by teachers in some countries. Evidence also suggests that it is easier for principals to distinguish teachers whose students' achievements were low or high on standardised tests, but it is harder for them to judge those in the middle (OECD, 2007).

WHO evaluates?

The availability of trained and competent evaluators has to be guaranteed. Effective evaluators should have at least: *i)* knowledge of the work teachers carry out; *ii)* training to make expected observations; and *iii)* autonomy in relation to the evaluated teacher (Mancera and Schmelkes, 2010).

Equity is also an important factor to consider in developing an evaluation system. While a comprehensive teacher evaluation framework allows the setting of common performance measures for all teachers, it should recognise the very different situations in which teachers work. This is especially true in countries with large disparities such as the United States and Mexico. Other issues that need to be taken into account are the status of the teaching profession, concerns about the quality and fairness of education, retention problems, the stress between internal and external evaluations, the definition of teacher evaluation as such,⁶ and relatively weak accountability for students' learning.

6.3 CONSIDERATIONS FOR MEXICO

Current reform efforts in Mexico focusing on aspects of teacher selection, accountability, and assessments for evaluation, may offer useful insights for countries facing similar contexts and difficulties: heterogeneous geography, income inequalities, an ongoing decentralisation and devolution process, and a large and complex basic education system (pre-primary, primary and secondary). In this context and to support current and future efforts, the following are summary recommendations for Mexico for an in-service teacher evaluation process that allows teachers at all levels of the performance spectrum to improve, to be recognised and to contribute to overall educational results:⁷

- *Establishing consensus among stakeholders on the importance of developing a comprehensive, transparent and fair in-service teacher evaluation framework is vital.*
- *A foundation for such a framework is the development of teaching standards that provide teachers with clear guidance as to what is considered good teaching practice, and opportunities for professional development and improvement.*⁸
- *It is essential to ensure that all teachers meet minimum levels of professional performance and results.*

As indicated in its *Education Sector Programme 2007-2012*, the Mexican government sees the creation of a standards-based, in-service teacher evaluation system as one of its education priorities. The Programme indicates that evaluation is a central tool for ensuring the quality of education. Evaluation is thus considered vital for accountability, as a communication tool and as a basis for designing public policies (SEP, 2007).

Mexico has made some progress in teacher evaluation in recent years, including *Carrera Magisterial*, *Escalañón Docente*⁹ and exploratory efforts aimed at developing teacher standards, all of which should be reviewed during the design of a comprehensive teacher evaluation system. Current efforts, however, are not necessarily articulated or comprehensive, and are not based on an accepted definition of performance standards for students or teachers. The effectiveness of past efforts is also relevant. A study made by the Rand Corporation showed that *Carrera Magisterial* has had little or no impact in increasing student achievement (Santibañez *et al.*, 2007). Evaluation has traditionally been the responsibility of school principals and, to a lesser extent, of supervisors, or other educational authorities. Table 6.1 summarises the evaluation practices currently used in Mexico.

A key question is therefore how Mexico could effectively begin to build a comprehensive in-service teacher evaluation system. As discussed in Chapter 2, a key step is to learn from international experiences but to adapt them to the conditions, constraints and opportunities of the Mexican educational system (also suggested by Mancera and Schmelkes, 2010). To accomplish this, Mancera and Schmelkes (2010) suggest implementation steps that are essentially sequential but that could also be done simultaneously depending on circumstances and opportunities (Table 6.2).

Table 6.1

General overview of teacher evaluation practices in Mexico

FUNCTION	INSTRUMENT	PURPOSE
Access to teacher education (Escuelas Normales)	Several	Selection/Admission
Initial studies certification	Exam presented at the end of initial teacher training – <i>EGEL/CENEVAL (Examen de término de los estudios en Escuelas Normales)</i>	Formative/Summative
Selection	Selection exam (<i>Concurso nacional de oposición para la obtención de plazas docentes</i>)	Summative = To obtain a teaching post (<i>plaza docente</i>)
Horizontal promotion (continuous training)	Exam – <i>Carrera Magisterial/ Examen de preparación profesional</i> Exam – <i>Carrera Magisterial/ ENAMS (Exámenes Nacionales para la actualización de los maestros en Servicio)</i> Exam (seeks to calculate student learning outcomes) – <i>Carrera Magisterial/ Evaluación de Aprovechamiento Escolar</i>	Summative Formative/Summative Formative/Summative
Recognition and stimuli	Calculation model (<i>Programa de Estímulos a la Calidad Docente</i>) (based on ENLACE results, school types and socio-economic context)	Recognition, with the intention of becoming an incentives programme in the future

Source: Translated and adapted from Zorrilla, 2009.

Table 6.2

Implementation steps in sequential order

IMPLEMENTATION STEPS	
A. General initial steps	1. Making the case for a teacher evaluation system 2. Involving stakeholders 3. Involving local authorities 4. Identifying a champion for the plan 5. Ensuring funding
B. Creation of the evaluation framework	1. Developing standards for teaching 2. Developing valid performance measures 3. Building a robust data management system 4. Training competent evaluators
C. Preparation in schools	1. Training school leaders and teachers on the evaluation system 2. Creating access to feedback and improvement 3. Implementing a communication plan for teachers and principals
D. Piloting	1. Designing a pilot programme to test design, instruments and evaluators
E. Full implementation	1. Preparing transitions from pilot to full implementation 2. Attaching consequences to the evaluation system 3. Creating of an evaluation plan
F. Monitoring and evaluation	1. Defining indicators 2. Establishing baselines 3. Defining information sources and process

Source: Developed and adapted from Mancera and Schmelkes, 2010, and OECD, 2009b.

Evidence on teacher performance and student learning outcomes should be gathered from multiple sources. In addition to interviews, portfolios, classroom observations and teacher knowledge tests, the use of standardised tests of students to evaluate teachers' performance is fundamental. As discussed in previous chapters, Mexico already applies a national student assessment, ENLACE, that could serve to identify schools with above-average teacher performance. ENLACE is a major asset of the Mexican education system and could be used as part of a collective assessment process, with the school as the unit of accountability. For individual teacher summative assessments, ENLACE would need to evolve to include value-added components, so that it can measure more clearly the contribution of schools and eventually teachers to the learning of their students in specific contexts.

Designing, piloting and implementing a comprehensive, transparent and fair in-service teacher evaluation system should be a gradual process. It took Chile ten years to design and implement its system, and education results are still lagging. Mexico also has a clear need to improve student performance. Based on discussions of the OECD Steering Group on Evaluation and Teacher Incentive Policies and the Steering Group on School Management and Teacher Policy, Mancera and Schmelkes (2010) have made 11 recommendations, as listed in Table 6.3, for Mexico to move forward in this matter.

Table 6.3

Summary of specific recommendations for Mexico regarding teacher evaluation

-
- | |
|--|
| Recommendation 1. Establish a leadership structure and clear rules for the governance of the evaluation system. |
| Recommendation 2. Establish a technical unit that will be responsible for the implementation of the evaluation. |
| Recommendation 3. Develop standards for teaching. |
| Recommendation 4. Design an in-service teacher evaluation model that gradually evolves from a purely formative system to one that combines formative and summative aspects. |
| Recommendation 5. Define the instruments for the in-service teacher evaluation system. |
| Recommendation 6. Develop a support system for school-based development that leads to the improvement of teacher practice, and a system that monitors this improvement. |
| Recommendation 7. Train evaluators. |
| Recommendation 8. Reduce administrative duties of supervisors and principals, and increase school autonomy. |
| Recommendation 9. Prepare a programme for ENLACE to enable the measurement of value added. |
| Recommendation 10. Gain momentum with key stakeholders towards establishing the teacher evaluation system. |
| Recommendation 11. Pilot and evaluate the design, instruments and evaluators in different contexts, before rolling out the evaluation system to the entire system of schools. |
-

Source: Mancera and Schmelkes, 2010.

In addition to the creation of standards, Mexico should reach a consensus on the importance of designing and implementing a comprehensive, transparent and fair in-service teacher evaluation system. Building and operating a framework for teacher evaluation is a long and complex endeavour. Political and administrative changes need to be navigated and stakeholders should be involved in the process. Both national and state-level authorities should take an active part in the design, along with input by local authorities. Teacher unions and civil society should also be involved in making major decisions. All relevant stakeholders should be represented on a designated body responsible for ensuring the implementation of a system that achieves trust and support from teachers and society in general. It is important to establish mechanisms for continuous teacher training and in the case of Mexico, formative evaluation should be established and tested before introducing significant consequences for individual teachers (Mancera and Schmelkes, 2010).

In the context of increasing accountability and providing opportunities for capacity-building and professional development for teachers, it is important that all teachers meet minimum levels of professional performance and results. Growth in student learning should be one of the evaluation criteria. In this way, a teacher would not be considered effective unless their students demonstrate satisfactory levels of student growth, while a teacher would not be rated ineffective if their students show satisfactory levels of student growth. It is also essential that basic issues such as attendance, punctuality and time-on-task can be included in the earlier stages of the teacher evaluation framework as a way of getting all teachers to perform at capacity. Including basic criteria such as these can produce considerable and timely gains for the teacher evaluation system in a cost-efficient manner (*i.e.* ensuring that all of the “low-hanging fruit” is collected first).

As discussed in Chapter 3, evidence from studies in the United States and developing countries shows that teacher absences can strongly impact student learning, especially in the poorest areas and remote communities. The *Teaching and Learning International Survey* (TALIS) conducted by the OECD shows that for a quarter of Mexican teachers, only between 40% and 60% of instruction time is actually spent teaching (OECD, 2009a). In addition, principals report that 70% of instruction in their school is hampered by teachers arriving late, by teachers being absent, or by teachers not having adequately prepared their lessons (OECD, 2009a). Bearing in mind the socio-economic disparity in conditions across Mexico, basic professional performance, as a sound basis upon which to build capacities, is one of the main issues that teacher policies and evaluation must address.

NOTES

1. This chapter draws on two expert papers commissioned as part of the Co-operation Agreement between Mexico and the OECD: “Report on In-Service Teacher Evaluation and Development Practices in a Comparative Perspective” by Jorge Manzi and Susan Sclafani (2010), and “Specific Policy Recommendations on the Development of a Comprehensive In-Service Teacher Evaluation Framework” by Carlos Mancera and Sylvia Schmelkes (2010).
2. Responsibility- and trust-based reforms involve the gradual building of a culture of responsibility and trust within the education system that values the professionalism of teachers and principals in judging what is best for students and in reporting their learning progress. They often involve channelling resources and support to schools and students who are at risk of failure or of being left behind.
3. An often cited reference is C. Danielson’s *Framework for Teaching*, Perrenoud (2004), although others may also be relevant: Rewards and Incentives Group (2009); Ontario Ministry of Education (2009); Khim Ong (2008); and Singapore Ministry of Education (2006).
4. Existing schemes of teacher evaluation in OECD education systems take multiple forms. They differ in terms of scope and methods of teacher evaluation, criteria and standards, and data-gathering instruments, according to the educational context and tradition, the actors involved in the design and implementation of the evaluation system and the primary purpose of the evaluation. The consequences of evaluations on teachers’ careers also vary. Although the single promotion table and the single salary scale remain widespread, several countries link their teacher appraisal system either to recognition and rewards, whether financial or not, or to professional development opportunities (OECD, 2009b).
5. Formative evaluations are essential to assess the quality of teacher performance and are a useful instrument to underpin teachers’ professional development. A formative evaluation system, designed and oriented mainly to improve the quality of instruction, can also serve to identify good teachers.
6. An important goal of teaching policies in Europe is to attract and retain the best candidates to the profession. In Latin America, it is to improve the quality of existing teachers (Manzi and Sclafani, 2010).
7. These recommendations should be considered in the context of the OECD recommendations on teacher professional development and school leadership (OECD, 2010b), as the quality of schools has an important impact on teaching. For a broad description of current teacher policy in Mexico, see Chapter 3 of *Improving Schools: Strategies for Action in Mexico* (OECD, 2010b).
8. Teaching standards should also reflect content standards. This is particularly important for Mexico given current reform efforts in this area.
9. Both *Carrera Magisterial* and *Escalafón Docente* serve mainly as mechanisms related to promotion. See Santibañez *et al.* (2007) and OECD (2010b).

References

Alcázar, L., H. Rogers, N. Chaudhury, J. Hammer, M. Kremer and K. Muralidharan (2006), "Why are Teachers Absent? Probing Service Delivery in Peruvian Primary Schools?", *International Journal of Educational Research*, Vol. 45, pp. 117-136.

Barber, M. and M. Mourshed (2007), *How the World's Best-Performing School Systems Come Out on Top*, McKinsey & Company, London.

Clotfelter, C., H. Ladd and J. Vigdor (2008), "Are Teacher Absences Worth Worrying about in the U.S.?", National Center for Analysis of Longitudinal Data in Education Research Working Paper 24 (May, 2008 version), accessed at www.nber.org/papers/w13648.pdf.

Danielson, C. (2007), *Enhancing Professional Practice: A Framework for Teaching*, Association for Supervision and Curriculum Development, Second Edition, Alexandria, VA.

Danielson, C. (2008), *The Handbook for Enhancing Professional Practice: Using the Framework for Teaching in Your School*, Association for Supervision & Curriculum Development (ASCD), 2nd ed., Alexandria, VA.

Delaware Department of Education (2010), "Race to the Top: Application for Funding", CFDA No. 84395A, Narrative, The State of Delaware.

Khim Ong, K. et al. (2008), "Teacher Appraisal and its Outcomes in Singapore Primary Schools", in *Journal of Educational Administration*, pp. 39-54.

Mancera, C. and S. Schmelkes (2010), "Specific Policy Recommendations on the Development of a Comprehensive In-Service Teacher Evaluation Framework", OECD Publishing, Paris.

Manzi, J. and S. Sclafani (2010), "Report on In-Service Teacher Evaluation and Development Practices in Comparative Perspective", OECD Publishing, Paris.

Miller, R., R. Murnane and J. Willet (2007), *Do Teacher Absences Impact Student Achievement? Longitudinal Evidence from One Urban School District*, National Bureau of Economic Research, Working Paper 13356, accessed at www.nber.org/papers/w13356.pdf.

Organisation for Economic Co-operation and Development (OECD) (2005), *Teachers Matter: Attracting, Developing and Retaining Effective Teachers*, OECD Publishing, Paris.

OECD (2007), "Teacher Evaluation: Current Practices in OECD Countries and Literature Review", M. Isoré, Education Working Paper, OECD Publishing, Paris.

OECD (2009a), *Creating Effective Teaching and Learning Environments: First Results from TALIS*, OECD Publishing, Paris.

OECD (2009b), *Evaluating and Rewarding the Quality of Teachers*, OECD Publishing, Paris.

OECD (2010a), *The High Cost of Low Education Performance*, OECD Publishing, Paris.

OECD (2010b), *Improving Schools: Strategies for Action in Mexico*, OECD Directorate for Education, Education Policy Implementation, OECD Publishing, Paris, accessed at http://www.oecd.org/document/4/0,3746,en_2649_39263231_41829700_1_1_1_1,00.html.

Ontario Ministry of Education (2009), "Overview of the Ontario Teacher Performance Appraisal (TPA) System".

Perrenoud, P. (2004), *Diez nuevas competencias para enseñar. Invitación al viaje*, Grao, Madrid.

Qualifications and Curriculum Development Agency (2010), *A Big Picture of the Secondary Curriculum*, Qualifications and Curriculum Development Agency, http://curriculum.qcda.gov.uk/uploads/BigPicture_sec_05_tcm8-15743.pdf.

Rewards and Incentives Group (2009), "Teachers' and Head Teacher's Performance Management: Guidance", accessed 24 April 2010 from www.teachernet.gov.uk/management/payandperformance/performancemanagement.

Rogers, F.H. and **E. Vegas** (2009), *No More Cutting Class? Reducing Teacher Absence and Providing Incentives for Performance*, World Bank Policy Research Working Paper 4847, World Bank, Washington, DC.

Sahlberg, P. (2009), *A Short History of Education Reform in Finland*, Helsinki, accessed 24 April 2010 from www.pasisahlberg.com/index.php?id=64.

Santibáñez, L., J. Martinez, A. Datar, P. McEwan, C. Setodji and **R. Basuto-Dávila** (2007), "Breaking Ground: Analysis of the Assessment System and Impact of Mexico's Teacher Incentive Program 'Carrera Magisterial'", *RAND Technical Report*, RAND Corporation, Santa Monica, CA.

SEP (2007), *Programa Sectorial de Educación 2007-2012*, accessed 24 April 2010 from http://upepe.sep.gob.mx/prog_sec.pdf.

Singapore Ministry of Education (2006), "Singapore Staff Appraisal (Education Service)", in MOE, Singapore.

Zorrilla, M. (2009), Presented during the Workshop: *TALLER OCDE-MEXICO Hacia un sistema de evaluación docente en México: Prácticas internacionales, criterios y mecanismos. Panorama general de prácticas de evaluación en México como parte de las políticas de profesionalización docente*, Mexico City, 1-2 December 2009.

CHAPTER 7

Incentives for In-Service Teachers

7.1 Types of teacher incentives.....	93
7.2 National guidelines and local implementation: Finding the right balance.....	103
7.3 Piloting, monitoring and evaluating incentives.....	103
7.4 Considerations for Mexico.....	105
Appendix 7A.....	107

As discussed briefly in the previous chapter, research evidence has confirmed the importance of quality teachers to student learning (OECD, 2009a, 2005). Research in the United States dating back almost 20 years found that students whose teachers are at the top of the effectiveness range achieve as much as an additional year of growth in learning when compared with those students whose teacher is near the bottom of the range (Hanushek, 1992). In another often cited study based on data sets from Tennessee, United States, researchers found that if two similarly performing students in Grade 2 of primary school are assigned to high and low-performing teachers for each of the subsequent three years, the difference in their performance at the end of the three years may be as much as 54 percentile points (Sanders and Rivers, 1996). The most recent thinking regarding effective educational systems has confirmed that the quality of an educational system cannot exceed the quality of its teachers (OECD, 2009a; McKinsey & Company, 2007). In the face of increasing international competition and economic downturn, however, governments are increasingly being forced to do more with less. In addition to attracting and retaining the most qualified professionals into the teaching profession, one important challenge becomes how educational systems can motivate and support teachers already in service to improve performance and increase student achievement.¹

Policies and programmes that evaluate and reward effective teachers are thus becoming more central to education reform. This is true for developed as well as developing economies where there is greater need for strategic and efficient use of compensation systems as a mechanism for improving teacher quality and student learning. Indeed, an organisation's compensation system is arguably its most important human-resource management system (Ehrenberg and Milkovich, 1987; Lawler, 1981).

Although performance rewards have been used effectively in other fields of employment, particularly in the private sector (Lazear, 1996), their recent use in the education sector, particularly for teachers, is still being explored, monitored and evaluated. Policy makers in education are now considering incentives that do not involve compensation, as well as those that do.² In-service teacher incentives are increasingly being considered, therefore, because they reflect the overwhelming importance of teachers, the focus on improving student achievement through improved teaching practices, and the cost-effectiveness of rewarding superior performance.

Incentives are used for a variety of specific purposes: to encourage teachers to work in difficult-to-staff schools or to teach certain subjects, to undertake training or to assume roles and responsibilities in the school (OECD, 2009a). Educational systems can also reward schools and teachers for acceptable or above-average performance, while sanctioning under-performers to encourage improvement efforts. The exact nature and design of the teacher incentives scheme will thus depend on the policy objectives established. Incentives in this sense can be an important element within a broader accountability framework focused on student achievement.

This chapter provides an overview of past and current international practices in relation to different types of teacher incentives, and attempts to identify lessons learned and best practices for the development of an effective teacher incentives policy.³ It begins by detailing various types of incentives, including input-based and performance-based incentives for teachers. In particular, an emphasis is placed on recent policy shifts away from input-based compensation systems to outcome-based systems. These systems focus on the measurable outcomes or products of education (such as those reflected in student test results) from those that reward characteristics or activities thought to improve teacher effectiveness (OECD, 2009a). The chapter thus provides an overview of the main design elements for teacher incentive policies. Educational authorities in a particular country, along with stakeholders, will need to determine the right combination of monetary and non-monetary incentives and stimuli that will be most effective. Regardless of the rewards and consequences for teachers, however, current practices indicate that for teachers to be considered effective, their students should demonstrate satisfactory levels of growth, while no teacher should be rated as ineffective if students show satisfactory levels of growth (Delaware Department of Education, 2010). Drawn from international experience and research practice, the appendix at the end of the chapter also provides detailed considerations

for the piloting, monitoring and evaluation of teacher incentive policies and programmes. This appendix is intended to provide educational authorities and related stakeholders with guidelines regarding how to design, plan and implement effective monitoring and evaluation processes.

7.1 TYPES OF TEACHER INCENTIVES

Providing financial incentives is believed to increase organisational productivity by encouraging less effective teachers to seek out more effective instructional strategies, reducing turnover among highly skilled teachers, and attracting teachers who are particularly good at the targeted activities. To the extent that incentives lead to a more effective teacher workforce, those incentives can have positive impacts on teaching practices well beyond the operational life of a particular programme.

Input-based incentives

Input-based incentives are those that reward teachers for activities believed to improve instructional quality or student outcomes. In other words, input-based incentives encourage teachers to pursue activities thought to improve effectiveness, but stop short of indicating whether the teachers' overall effectiveness actually improves. This feature has spurred many to criticise, and in some case move away from, these types of incentives. Among the most common examples of input-based incentives are incentives for teacher education, knowledge- and skill-based pay, and career ladders. The *Carrera Magisterial Programme* in Mexico, before 2006, is an illustrative example.

Knowledge- and skills-based incentives, and career ladders

Knowledge- and skills-based incentives reward teachers for the acquisition of additional knowledge and skills thought to improve a teacher's overall effectiveness. Knowledge- and skills-based pay is distinct from the traditional salary schedule's pay increase for advanced degrees. Rather, the incentives focus on the ongoing improvement and growth of teacher skills and competencies throughout their career. The premise behind this pay is to motivate teachers to improve their knowledge and skills, thereby improving (in theory) their instructional practices, increasing their effectiveness and improving student achievement (Odden and Kelley, 1997).

Knowledge- and skills-based incentives may reward teachers above and beyond the traditional salary schedule increase for pursuing an advanced degree. Other examples include taking additional professional development coursework, pursuing dual certification, and completing a teaching portfolio. In Singapore, in addition to 100 hours of professional development offered each year, teachers are eligible for reimbursements of SGD 400-700 for expenses related to improving their knowledge and skills. Qualifying expenses include purchasing software, taking courses or training, subscriptions to journals and joining professional organisations.

It is not uncommon for knowledge- and skills-based pay to involve linking the salary increase with external evaluations and assessments. In the United States, for example, such bonuses are frequently awarded for teachers participating in and completing National Board for Professional Teaching Standards Certification, or for passing other teacher assessment exams (Podgursky and Springer, 2007). Depending on the design of the system, teachers can receive single-year payments, payments for a pre-determined number of years or payments for the duration of their professional careers.

Akin to knowledge- and skills-based pay, career ladders create different categories, or levels, that reward teachers with higher salaries. Each level is associated with increased mastery or competence, and distinguishes novice teachers from expert or master teachers. In most instances, career ladders require teachers to pass either formal or informal credentialing, or to assume additional responsibilities, in order to advance on the career ladder. Generally, teachers must meet multiple criteria in order to advance. In the United Kingdom, for instance, the upper scale of the salary schedule includes performance-pay based upon eight nationally agreed teaching standards.

Teachers who have reached the top of the salary scale (which can be reached in five steps) may apply to pass the “threshold” by being assessed against the standards, which are grouped into five categories: knowledge and understanding (one standard); teaching and assessment (three standards); pupil progress (one standard); wider professional effectiveness (two standards); and professional characteristics (one standard). Principals are trained to complete the assessment, which is then confirmed by external reviewers. In the upper pay scales, along which teachers can progress every two years based on annual evaluations, teacher salaries are 12-25% higher. Career ladders provide new roles for teachers with additional pay and responsibilities as they increase their knowledge and skills. In addition to offering advancement possibilities for good teachers (thereby encouraging their continued service), career ladder programmes “counteract stagnation by varying teachers’ responsibilities and activities at each level” (Cresap *et al.*, 1984, p. 22).

Performance-based compensation

In part due to criticism of input-based incentives, and also as a result of the increasingly widespread use of standardised assessments, incentive pay plans have become increasingly focused on rewarding teachers or groups of teachers based on their performance or outputs. Performance-based compensation systems have taken on heightened relevance in light of the strong evidence that many input-based systems are financially inefficient. In the United States, for example, researchers Roza and Miller (2009) found USD 8.6 trillion was spent paying for master’s degrees, accounting for upwards of 3.3% of total education expenditures in some states despite research evidence demonstrating that non-subject-specific master’s degrees bear no relation to improved teacher effectiveness.

Performance-based incentives are more directly associated with student learning than input-based incentives, and are generally grouped into two categories: those based on educational outcomes and those based on educational processes. While these programmes may use multiple measures to evaluate teacher performance, and incorporate elements found in career ladder or knowledge- and skills-based pay plans, student outcomes on standardised assessments remain paramount in determining bonus award eligibility.

Performance outcome incentives

Performance-outcome incentives are those that rely on quantifiable measures of student achievement as the primary evaluative tool. Such awards are distinguished in that the reward “hinges on student outcomes attributed to a particular teacher or group of teachers” (Podgursky and Springer, 2007). Student scores on standardised assessments, graduation rates, dropout rates, attendance rates and grade-promotion rates are examples of the types of measures that can be used. Outcome incentives can be thought of as those that reward the products of teaching.

Increasingly, performance-outcome incentives are based on student performance in standardised assessments. This is due in large part to improvements in the development and implementation of both standardised assessments across nations and longitudinal data systems that link individual students and teachers, as well as more sophisticated measurement of student, teacher and school performance (OECD, 2010b). Results from student assessment are also used for teacher feedback and appraisal. According to the OECD’s Teaching and Learning International Survey (TALIS), of over 70 000 teachers in 23 countries, 65% of teachers report that student test scores were a moderate or highly important criterion in the appraisal or feedback that they received. In Bulgaria, Malaysia, Mexico and Poland, this percentage was well over 80% (OECD, 2009b).

In the United States, the *No Child Left Behind* law, passed in 2002, requires all states to test students annually in Grades 3 to 8, and once in upper school, in both reading and mathematics. Each state is also required to publish the school level results of these assessments. As a result of this and other initiatives, many states have developed expansive data systems of student achievement, including student-teacher linkages. Combined with policy initiatives that attempt to improve student achievement, federal-, state-, district-, and school-level incentive pay programmes based on student achievement have become widespread in the United States.

As of the late 1980s, the United Kingdom introduced a new school accountability policy through the Education Reform Act of 1988. The system instituted a set of national educational reforms aimed at raising achievement. England's school accountability system has evolved into a broad and more complex framework of accountability including school improvement partners, continuous learning requirements for senior teachers, rewards and sanctions for schools determined to be performing poorly, school choice, internal and external assessment of instructional and organisational processes, and an emphasis on ongoing professional development and support for teachers and leaders.

One significant advance in student assessment data is the use of statistical methods known as value-added modelling (VAM), as discussed in Chapter 5, which can isolate the contribution of schools, school groups and, in some cases, individual teachers, to student learning. Although raw scores may provide useful information on the learning achievement of students and can support accountability systems and improvement efforts, raw scores alone may present distorted and biased results of school performance (see Chapter 5 for a more detailed discussion). With value-added models, however, researchers are able to measure more precisely student growth, and how much or little of that growth is attributable to schools and to teachers, independent of the students' socio-economic background and other family characteristics. This type of information provides much greater insight into the effectiveness of schools and, indirectly, teachers. While value-added models are still at the forefront of innovation and policy initiatives in only a handful of countries (OECD, 2010b), they offer an important option for systems that require more precise and accurate measures of school performance and teacher effectiveness than can be obtained from raw test scores or other measures that may suffer from bias and issues of validity.

Examples of performance-outcome incentives

In the United States, the Teacher Incentive Fund (TIF) offers federal grants to support the development and implementation of performance-based teacher and principal compensation systems in high-needs schools. This Fund has awarded grants since 2007 to some 95 recipients, including 62 awards in September 2010. Designed and implemented at the local level, all programmes must include student academic achievement as well as classroom evaluations among other factors in determining incentive awards. In a 2010 analysis of the first 33 Fund grants, Heyburn *et al.* found, as expected, that employees could receive awards based on student performance, with slightly more than half relying on some form of value-added measures. Originally funded at USD 99 million in 2006, the Fund currently receives USD 400 million, with grant awards starting from USD 500 000. Additionally USD 950 million has been sought by the United States Department of Education for 2011 for a new Teacher and Leader Innovation Fund that would support the development and implementation of performance-oriented approaches to recruiting, retaining and rewarding highly effective educators (Springer, 2010).

In addition to federal programmes, several states and districts have implemented incentives programmes based, at least in part, on student achievement. In Texas, the Governor's Educator Excellence Award Program (GEEAP) consisted of three distinct programmes and was the largest state-level system with funding at nearly USD 250 million. The first of these programmes, the Governor's Educator Excellence Grant, provided three-year grants to 99 high-achieving schools that were in the top third of schools serving highly disadvantaged students. Participant schools designed their own plans, but were required to include student achievement scores and measures of teacher collaboration. For the student achievement component, awards were most typically based on an individual teacher's performance, and were based on student attainment levels rather than on student growth scores. Over the course of the programmes, participant teachers received awards ranging from less than USD 100 to more than USD 10 000, with an average award amount of approximately USD 2 300 across all three years (Springer *et al.*, 2009).

The second Texas programme, Texas Educator Excellence Grant, awarded more than 1 000 one-year grants in each of two years to high-achieving schools serving a high proportion of disadvantaged students. As with the Governor's Educator Excellence Grant, plans were designed at the school and district level, though student achievement and teacher collaboration were again required measures. Schools participating in the

Texas Educator Excellence Grant typically used student attainment scores rather than student growth scores in establishing the performance measure. Interestingly, roughly 90% of schools in both years of the programme relied solely on state-standardised assessments to measure student outcomes, despite an array of alternative sources. Average bonus awards in the first year of the programme were USD 1 982, while in the second year average awards were USD 2 094.

The District Awards for Teacher Excellence Program, the third Texas programme, provided funds at the district level and was open to all districts in the state. As with the first two programmes, plans were designed locally and were required to include student achievement measures. In total, 203 districts representing 16% of all districts elected to participate in the District Awards Program (Springer *et al.*, 2009). Of those who chose not to participate, a clear majority was concerned about the programme's impact on school culture and professional collegiality (Springer, 2010).

In New York City, the School-Based Performance Bonus Pay Programme, implemented in approximately 200 K-12 public schools during the 2007/08 school year, provides school-level awards bonuses of up to USD 3 000 per full-time union member working at the school for meeting performance targets defined by the city's accountability system. A randomised experiment, this programme is unique in that the performance outcome of interest is the outcome of the school as a whole. Indeed, measurement and distribution of awards at the school level were paramount to securing support from the local teachers association. In addition to monitoring student achievement scores, the district collects survey data on student, parent and teacher perceptions of the school learning environment, including items on academic expectations, communication, engagement and safety. Teams of experienced educators also conduct two- to three-day on-site visits to review the quality of a school's institutional and instructional programme. The school learning environment survey and external quality reviews provide means for studying the causal effect of this programme on intermediate outcomes. Unlike the great majority of pay-for-performance programmes, this one has been designed as a randomised experiment, thereby allowing researchers to measure its impact on student learning.

The research evidence about the effect of these programmes on student achievement is still in its infancy. As with input-based programmes in the past, more research is needed to assess the effectiveness of these programmes as a mechanism for improving student learning, both in the short- and long-term. Several comprehensive studies are ongoing or in their initial stages, including a randomised evaluation of the Teacher Incentive Fund Program in the United States (Springer, 2010). Moving forward, it will be important to garner lessons learned from research to refine and enhance these programmes.

Other measures of student learning

As noted in earlier chapters, test scores are not the only outcome measure of student performance. Student attendance rates, graduation rates, drop-out rates and portfolios are other common examples of criteria that can be included in incentive plans. In Austin, Texas, Student Learning Objectives (SLOs) are the primary measure of individual teacher success in their pay-for-performance initiative, known as the Austin Independent School District (AISD) REACH programme (Springer, 2010; AISD, 2010). SLOs are teacher-designed, data based instructional goals. Through examination of student achievement data, teachers work with their principal to develop two SLOs based on student need. At least one SLO must target the teacher's class as a whole, while the second SLO can focus on a particular sub-group of students. SLOs must be based on standards established by the state government, address classroom needs, align with the goals of the Campus Improvement Plan, and satisfy standards of rigour for both performance and assessment. Each SLO must also be approved by both the teacher's principal and AISD REACH staff to ensure objectives are appropriate and rigorous. At the end of the school year, students are evaluated to determine whether the performance objective has been met. Teachers can earn up to USD 3 000 for meeting both SLO goals, while principals can earn up to USD 4 500 for facilitating the process. A nearly identical system of Student Growth Objectives (SGOs) provides a primary student achievement outcome variable in Denver, Colorado's pay-for-performance programme, ProComp (Springer, 2010).

While SLOs and SGOs engage teachers in using data to drive instruction and curricular goals, they lack the objectivity of a standardised assessment. Indeed, an evaluation of AISD REACH found 83% of teachers in the first year and 81% in the second year met at least one of their SLOs (Cornetto *et al.*, 2010). This high level of attainment raises questions about the overall rigour of the goals and whether they are appropriate or true proxies for teacher effectiveness. It underscores the importance of ensuring that goals and measures are both rigorous and objective in designing an incentive programme.

Process incentives

Unlike incentives based on performance outcomes, performance process incentives focus on teacher practices and the skills of the teacher in the classroom. In some states in Germany, for instance, teachers are able to advance through the salary levels more quickly than normal, when they undertake performance evaluations based on classroom observations. In Baden-Wuerttemberg, 10% of all schools' teachers can advance a salary step based on outside evaluations, whereas in North-Rhine Westphalia, teachers may only advance in salary steps based on performance evaluations.

In an effort to establish objective, valid and reliable measures of teacher performance that do not involve student performance data, systems have sought to incorporate measures of teacher effectiveness that are not negatively influenced by teachers' unions, local loyalties or other interests (Buck and Greene, 2010). In Singapore, for example, teacher evaluation sessions include an extensive planning and goal-setting meeting at the start of the academic year, a mid-year review and meeting, and a final evaluation including recommendations and feedback from senior teachers and department/subject area chairs who have worked with the teacher. Teachers who receive high ratings on their performance evaluation are eligible for bonuses equal to one to three months of their salary.

In Switzerland, Zurich's salary-effective qualification system reviews teachers in the middle phase of their career. In this case, a team of specially trained professionals observes and interviews the teacher, and prepares an extensive report of the teacher's pedagogical approach, with successful teachers eligible for pay increases of 1-4% for the following four years.

Performance process evaluations can provide useful insight on a teacher's classroom practices, and can be beneficial in providing tailored feedback to the teacher. As previously mentioned, however, it is critical to link such observations with objective measures of students learning. In 2010, the Gates Foundation gave USD 45 million to launch the Measures of Effective Teaching Project (Springer, 2010). The two-year project, led by independent education researchers, "seeks to uncover and develop a set of measures that work together to form a more complete indicator of a teacher's impact on student achievement" (Gates Foundation, 2010). The researchers are working with 3 000 teachers in seven school districts across the United States and are collecting information including student feedback through surveys, samples of student work, supplemental student assessments, videotaped classroom lessons, teacher reflections on their videotaped lessons, assessment of teachers' ability to recognise and diagnose student problems, and teacher surveys on working conditions. The aim of the project is not to investigate the design of incentives, but rather to determine the actual instructional practices of effective teachers. The results of this comprehensive study may provide invaluable information for evaluation systems internationally, including more rigorous criteria on which to link teacher pay and performance.

Market-oriented reforms

Market-oriented compensation rewards teachers based on market factors including teaching location and subject, and is characteristically a response to broader supply and demand considerations within the particular school, district or state. The most prevalent forms of market-oriented compensation reward teachers assigned to a hard-to-staff school or teaching in a hard-to-staff subject.

A common mechanism for awarding market-oriented compensation is payment of recruitment and retention bonuses. Recruitment stipends are awarded to new teachers at a school, and are used as a tool for motivating the teacher to accept a teaching position in the school (generally a hard-to-staff school). Recruitment stipends are most often distributed regardless of past teaching experience (meaning a teacher can be an experienced teacher moving to a new assignment or a novice teacher), and can be distributed at the individual school or district level. In some locations, recruitment stipends are awards for multiple years after a teacher accepts the new position. Retention stipends are distributed to teachers who continue working in their position and/or school following the summer vacation. Retention stipends are generally thought to improve student achievement in hard-to-staff schools by decreasing the continual turn-over and overabundance of novice teachers emblematic of these schools.

In Singapore, the Continuity, Experience and Commitment programme (CONNECT) offers retention payments to retiring teachers in addition to the social security assistance available to all teachers. The payments range from SGD 4 200-6 200 per year for teachers with 1 to 15 years experience, and SGD 3 200 per year for more experienced teachers. Funds are also available for withdrawal every three to five years to help encourage retention until retirement. In Queensland, Australia, in addition to recruitment incentives, teachers are offered AUD 5 000 per year to remain in their teaching position.

Hard-to-staff subjects

Hard-to-staff subject incentives are those offered to teachers or prospective teachers qualified to teach subjects that experience shortages of qualified teachers, and can be based on school, district or state needs. A 2005 OECD report found that over 30% of students in participating countries attended schools with shortages of qualified foreign language, mathematics, science and technology teachers (OECD, 2005). More generally, science, technology and mathematics experience the largest shortages internationally as fewer undergraduate students majoring in these fields enter the profession. Due to the often significant pay differentials between public and private sector workers in these fields, incentives have become a popular mechanism for narrowing this gap. In the United States, hard-to-staff subject incentives are offered most frequently for mathematics, science and special education teachers, with stipends typically ranging from USD 1 500 to USD 3 000. Additionally, many districts and states offer tuition assistance and loan forgiveness for teachers of designated shortage areas. In Utah, qualified mathematics and science teachers receive a USD 5 000 signing bonus for agreeing to stay in the district for at least four years. In New York City, where housing costs exceed the national average, qualified mathematics, science and special education teachers can earn housing assistance of up to USD 15 000 in moving expenses, down payments or security deposits on homes, and a housing stipend of USD 400 for two years (Springer, 2010).

In England and Wales, loan forgiveness programmes repay up to GBP 16 000 in tuition expenses over the course of ten years for qualified mathematics, science, special education and technology teachers. The Golden Hello programme offers an additional GBP 4 000 for teachers who complete their first year of teaching in a shortage area. In Australia and Ireland, incentives are offered to indigenous language teachers, while Brussels offers incentives to French-language instruction teachers (OECD, 2009a).

Hard-to-staff schools

Incentives for teaching in hard-to-staff schools are those targeted towards schools, districts or geographic regions that experience a chronic dearth of qualified teachers in all or most subject areas. These schools are generally located in urban or remote rural areas, and commonly serve a high proportion of economically disadvantaged, minority or low-achieving students, and are characterised by a high turn-over rate among teachers as well as relatively inexperienced teachers.

Many countries and states offer incentives for recruiting and retaining teachers to remote rural regions. In Australia, the Queensland Remote Area Incentive Scheme (Sclafani and Tucker, 2006) offers rural teachers AUD 5 000 for travel to and from metropolitan areas, in addition to retention incentives of up to AUD 5 000.

In Chile, teachers working in areas of geographic isolation, extreme poverty or difficult access can receive incentives up to 30% of their annual salary. In other countries, including Korea, China and, at one time, Bolivia, teachers are required to spend time teaching in remote areas in order to be eligible for promotion.

While recruitment to remote rural environments is a challenge in some countries, other countries and states experience the opposite: extreme shortages in urban settings. The Staffing Incentive Allowance in New Zealand, for instance, provides NZD 1 000 for teachers who serve for three years in designated high priority city schools. Similarly, in the United States, teachers in Los Angeles, California, Charlotte, North Carolina, and New York City, New York, were eligible for incentive stipends of USD 2 000 to USD 3 000 for serving in high priority schools.

Beyond monetary incentives

Incentive programmes can be designed to meet local, regional or national needs, and can take a variety of forms, including offering both financial and non-financial incentives. In fact, Vegas and Umansky (2005) identify a range of possible teacher incentives, including recognition and prestige, gratification of intrinsic motivation, job stability, working conditions with adequate resources and support, and so on. Similarly, Darling-Hammond (1997) found a strong association between teachers' views of the quality of support provided by their school administration and their plans for staying in a teaching position. Darling-Hammond also found the availability of resources and teachers' ability to have a voice in school decisions were associated with their plans to stay. Likewise, a lack of positive work environments has been found to contribute to high attrition rates from schools in high-poverty areas and schools with high-minority student populations (Loeb, Darling-Hammond and Luczak, 2005).

In the United Kingdom, the Fast Track Teaching Programme provided incentives to promising new teachers who exhibit leadership potential. Designed to identify school leaders earlier in their career, teachers in the programme are eligible for GBP 2 000 per year, as well as coaching, mentoring and leadership development activities. The Raising Standards and Tackling Workload – A National Agreement programme, also in England, seeks to address teacher dissatisfaction with working conditions. The programme aims to improve overall teacher satisfaction and retention by reducing the hours in the teacher contract, providing guaranteed planning time, reducing paperwork, and providing various support staff including bursars, administrative, technical and classroom support staff. The programme also created new career paths for teachers (Springer, 2010).

In Singapore, programmes offer both financial and non-financial incentives for teachers. For instance, teachers are allowed to accrue a month of full-pay for half-time service or half-pay for full-time study for each year of service, up to 12 months. Teachers are then able to use their time in any way that benefits the education system, including teaching in an independent or foreign school, or working in the private sector in a field related to their teaching subject area. Along similar lines, deferred salary leave plans in Canada and in some parts of the United States allow teachers to defer a part of their current salary to use for paid leave at a later time. Programmes in Norway, Germany and the Netherlands allow teachers to reduce their course load for a small reduction in salary (OECD, 2009a).

While there is no standard design for these programmes, the intent is to improve the overall working conditions for teachers, thereby in theory improving their overall effectiveness and rate of retention. Milanowski (2007) found prospective teachers to be more concerned with the qualities of their teaching assignment, including supportiveness of the principals, availability of induction programmes and curricular flexibility at the school than they were with increases in salary.

The importance of evaluating evaluations

The need for comprehensive evaluations of teacher evaluation and incentive programmes cannot be overstated. While the preceding sections provide an overview of the diversity of incentive programmes that

rely on student achievement outcomes, rigorous evaluations are needed to ensure the effectiveness of the most recent programmes, and that they are cost-effective (Springer, 2010). In an evaluation of the Texas Educator Excellence Grant programme, Springer *et al.* (2009) found that while the attitudes and behaviours of school personnel, school environment and teacher turnover were affected by the programme, the evidence suggested the programme had no strong, systematic effect on student achievement gains. Nor were there consistent associations between design features of the programme and student achievement gains. A recent study conducted in the United Kingdom, on the other hand, found that performance incentives improved student learning on average by 40% of a grade, as measured by standardised testing (Atkinson *et al.*, 2009). This study, however, was based on data that matches students to individual teachers, over time; this cannot be done with current data systems in most countries, although efforts are underway (OECD, 2010b). Thus, specific teacher incentive programmes will need to be developed and implemented by education authorities and stakeholders with careful consideration of the local context, constraints and opportunities as discussed in Chapter 2.

Design considerations

While determining the optimal type of incentive programme (input-based, performance-based or a hybrid) is important, equally important is the design of the programme. For instance, should awards be targeted at the individual teacher, the team or the school? What minimum performance thresholds should be met to earn an award, and what are appropriate ranges for incentive awards? Should all teachers be eligible for the award or only top performers? The ways in which programmes are structured to answer these questions can have a significant impact on the effectiveness of the programme as a whole. This next section identifies the key design features to consider in developing an incentive programme.

Incentive structure

Incentive structure refers to the scheme or mechanism that guides the allocation of awards in a pay-for-performance system. In some cases, only a limited number or proportion of teachers can earn an award while in others, any teacher who meets a predetermined performance standard will receive an award. The two main forms of incentive structure are *rank-order tournaments* and *fixed-performance contracts*.

Rank-order tournaments are incentive structures in which a fixed proportion of teachers, teams or schools earn an award based on their performance ranking relative to the ranking of other teachers, teams or schools. The key feature that distinguishes tournaments from other incentive pay structures is that compensation depends on relative performance rather than absolute performance. By fixing in advance the total number of individuals or teams that can earn bonus awards, rank-order tournaments reduce the potential for unknown financial exposure. On the other hand, they are more likely to create competition within schools, and may disrupt the collaborative culture of teaching by placing teachers in competition with one another. Moreover, if the performance standard is unclear or perceived as unattainable, lower performers may not respond to the incentive.

A *fixed-performance contract* defines the performance standard teachers, teams or schools must meet to earn an award. In contrast to a rank-order tournament, any individual who meets the predetermined performance standard is eligible to receive the benefits, regardless of the performance of others. The potential financial exposure with fixed performance contracts is significant, but the structure is generally more palatable to teachers and teacher associations as they are not placing teachers in direct competition with one another.

Unit of accountability

As discussed in earlier chapters, the unit of accountability refers to the entity responsible for a measurable product or service. Performance on that measurable dimension determines bonus eligibility. The unit of accountability can be defined in multiple ways, including the *individual* teacher, a grade-level or departmental *team* of teachers, all employees within a *school*, or some combination thereof (*hybrid*).

An *individual* unit of accountability refers to an incentive pay policy in which the performance of the individual teacher determines award eligibility. The *team* unit of accountability refers to award eligibility being the product of aggregated performance among members of a group, where the size of a group can range from as few as two employees to all employees within a school. Teams may include grade-level teams and disciplinary or inter-disciplinary departments. The *school* unit of accountability refers to award eligibility being the product of aggregated performance among all employees within a school.

While individual awards create the strongest connection between variations in award size and performance, and help offset free-rider issues that can be emblematic of team- and school-based awards, it is more difficult to identify effective individual teaching than it is for a team or school. This is particularly the case as some standardised test results are less readily attributable to a single teacher (e.g. if reading and language arts are taught by different teachers). When awards are distributed at the team or school level, however, such attribution is generally less problematic. Additionally, team- and school-level incentives generally promote greater social cohesion and collegiality, as well as feelings of fairness, as workers come together around shared goals. Because the connection between awards and performance is weaker in team- and school-based programmes, it is important to consider this trade-off in light of the stated goals of the programme.

Performance standards and thresholds

Performance standards and thresholds determine the required level of performance that secures a bonus award for an individual teacher and/or team of teachers. These design components affect the likely number of teachers that may earn a bonus as well as what standard they must meet to earn an award. Given the number of different models available, and the complexities involved when trying to operationalise different performance standards and threshold models, we focus here on two sound models: the *limited linear model* and the *limited step function model*.

A *limited linear model* is one in which a baseline performance threshold establishes the minimum possible level of performance that is associated with a bonus. Beyond this point, a positive linear relationship exists between increasing performance and the size of a bonus award. At some bonus “cap”, further increases in performance are not associated with a larger bonus amount. The range between the minimum and maximum thresholds is referred to as the incentive zone. The model allows for progress at any point in the incentive zone to increase the award amount, yet still limits bonus size of low performers and reduces the financial risk of extremely high performance. Similar to the limited linear model, the *limited step function model* uses minimum and maximum thresholds to determine the incentive zone. In this model, however, the incentive zone is comprised of different performance thresholds, each of which is associated with a particular bonus amount. Increasing award amounts are only realised by meeting the different thresholds within the incentive zone.

Both the limited linear and limited step function models have the benefits of maintaining a clear linkage between performance goals and bonus award payouts, while minimising financial risk. These advantages do not free those designing an incentive pay system, however, from needing to consider the right balance between motivating performance and keeping performance goals attainable. Performance standards and thresholds must consider both attainability (*i.e.* the degree to which they are attainable by average teachers) that motivates behaviour and exclusivity that can discriminate performance.

Size and distribution of performance award

The size of bonus or payout level refers to the amount of the total bonus award a school, team of teachers or individual can earn. Distribution refers to the guidelines that determine the share of teachers who receive a bonus award and how bonuses vary among teachers. While no clear guidance exists on the optimal size of a bonus in a teacher incentive pay programme, several studies suggest that bonus awards for teachers have often been so small that the motivational value has been compromised (Malen, 1999; Chamberlin *et al.*, 2002; Heinrich and Marschke, 2007; Taylor *et al.* 2009). For instance, in two of Texas’s programmes, minimum bonus

amounts were as low as 0.4% and 1.5% of teachers' monthly salary. Similarly, bonuses in Bolivia's *Incentivo Colectivo a Escuelas* began at 5% of monthly salary, while Chile's *Sistema Nacional de Evaluación de Desempeño de los Establecimientos Educativos* began at 4.7% (OECD, 2009a; Springer, 2010).

The magnitude of the maximum bonus award, however, varies. Mexico's *Carrera Magisterial* programme offers bonuses of up to 224% of a teacher's monthly salary. Similarly, the POINT Experiment in Nashville, Tennessee, and the Governor's Educator Excellence Grant programme in Texas both offered maximum awards equal to 270% of monthly salary (Springer, 2010). These levels of maximum awards suggest that these increases are both a response to past reform efforts providing unappealing award amounts, and a result of increased interest in finding out whether teachers respond to substantial bonus awards, even if the incentive system lacks a complete array of measures.

Bonus award distribution systems determine how evenly an incentive pay programme distributes rewards to eligible employees. An egalitarian distribution plan distributes incentive money widely, in contrast to plans that reward larger sums of money to fewer schools, teams of teachers or individuals. However, there is no clear guidance as to whether an incentive pay programme should make a large number of relatively small awards to teachers in a school, or reward a smaller number of teachers with a relatively large award (Taylor, Springer and Ehlert, 2009). Proponents argue that individualist reward plans help create a meritocracy able to retain an organisation's highest performers, attract similar talent over the long run, send a clear signal to the lowest performers to improve or move elsewhere, and are more cost-effective (Milgrom and Roberts, 1992; Zenger and Marshall, 1992; Ehrenberg and Smith, 1994; Pfeffer and Langston, 1993). At the same time, a growing body of research suggests egalitarian distributions promote co-operation and group performance, which are critical in participative organisations. Milgrom and Roberts (1992) suggested, moreover, that greater pay dispersion may elevate the performance of the lowest performers. The lack of solid evidence as to the most effective strategy for determining bonus levels further underscores the importance of piloting programmes so as to determine the most effective thresholds and components for the locality in which the programme is being implemented.

Payout frequency

Payout frequency is the rate of award distribution and the time interval between assessment of the targeted activity and the distribution of awards. Most incentive pay systems in the education sector distribute awards on an annual basis, usually corresponding with the academic year. This is generally due, at least in incentive systems that rely on student test scores, to there being a yearly testing schedule. In practice, there is often a large time gap between performance and payout. After testing of students, it can take months to analyse results, calculate payouts and distribute bonus awards to teachers. It is not uncommon for teachers to receive bonus awards mid-way through the school year following the year in which performance was evaluated.

Psychology literature suggests incentives are most effective when performance is rewarded with minimum time between the desired action and award distribution (Skinner, 1981). Too much time between those events can reduce a teacher's association between what they did well and receiving the award. Consistent reinforcement is a stronger motivational force, suggesting multiple award payouts throughout the year may be optimal from a motivation perspective. Paying incentives more than once a year, however, places greater demand on resources to monitor performance in an ongoing manner.

Because there is no model of in-service teacher incentives that can be transported wholly into a different context, the relevant and best practices from international examples should be analysed and considered in light of local constraints and opportunities (as discussed in Chapter 2). In addition, piloting of incentive schemes before a full roll-out of a programme is also important to ensure effectiveness and proper policy design (see Appendix at the end of this chapter). As most OECD member countries have decentralised education systems that have undergone varying degrees of devolution of authority to local bodies, incentive policies for teachers must address issues around the location of authority. The following section presents a brief overview of the importance of this issue for in-service teacher incentives.

7.2 NATIONAL GUIDELINES AND LOCAL IMPLEMENTATION: FINDING THE RIGHT BALANCE

For decades, decentralisation of the education sector has been part of larger political reforms undertaken in several countries (e.g. in Latin America, India and China). Governments striving for enhanced coverage, efficiency and quality in the public sector, including education, have sought an appropriate balance of decision-making authority and functions between central, local and regional governments: provincial (Argentina), state (Mexico, United States), municipal (Chile), districts (United States) or a combination thereof (Brazil and the United States) (OECD, 2007; World Bank, 2004b). These reforms have produced varied institutional arrangements, with roles and responsibilities spread across different bodies and jurisdictional levels. Not surprisingly, the focus has often centred on the effectiveness of different forms of education decentralisation to increase student learning outcomes.

In this context, one of the most significant findings is that with increased decentralisation, devolution and indeed school empowerment and autonomy, clear mechanisms of accountability become more important. Reviews of reform efforts in different countries have shown that decentralisation, devolution and increased autonomy (e.g. for schools) will not necessarily lead to increased student achievement unless there are incentives to local authorities which are held accountable (World Bank, 2004a, 2004b). Incentives, therefore, can be considered an operationalisation of a wider accountability framework. From research and international practices, it is apparent that devolution of educational processes can be more favourable to improved student achievement and less detrimental to equity of educational opportunities for all students, even those from disadvantaged socio-economic backgrounds, where there is a clearly established framework of accountability.

The specific arrangements for teacher incentives across and between levels of authority can take multiple forms. In four of the most studied and well-known incentives systems in the United States, for example, two are arrangements between the state and multiple districts (Texas and Ohio), one is an arrangement between a district and the local chapter of the teachers' union (Denver ProComp), and one is an example of a federally funded⁴ single-district system (Chicago). The challenge is therefore for countries to identify and design the most effective and appropriate devolution arrangement between relevant jurisdictions and bodies. This is particularly relevant for the establishment of content and performance standards. Although good teaching practices may vary depending on local conditions, certain elements, such as content standards and curricular reforms, may be better undertaken by the central decision-making body. This will ensure that there is a unified vision of what students are expected to know and know how to do. Similarly, teacher standards may serve a fair and valid teacher evaluation framework if they are clearly established at the central or national level, albeit with input and considerations from a wide variety of teachers and teaching contexts. Thus, for federal and decentralised systems, policy reforms must necessarily address the issue of *de jure* and *de facto* devolution. This implies considerations that go beyond simple bilateral accountability relationships based on funding (e.g. federal funding and state/local implementation). More importantly, perhaps, educational systems must find the right balance between national and centrally established standards and guidelines, and local application and innovation.

7.3 PILOTING, MONITORING AND EVALUATING INCENTIVES

In addition to the ongoing discussion regarding the design, effectiveness and consequences of incentive schemes for teachers, international practices clearly suggest the value of careful piloting, evaluating and monitoring of such reform initiatives prior to and during the course of full-scale implementation. Based on international practices regarding the design, planning, implementation and evaluation of incentive programmes, the Appendix at the end of this chapter provides examples of best practices and provides specific recommendations regarding piloting, monitoring and evaluating of these policy initiatives. The following elements highlight the importance of conducting pilots before implementing full-scale national programmes and are summarised in Table 7.1.

- **Carefully designed, monitored and evaluated piloting exercises can provide objective evidence regarding the most effective and efficient ways to implement teacher incentive schemes.** Incentive pilots conducted in India and Israel,⁵ for example, show that rigorously evaluated pilots can provide evidence regarding the effectiveness of teacher incentive programmes. Objective and reliable evidence can thus inform discussions and engagement among education stakeholders regarding specific programme elements. Those elements that are points of disagreement with teachers' unions, for example, could be tested empirically under properly designed piloting conditions in participating jurisdictions.
- **Pilot experiences can identify and address potential undesirable results of teacher incentive schemes before a full rollout of a national programme.** Experience from India, Israel, Kenya, the United Kingdom and the United States shows that perverse effects that can be properly addressed through pilots include "teaching to the test", exclusion of low-performing or challenged students, corruption or falsification of performance measures such as attendance and test scores, and the neglect of non-tested subjects such as citizenship, music and art.
- **Piloting can identify and quantify unexpected expenditures during implementation.** Pilots can also identify the more realistic extent of expenditures of a programme, at a fraction of what a national program would cost to implement and evaluate. Ineffective or inefficient programme elements can be identified in a pilot and corrected or discarded in the design of a broader national programme, making it more cost-effective.
- **Piloting exercises conducted in specific contexts can be designed to compensate for prerequisite conditions that may not be in place on a national level,** while also providing insight into how these conditions can be gradually put in place on a larger scale. Updated and accurate information regarding school teaching staff and student groups, for example, would be collected and validated rigorously for those schools participating in the pilot samples.

Table 7.1

Summary of benefits of conducting pilots

Opportunity to assess feasibility of programme elements (e.g. teacher performance measures, criteria and type/amount of rewards).

Opportunity to identify effectiveness of programme in achieving desired results on teacher behaviours and student learning.

Limited scale allows for a controlled and robust evaluation process.

Opportunity to test innovative practices, if pilots are properly designed to be evaluated.

Opportunity to identify and evaluate unintended negative effects of programme elements.

Opportunity to assess true cost of programme elements, as well as unforeseen expenses of implementation.

Opportunity to identify data quality control needed for effective implementation of programme.

Opportunity to implement different pilot models in parallel and in different contexts to compare and contrast effectiveness.

Opportunity to identify and gauge the degree of support or resistance from stakeholders (e.g. teachers, principals, administrators and local chapters of unions).

Can provide evidence and data to support larger-scale programmes (i.e. state or national).

Source: Developed and adapted from H. Rogers, 2009, OECD, 2009a, and the work of the OECD Steering Group on Evaluation and Teacher Incentive Policies.

7.4 CONSIDERATIONS FOR MEXICO

Although performance rewards have been used effectively in other fields of employment, their recent use in the education sector, particularly for teachers, is still being explored, monitored and evaluated. Thus, SEP, state educational authorities and stakeholders will need to determine the specific combination of monetary and non-monetary incentives and stimuli that will be most effective in Mexico. Regardless of the rewards or consequences that are linked to results, however, for teachers to be considered *effective*, their students should demonstrate satisfactory levels of growth, while no teacher should be rated as *ineffective* if students show satisfactory levels of growth. Based on the review of the main characteristics and examples of teacher incentives and stimuli from different countries, the following are the main summary recommendations for Mexico for the establishment of an effective policy of in-service teacher incentives and stimuli.

- ***For an effective and sustained in-service teacher incentive policy, the following five principles should guide its development:***
 - i) Incentives should reflect the quality of teaching. The criterion for success of the incentive programme would not simply be better pay for better performing teachers, but the contribution of teachers to improved student learning outcomes.
 - ii) The incentive system should recognise and support the individual teacher, the team of teachers at the school and the profession as a whole. Incentives should be embedded in a system that supports the continuous improvement of students, teachers, schools and the education system. In the longer-term, incentives based on tests should be complemented by a sound human-resource management capacity in schools and at local levels that can accurately assess the quality of work, with robust external validation and corroboration methods jointly owned by government and the teaching profession.
 - iii) The incentive system should build on a sound understanding of what motivates teachers and should embrace multiple dimensions of motivation, with the aim to foster an attractive work environment, create and facilitate advancement along a career path, provide access to professional development, and identify and promote effective teaching practices. Incentives and stimuli should therefore consider financial and non-pecuniary incentives, such as working conditions, material inputs for schools and classrooms, social recognition, enhanced training and professional development opportunities, or a combination thereof.
 - iv) The incentive system should provide good feedback mechanisms and access to professional development, to ensure that teachers who do not receive the incentive understand what they can do to improve performance and have incentives to change behaviour. It should foster a culture based on evidence and data.
 - v) The incentive system should reward both good performance and relative improvement, and consider the value added by teachers and schools, net of socio-economic factors. While value-added analytical models are being developed, however, simpler methods can be employed to ensure that students, schools and teachers are compared with those in similar contexts (e.g. socio-economic stratification and/or contextualised attainment models).
- ***It is important to clearly distinguish an in-service teacher-incentive policy from other teacher-related programmes that may appear to be similar, but that do not fundamentally provide incentives to teachers to improve performance.*** Incentive policies should be communicated clearly to teaching professionals in advance of the assessments and measures that will be used for the awards. In addition, each eligible teacher should have a probability of being rewarded for outstanding performance that is greater than zero, which is currently not the case for similar programmes in Mexico. Of particular importance will be finding a balance between national guidelines for the incentives (and at least partial funding), and state-level flexibility and co-participation in resources (financial or otherwise) for incentives and stimuli to teachers.

Finally, a pilot of possible incentive programmes is highly recommended to ensure viability and cost-effectiveness of policy design. Pilot exercises should be rigorously monitored and evaluated in order to be most useful and worthwhile, with a base line and as much control as possible (e.g. randomised or quasi-experimental trials, if conditions allow).

- ***In-service teacher incentives in Mexico should motivate individual teachers to improve performance, but use the school as the basic unit of accountability***, given the current state and prospects of data systems, and the quality of information available. Even with insufficiently robust data and information systems, however, local education authorities can develop measures to confirm and validate the eligibility of teachers for incentive awards (e.g. with on-site inspections and data validation of student, school and teacher information). As a robust and credible individual teacher evaluation system is developed, incentive policies could be modified to ensure that teachers are able to receive incentives individually in the future.⁶ For school incentives, schools should be made publicly accountable for the additional resources received. If schools have discretion over the allocation of resources provided by the incentives, mechanisms should ensure transparency and the progressive involvement of relevant stakeholders, including parents and local school councils.
- ***Financial and non-financial incentives and stimuli to teachers should be based on a fair and adequate assessment and evaluation process***. Given the wide diversity of the Mexican educational system, a valid and reliable assessment process to identify eligible teachers for incentives needs to be developed. The success of incentives is directly linked to the credibility and fairness of the assessment and evaluation process upon which they are based. Models that take into consideration the socio-economic diversity of Mexican students, as well as other factors that can largely influence student performance, such as Spanish as a second language and ethnicity, for example, should be used when making comparisons among schools and their teachers. Special-education schools and programmes, as well as pre-primary schools, could be evaluated on the basis of appropriate measures of teacher performance and student learning, where possible. Given the diversity between and within states, the incentives policy should also consider a relative premium for disadvantaged rural schools, as opposed to non-disadvantaged urban or rural schools. Incentives should also support continued improvement of schools and teachers across the entire performance spectrum.

APPENDIX 7A

PILOTING, MONITORING AND EVALUATING TEACHER INCENTIVE PROGRAMMES: RECOMMENDED PRACTICES

The purpose of this Appendix is to provide policy makers and their staff with an overview of issues relating to piloting, evaluating and monitoring of teacher incentive programmes. From the beginning of a programme, appropriate questions need to be asked about the goals of the programme: How is it being implemented? How are teachers responding to particular elements of the programme? What are some of the unforeseen consequences that are being observed? Combined with causal questions, which seek information on the results of the programme, each of these will provide a solid understanding of what the programme is designed to do, what it is actually doing in practice, and its results. Piloting, monitoring and evaluation are distinct yet interdependent processes for a well-planned and well-implemented incentive programme. This is particularly true given the fact that teacher incentives are an innovative field and some of the most relevant examples are only just starting to be evaluated to help explain what works, in which settings, and under what conditions.

Box 7A.1 Important definitions

What is meant by piloting? In essence, a pilot is a controlled, systematic roll-out of a programme in a restricted setting prior to full-scale implementation across an entire education system. A pilot allows examination of how a programme translates from design to implementation and the impact of implementation for outcomes such as teacher attitudes and behaviours, as well as student learning. Through a pilot, there is a chance to learn how reality may differ from what was envisioned and how to adapt implementation and design to better meet those intended goals. Additionally, an education system can identify resources that might be needed to support better implementation, such as training for educators, a communications plan or more technical expertise.

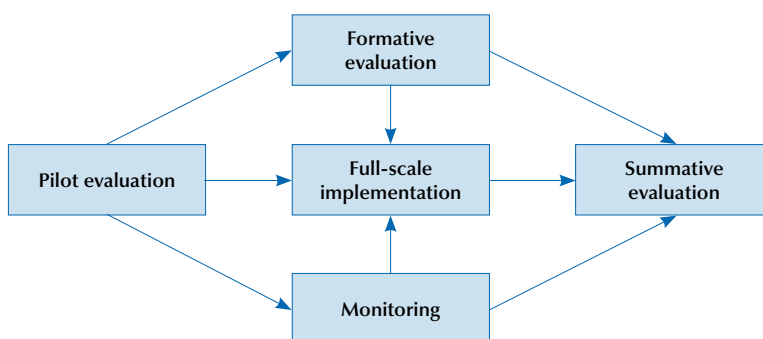
What is meant by evaluation? Programme evaluation is not entirely separate from piloting: any successful pilot must be accompanied by a thorough and rigorous evaluation in order to learn from the experience and adapt prior to launching the full programme. The intent of evaluation – whether for a pilot programme or for the full scale version – is to learn empirically *what* works and *why* using appropriate and justifiable methods. There are three categories of evaluation: descriptive evaluations focus on describing the context, while process evaluations illustrate the nature of programme implementation and experiences, while causal evaluations identify the programme's impact. Ultimately, a well-designed evaluation can explore the nature of programme implementation and design, what the impacts are for students, teachers and other stakeholders, and under what conditions.

What is meant by monitoring? While evaluation addresses empirical questions related to the nature, quality and outcomes of a programme, monitoring is a process used to ensure that programmes are implemented according to the design and to provide quality assurances on programme operation. Monitoring systems ensure the incentive programme is running as intended, including such aspects as communication campaigns and timeliness of award payouts. It can also serve as a quality control process, to ensure, for example, that data systems necessary for incentive pay implementation are working appropriately.

The lessons learned from a pilot evaluation influence the way in which a programme is implemented, provide insight into areas of implementation that might require close monitoring, and offer considerations for full-scale evaluation including areas of focus, methods and forecasting of potential challenges. Once the programme is fully implemented, the monitoring process, together with lessons from formative evaluation, can be used to improve upon implementation. Additionally, formative evaluation results may shed light on areas needing more or less monitoring whereas issues identified through monitoring should also be considered in crafting evaluation research methods and questions of interest. All of these inputs and processes are finally considered in a summative evaluation that identifies programme outcomes given the reality of programme implementation.

Figure 7.1

Model of programme piloting, evaluation and monitoring



Source: Springer, 2010.

FOCUS AND PURPOSE OF EVALUATION

As presented in Box 7A.1, there are three categories of questions that can be addressed by an evaluation: descriptive, process and causal (McEwan, 2008; Shavelson and Towne, 2002). Table 7A.1 below provides some examples of how these categories of questions might apply to an evaluation of an incentive pay programme and includes questions commonly addressed by researchers in the field.

It should be noted that these categories should not be considered in isolation. For example, a stronger causal argument can be made when an evaluation has taken into consideration the context in which and the processes through which a programme operates. In fact, some argue that the success of a programme may be just as dependent on the reality of programme implementation and support as it is on the design of the programme itself (Goldhaber, 2009).

DESIGN AND METHODS FOR EVALUATION

Overview of evaluation methods

To address *descriptive questions*, evaluators typically use basic quantitative techniques, such as descriptive statistics that measure the central tendency of variables, the dispersion and frequency of variables, and the correlation (or association) of variables with one another (McEwan, 2008; Tukey, 1977). In the context of describing, for example, the characteristics of schools participating in an incentive pay programme, such descriptive techniques can explain the dispersion of schools by poverty level or record of past performance, or – on average – what is the type of student served by the participating schools. Using correlation, one might address whether or not the characteristics of schools are related to the likelihood of participating in the incentive pay programme.

Table 7A.1

Focus and purpose of evaluation questions in the context of incentive pay

Type of question	Questions to be addressed
Descriptive	– What are the characteristics of sites participating in the incentive pay programme? (e.g. poverty status, record of performance, type of student served?) – What is the policy and political context in which the incentive pay programme is being implemented? (e.g. recent reforms initiatives, political key players' stance on incentive pay, legislative changes?) – What are the policy guidelines that shape design features of the incentive pay programme? (e.g. rank-order versus performance contract, size of bonus award, measures of performance, individual or group awards?) – What is the nature of funding for the incentive pay programme? (e.g. sources of funding, how much, sustainability?)
Process	– What are the experiences of participants during the implementation process? (e.g. how was the programme designed, what went according to plan and what unpredicted issues arose?) – What changes were made to system infrastructure or supports/resources to implement the incentive pay programme (e.g. modifications to data systems, human resource practices, technical support and training?) – What are the features of programme implementation and do participants understand them? (e.g. how is performance measured, how are award amounts determined, when payment is expected)
Causal	– What is the incentive pay programme's impact on teacher attitudes and behaviours? (e.g. perceptions about the programme's fairness and utility, Influence on instructional practice, teacher attendance and retention?) – What is the programme's impact on the organisational dynamics within schools? (e.g. collaboration or competition among educators, leadership practices and decision-making?) – What is the programme's impact on student learning outcomes? (e.g. gains in student achievement, types of students most impacted, subject areas most impacted?)

Source: Springer, 2010.

Process questions are typically addressed using a combination of methods but particularly qualitative methods (McEwan, 2008; Patton, 2002). Case studies, interviews and site visits/observations are invaluable for gathering details about how an incentive pay plan is developed and implemented that may be difficult to gather from quantitative data sources (e.g. administrative records, survey results) alone. A more complete and validated picture of implementation, however, certainly benefits from the triangulation of numerous data sources, both quantitative and qualitative. As an example, to examine features of programme implementation, evaluators might conduct interviews with a programme development and/or oversight team to learn about the process of programme design and the phases of implementation. Teacher interviews might indicate whether their perceptions mirror the actual intent of the programme, while a review of administrative records might reveal how components of the incentive pay plan (e.g. results from educator evaluation, payout of bonus awards) actually occur in practice.

Causal questions typically garner the greatest attention because they aim to determine whether or not a programme works. For policy makers and other practitioners, this is of utmost importance in determining the life of a programme, particularly in the current educational policy context driven by accountability and results. In the context of incentive pay, causal research might ask whether a school's participation in an incentive pay programme results in higher retention of effective teachers than in a comparable non-participating school. The underlying goal of causal research in such an example is to identify the difference in retention of effective teachers between the school taking part in the intervention and the same school if it had not been part of an incentive pay programme.

Methods for causal research are almost always quantitative in nature and can be grouped into one of three categories: non-experimental, quasi-experimental and experimental. McEwan (2008) summarised the fundamental difference between these methods succinctly: "The essential difference among categories is the degree of control exerted by the researcher over who is assigned to the program/treatment and who is assigned to the control/comparison group." Table 7A.2 presents a summary of causal research categories, rated from highest to lowest quality.

Table 7A.2

Evaluation designs to investigate the impact of programme and policy interventions

Category	Grade	Description
Experimental designs	Highest quality	Random assignment to control and treatment condition (e.g. as in a researcher's flip of a coin).
Quasi-experimental designs	Moderate quality	Use of matching, statistical controls or similar strategy to establish treatment and comparison conditions in absence of random assignment.
Non-experimental designs	Lowest quality	Correlational or observational study, with no random assignment or comparison condition. Researcher has no influence on assignment, which is completely due to self selection.

Source: Springer, 2010.

Statistical methods

Two commonly used statistical controls are regression analysis and propensity score matching. With regression analysis, factors that might explain differences in outcomes (other than the treatment alone) are statistically controlled for. For example, if trying to isolate the effect of a school-based incentive pay programme on student achievement, the regression analysis might integrate variables such as student poverty, past student performance, teacher experience and teacher attrition at the school level. Nonetheless, the controls used in regression are only those that can be observed.

Propensity score matching involves the evaluator developing a score that represents the predicted probability that a unit receives a treatment. Each unit in the treatment group is then matched to a unit in a comparison group based solely on the developed propensity scores. If a match cannot be found for a unit in the treatment group, that particular unit is removed from analysis so as to minimise differences between the treatment and comparison groups. The evaluator then estimates the effect of a programme or policy by comparing the average difference in outcomes between matched units in the treatment and comparison group (McEwan, 2008). To put propensity score matching in the context of an incentive pay programme, imagine that an evaluator develops a score (based on observable variables) that represents a school's likelihood of participating in the programme. An evaluator would then, for each school in the programme, find a similar school not in the programme but with a matched score value. At that point, an analysis of differences in teacher attendance rates, as an example, between the matched schools might be the result of the incentive pay programme. As with regression analysis, however, propensity score matching cannot guarantee that all unobserved variables are accounted for.

PREREQUISITES FOR AN EVALUATION

While few argue against the merit of a rigorous programme evaluation in providing useful feedback about whether a programme works, in which settings and under what conditions, the practical needs and pre-conditions for conducting an evaluation often lead to them being done poorly or not at all. One of the main ones is that an education system must be committed to ongoing learning driven by soundly derived evidence. In such a system, stakeholders welcome systematic, rigorous and transparent efforts to understand what works and why in order to make necessary changes over time. Stakeholders will also understand that learning about programme effectiveness takes time. In the short-run, an evaluation might be able to identify changes in professional practice or attitudes within a school, but long-run outcomes such as the quality of a teacher workforce or student achievement may take longer to determine.

Pilots and experiments of teacher incentives: International examples

One of the most common strategies is to implement and evaluate a pilot programme prior to full-scale implementation. The following are selected international examples.

- The Project on Incentives in Teaching (POINT) experiment was a three-year experimental study of middle school mathematics teachers and their students and schools conducted by the federally-funded National Center on Performance Incentives (NCPI). The POINT experiment was conducted in a large urban public school district in the State of Tennessee, United States, and sought to study the effects on student outcomes of paying teachers incentive pay bonuses of up to USD 15 000 per year on the basis of student test-score gains. Teacher volunteers were randomly assigned to either the treatment or control condition for the duration of the study. Treatment condition teachers' bonuses were based on two criteria: the progress of a teacher's mathematics students over the year as measured by their test-score gains, and the progress of a teacher's non-mathematics students over the year as measured by their test score gains. Results from this experiment are forthcoming (Springer, 2010).
- NCPI is also currently in the second year of a team-level teacher incentive pay experiment in the Round Rock Independent School District (RRISD) in Round Rock, Texas. The district organises its nine middle school teachers into grade-level interdisciplinary teams that oversee the learning experiences of distinct student groups. This study allows for exploration of how team-level awards impact student achievement in core subject areas (*i.e.* mathematics, reading, writing, social studies and science). It also allows for examination of the impact of team-level awards on teacher behaviour, institutional and organisational dynamics, and many other aspects of schooling processes. The study is especially relevant to current policy discussions as many individuals and groups argue that a grade-level team is the most appropriate level for rewarding performance in the education sector.

During the 2007-08 school year, New York City (NYC) and the United Federation of Teachers (UFT) designed the city's first school-wide incentive pay programme. Approximately USD 20 million in private funds was raised to support the pilot initiative. In November 2007, 240 (15%) NYC public schools were randomly selected for participation from a set of high-needs schools, defined by the average proficiency rating in core subject areas, poverty rates, student demographics and percentages of English language learner and special education students. Of those 240 schools, 205 (86%) agreed to participate. Beginning in the 2008-09 school year, the programme became publicly funded and expanded to include more than 400 schools (30% of all NYC public schools).

Eligible schools opted into the programme through a school compensation committee vote taken during the 2007-08 school year. Each school designed progress report targets to determine eligibility for school-wide incentive awards, which are distributed at the end of the school year. Schools meeting all performance targets can earn enough funds for all full-time UFT-represented employees to receive USD 3 000. Schools meeting 75% of targets can earn enough funds for those employees to receive USD 1 500 each. Each school's compensation committee decides how incentive awards will be distributed among employees.

An evaluation of this experimental pilot is in its early phases, with results from the first year of programme implementation available. Springer and Winters (2009) found that the school-based incentive programme had little impact on student proficiency or the environment within schools after one year of implementation. The researchers caution, however, that it is difficult to determine the effectiveness of a programme after only one year of implementation. Results from subsequent years of NYC's programme are forthcoming.

Evaluators with Mathematica Policy Research, Inc. designed an experimental evaluation of the Teacher Advancement Program (TAP) in a subset of public schools in the Chicago Public Schools (CPS) system in the United States. A commonly used model in the United States, TAP is a comprehensive system to improve teacher quality and includes an incentive pay component for teachers. The other teacher quality components of TAP include a career ladder and ongoing data-driven professional development. Under the TAP model, teachers can earn extra pay and responsibilities through promotion to Mentor or Master Teacher and can earn annual incentive bonuses based on a combination of their value added to student achievement and observed performance in the classroom. The model also includes weekly teacher cluster group meetings and regular classroom evaluations by a school leadership team to help teachers meet their performance goals. TAP has been implemented in more than 200 schools around the country (Springer, 2010).

Of the 16 CPS elementary schools that voluntarily applied for Chicago TAP and successfully completed the selection process, eight were randomly assigned to a treatment group that began implementing TAP in the 2007/08 school year and the other eight to a control group that delayed implementation until the following school year (2008/09). To complement the experimental analysis, researchers used propensity score matching procedures to create an additional comparison sample of more than 200 schools matched according to pre-intervention characteristics such as school size, school demographics, student achievement and teacher retention (Glazerman, McKie and Carey, 2009; Glazerman and Seifullah, 2010).

In the experiment's year one report, Glazerman *et al.* (2009) found that the implementation of TAP had resulted in some changes to the organisational dynamics within schools, namely the heightening of support for teachers by mentors. There was no detectable impact on student achievement scores, but there was increased teacher retention in TAP schools, although not in the district overall. After the second year researchers still did not detect any difference in student achievement gains between the treated and control group schools. There was also no significant impact on teacher retention in TAP schools or across the district (Glazerman and Seifullah, 2010). Research will continue to examine the outcomes in forthcoming years.

With funding from the US Department of Education, a randomised experiment called the Talent Transfer Initiative (TTI) has been designed in partnership with Mathematica Policy Research Inc. to test an incentive pay policy that offers additional pay to highly effective teachers if they transfer to low achieving schools.⁷ In seven large, diverse school districts throughout the country value added analysis of three years of student achievement growth data has been used to identify the top 20% of elementary, middle school English and middle school mathematics teachers (Springer, 2010).

Implementers of TTI offered teachers USD 20 000 to transfer to low achieving schools and facilitated their transfers. At the same time, the programme identified a pool of low achieving schools with teaching vacancies in the targeted grades and subjects, and the research team randomly assigned half of them to a treatment group and half to a control group. Treatment schools were eligible to hire a top-tier TTI teacher who had been offered USD 20 000 to transfer. Control schools had to fill their vacancies in the normal way. While still in its early phases, the study will follow the transfer teachers and the corresponding control teachers for two years.

In India, a series of experiments has identified positive outcomes for incentive pay policies, particularly for improving teacher attendance and students' performance on examinations. Duflo, Hanna and Ryan (2007) used random assignment to study the effect of incentive pay in rural, non-formal education sites in which the attendance of teachers was traditionally very low. The incentive pay was found to decrease teacher absences by as much as 19 percentage points and, ultimately, student test scores and graduation rates improved as well.

Another randomised experiment in India was conducted by Muralidharan and Sundararaman (2008). In this experiment of incentives in rural government primary schools, four experimental groups were designed. The first two groups received additional inputs, one in the form of an additional paraprofessional teacher, and the other in the form of a cash block grant for purchase of additional school resources. The third and fourth experimental groups received incentive pay bonuses for increased student test scores, one based on the performance of a group of teachers and the other based on individual teacher performance. All treatments (*i.e.* paraprofessional, cash block grant, incentive pay bonuses) were of equal value to the schools. Muralidharan and Sundararaman (2008) found that the incentive pay had the greatest impact on student achievement compared to control schools, but student achievement was still higher in the additional input schools as compared to the control schools. Overall, both input and incentive pay were found to be more cost effective than the status quo in control schools, with incentive pay being the most effective of the treatments.

Two studies of performance "tournaments" have been conducted in Israel (Lavy, 2002; 2004).⁸ In both of these studies the programme was designed to raise pass rates on high school exit exams in low socioeconomic high schools. Although schools were not randomly assigned to a control or treatment condition, both programmes were implemented using three formal assignment rules (*e.g.* grade range, past performance and matriculation

rate) permitting for a more rigorous evaluation design.⁹ The Israeli Teacher-Incentive Experiment was also carefully designed to minimise gaming or other opportunistic behaviour on the part of teachers and school administrators: performance measures were based on the size of the graduating cohort in order to discourage schools from encouraging transfer or dropout of poor students and from placing poor students in non-matriculation tracks).

Lavy's (2002) first study considered a tournament in which a selected group of low-performing high schools competed on the basis of school-wide performance. The top third of schools as determined by their year-to-year improvement in test scores were given awards ranging in size from USD 13 250 to USD 105 000. Teacher bonuses ranged from about USD 250 to USD 1 000, and were distributed equally to all teachers in the "winning" schools. Lavy found a positive effect on participating schools relative to a non-participating comparison group of low-performing schools. He also concluded that endowing schools with additional resources (25% of school awards had to go to capital improvements) contributed to increased student performance.

The second study examined an individual teacher bonus programme, also run as a tournament (Lavy, 2004). Essentially, teacher participants were ranked on the basis of value-added contributions to student achievement on a variety of exit exams, and bonuses were given to top performing teachers. The programme included 629 teachers of whom 302 won awards. The bonuses were substantial: as large as USD 7 500 per class on an average base pay of USD 25 000. Results indicate a positive effect in that the performance of participating teachers (*i.e.* both bonus recipients and non-recipients) rose relative to a comparison group of teachers who did not participate in the incentive programme.

In Kenya, Glewwe, Ilias and Kremer (2008) examined a group-based incentive pay programme in eight randomly selected primary schools. Incentive pay was given to schools that were ranked as top scoring or to those most improved. The authors found that student test scores did improve during the course of the incentive pay programme, but those improvements diminished after the programme's completion (*i.e.* there was no longer a detectable difference between student scores in the treatment and control schools).

DESIGN AND KEY CONSIDERATIONS FOR MONITORING

So far the chapter has addressed the role of and methods for evaluating incentive pay programmes and policies. Also important is whether a programme is being implemented according to its design. If an education system does not have the resources or ability to implement a programme or is not implementing it with fidelity, it becomes increasingly difficult for an evaluator to know whether the programme or policy is actually effective.

This section sets out the key considerations for monitoring an education system's ability to implement an incentive pay programme. It also integrates several lessons from recent examples in various countries.

Researchers and policy groups have availed themselves of the opportunity to explore the development and implementation experiences of these incentive pay initiatives. The lessons learned – from programmes successful in implementation and those that are not – can be grouped into three broad categories.

- *Policy and leadership.* This category represents the extent to which the policies of an education system are aligned and focused on a common vision for teaching and learning. It also addresses the need for leadership, at the school-level and above, that is not volatile and works in concert with a coherent policy framework.
- *Stakeholder engagement.* To design and implement an incentive pay programme well, an education system must, from the start and throughout the life of the programme, engage a broad set of stakeholders, including teachers. This level of engagement should include frequent and transparent communications, involvement in design and decision-making, and opportunities to gather feedback from stakeholders to gauge their experience.
- *Technical capacity.* This category represents the technical supports and infrastructure that must be in place to implement and sustain an incentive pay programme. An education system must have, for example, the human resource systems to evaluate and track teacher performance, data systems to accurately measure performance, along with transparent and accurate payroll systems.

Box 7A.2 Importance of stakeholder engagement

The early and ongoing involvement of a broad set of stakeholders in the design and implementation of an incentive pay programme is paramount. Meaningful and ongoing engagement can be secured through frequent and transparent communications, broad stakeholder involvement in design and decision-making roles, and opportunities to gather feedback from stakeholders to gauge the participant experience (Eckert, 2010; Lewis and Springer, 2009; OECD, 2009). Stakeholders should include representatives from groups such as teachers, teacher unions, school and district leadership, school board members, other school and district staff, and other community members (e.g. business leaders, parents). By involving these multiple perspectives, it is more likely that an incentive pay programme will reach consensus on core objectives and will be accepted upon implementation (OECD, 2009; Raue *et al.*, 2008).

Table 7A.3 provides an overview of the main quality checks described previously and examples of questions and strategies that can be used to monitor those quality controls prior to and during programme implementation.

Table 7A.3**Strategies to monitor quality of incentive pay design and implementation**

Purpose	Questions to Monitor
Leadership support and consistency	Are school, district, and/or state-level education leaders supportive of and committed to the incentive pay programme? Are they “champions” of the programme?
Policy coherence	Does the education system have a comprehensive strategy for improving teacher quality? Does the incentive pay programme align with those goals? Does the incentive pay programme conflict with or detract from other teacher quality initiatives?
Financial support and sustainability	Is there a steady funding stream for the incentive pay programme? Is the education system considering how to reallocate funds to address incentive pay as one component of a broader teacher quality improvement effort?
Communication with stakeholders	How often and through what mechanisms does the education system communicate with stakeholders? What is the target audience and what is the content of the communication material? Are stakeholders reading the material, do they understand it and value it?
Stakeholder awareness and understanding of programme	Are stakeholders – particularly those eligible to receive awards – aware of the programme? Do they understand how the programme operates and particularly how their performance is evaluated?
Stakeholder involvement in design and implementation	What kind of stakeholders are on the design committee or involved in some other active role? What mechanisms are used to engage stakeholders actively in design and implementation of the incentive pay programme? How does the education system gather feedback from stakeholders and what is done with the feedback?
Data identifying who is in the education system	Does the data system have a unique identifier for each student and for each teacher? Does the data system allow for matching of teachers to students? Is there an audit system to check the quality of these data?
Data identifying outcomes for students	Does the data system track outcomes for students at the individual level and is it able to attribute those outcomes back to the matched teacher? Is there an audit system to check the quality of these data?
Award payout system	Is there an award payout system that is accurate and linked to educator evaluation results? Are awards paid out in a manner that is understood by educators and in a timely manner? Does the education system have a grievance procedure for educators who dispute their award payout?

Source: Springer, 2010.

CONCLUSION

As the number of countries and local educational authorities considering strategies for improving student learning and educator effectiveness grows, compensation and teacher incentive initiatives are increasingly assuming a key role in reform agendas. The question of whether to implement incentive pay or not is quickly becoming less prominent than the question of how best to implement an incentive pay programme. Moving forward, it will be more important than ever that states and countries implementing incentive pay programmes include rigorous pilots, evaluations and quality control mechanisms as central components of their work.

Balancing the needs of research, policy and practice can be difficult. While rigorous research and evaluation is critical for improving programme design and implementation, time and resource constraints may not allow for thorough piloting prior to full-scale implementation. When a pilot is not feasible, it is at the very least critical that an education system put in place the mechanisms for ongoing, formative evaluation, quality monitoring and objective summative evaluation to understand and modify, as necessary, the nature and consequences of an incentive pay programme.

Box 7A.3 The Data Quality Campaign

The Data Quality Campaign (DCQ), a multi-partner collaborative effort in the United States to enhance states' data systems in public education, has made a concerted effort to identify data needs and educate states about the strategies to attain (what the campaign deems) the "10 Essential Elements" for a state longitudinal data system. These elements can be broken into three categories: *i)* identifying who is in the public education system; *ii)* understanding outcomes for students; and *iii)* monitoring data system quality.¹⁰

Who is in the public education system? The DCQ recommends that a state's longitudinal data system include unique identifiers for each student and for each teacher. The student identifier connects student data across various databases and captures student enrolment in schools and courses, their demographic information and their programme participation, such as in special education or poverty assistance programmes. The unique identifiers for teachers should match teachers to students by classroom and by subject. The data system should capture teacher information such as education, experience and training, as well as track their movement in and out of schools/districts.

What are outcomes for students? Each state's education data system should also capture information on student outcomes, including performance on assessments, course completion, graduation or drop-out, as well as their pathways following secondary education. More specifically, the DCQ recommends that student assessment results be collected at the student-level and allow the tracking of individual student test scores from year to year to measure academic growth. Similarly, course completion and course grades should be captured at the individual student level over time, as should outcomes for students following secondary education (*i.e.* whether they pursue post-secondary education and of what type). Finally, a quality data system must also identify those students who are enrolled but not tested, so as to understand patterns and types of students that are not assessed over time.

What is the quality of data? The DCQ points out that any state data system should be accompanied by a data audit system to assess the quality, validity and reliability of data integrated into the longitudinal system. Creating such an audit system involves the establishment of data standards, training for local school personnel in charge of providing data, statistical tests to identify errors, intermittent on-site audits of local school systems' data records, and a set of consequences and/or rewards to impress upon local school systems the importance of quality data recording and input.

NOTES

1. Examples, practices and recommendations regarding policies relating to teacher education, selection and training fall beyond the scope of this chapter.
2. In-service teacher incentives based on performance are an innovative approach and experiments, trials and experiences are still being monitored and evaluated. Research is conclusive, however, regarding the different problems with the single-salary schedule (OECD, 2009a).
3. This chapter draws largely on the sister publication of this report *Evaluating and Rewarding the Quality of Teachers* (OECD, 2009), which brings together the most current thinking and practices regarding incentives and stimuli for teachers, based on a review of examples from 23 countries.
4. Teacher Incentive Fund (TIF) administered by the US Department of Education (USDOE) with a USD 400 million budget in 2010 to provide districts funds to implement merit-based pay systems (USDOE, 2010).
5. Muralidharan and Sundararaman (2009) present findings from the Andhra Pradesh Randomised Evaluation Study (APRESt) conducted in India that show that teacher incentives improved results in student learning in both maths and language, after only two years of the programme. Lavy (2004) shows that a one-year pilot in Israel improved student learning and had positive impacts on teacher behaviours such as teaching methods, after-school teaching and increased responsiveness to student needs.
6. Regarding the appropriate amounts for incentives, a review of international programmes shows that individual teacher incentives can range from less than 1% to more than 360% of monthly salary (OECD, 2009), although experts suggest that between 4% and 8% of annual salary can be adequate for incentives to be meaningful but not cause unwanted behaviour.
7. See <http://talenttransferinitiative.org/index.html> to learn more about the Talent Transfer Initiative.
8. Tournaments award prizes not on the basis of an absolute standard but on the basis of relative performance.
9. Lavy used a regression discontinuity design in his studies of the effects of incentive pay in Israel. This design allows for more precise measurements of the effects of an intervention before and after it is implemented.
10. To learn more about the Data Quality Campaign and the 10 Essential Elements, visit www.dataqualitycampaign.org/survey/elements.

References

- Atkinson, A., S. Burgess, B. Croxson, P. Gregg, C. Propper, H. Slater and D. Wilson** (2009), "Evaluating the Impact of Performance-Related Pay for Teachers in England", *Labour Economics*, Vol. 16, pp. 251-261.
- Austin Independent School District (AISD)** (2010), *AISD REACH Supporting and Rewarding Success in the Classroom Year 2: Evaluation Report II 2008-2009*, Austin, TX.
- Battelle for Kids** (2009), *The Importance of Accurately Linking Instruction to Students to Determine Teacher Effectiveness*, Battelle for Kids.
- Boruch, R., D. DeMoya and B. Snyder** (2002), "The Importance of Randomized Field Trials in Education and Related Areas", in F. Mosteller and R. Boruch (eds.), *Evidence Matters: Randomized Trials in Education Research*, Brookings Institution Press, Washington, DC, pp. 50-79.
- Buck, S. and J.P. Greene** (2010), *Blocking, Diluting, and Co-opting Merit Pay*, PEPG Working Paper No. 10-14, Programme on Education Policy and Governance, Cambridge, MA.
- Buddin, R., D.F. McCaffrey, S.N. Kirby and N. Xia** (2007), *Merit Pay for Florida Teachers: Design and Implementation Issues*, Working Paper WR-508-FEA, RAND Education.
- Chamberlin, R., T. Wragg, G. Haynes and C. Wragg** (2002), "Performance-Related Pay and the Teaching Profession: A Review of the Literature", *Research Papers in Education*, Vol. 17, No. 1, pp. 31-49.
- Coates, Y.D.** (2009), *A Focused Analysis of Incentives Affecting Teacher Retention: What might Work and Why?*, American University, Washington, DC.
- Community Training and Assistance Center** (2004), *Catalyst for Change: Pay for Performance in Denver*, CTAC, Boston, MA.
- Cornetto, K.M., L.S. Schmitt, K. Malerba and A. Herrera** (2010), *AISD REACH Year 2 Evaluation Report II, 2008-2009*, Austin Independent School District, Austin, TX.
- Cresap, McCormick and Paget, Inc.** (1984), *Teacher Incentives: A Tool for Effective Management*, Reston: National Association of Elementary School Principals, Washington, DC.
- Darling-Hammond, L.** (1997), *Doing What Matters Most: Investing in Quality Teaching*, National Commission on Teaching & America's Future, New York.
- Delaware Department of Education** (2010), "Race to the Top: Application for Funding", CFDA No. 84395A, Narrative, The State of Delaware.
- Duflo, E., R. Hanna and S. Ryan** (2007), *Monitoring Works: Getting Teachers to Come to School*, Working Paper 11880, National Bureau of Economic Research, Cambridge, MA.
- Eckert, J.** (2010), *Performance-Based Compensation: Design and Implementation at Six Teacher Incentive Fund Sites*, The Bill and Melinda Gates Foundation and The Joyce Foundation.
- Ehrenberg, R. and R. Smith** (1994), *Modern Labor Economics: Theory and Public Policy*, Fifth Edition, Harper Collins, New York.
- Ehrenberg, R.G. and G.T. Milkovich** (1987), "Compensation and Firm Performance", in M. Kleiner *et al.* (eds.), *Human Resources and the Performance of Firms*, Industrial Relations Research Association, Madison, WI.
- Gates Foundation** (2010), *Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching*, Gates Foundation, Seattle, WA.
- Glazerman, S., A. McKie and N. Carey** (2009), *An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year One Impact Report*, Mathematica Policy Research, Inc.

- Glazerman, S.** and **A. Seifullah** (2010), *An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report*, Mathematica Policy Research, Inc.
- Glewwe, P., N. Ilias** and **M. Kremer** (2008), *Teacher Incentives in Developing Countries: Recent Experimental Evidence from Kenya*, Working Paper, Harvard University, Cambridge, MA.
- Goldhaber, D.** (2009), "The Politics of Teacher Pay Reform", in M.G. Springer (ed.), *Performance Incentives: Their Growing Impact on American K-12 Education*, Brookings Institution Press, Washington, DC.
- Gonring, P., P. Teske** and **B. Jupp** (2007), *Pay-for-Performance Teacher Compensation: An Inside View of Denver's ProComp Plan*, Harvard Education Press, Cambridge, MA.
- Hanushek, E.A.** (1992), "The Trade-off between Child Quantity and Quality", *Journal of Political Economy*, Vol. 100, No. 1, pp. 84-117.
- Heinrich, C.** and **G. Marschke** (2007), *Dynamics in Performance Measurement System Design and Implementation*, La Follette School of Public Affairs, University of Wisconsin, Madison, WI.
- Heyburn, S., J. Lewis** and **G. Ritter** (2010), *Compensation Reform and Design Preferences of Teacher Incentive Fund Grantees*, National Center on Performance Incentives, Nashville, TN.
- HISD** (2009), *Houston Independent School District*, Michael & Susan Dell Foundation, Houston, TX.
- Jacob, B.** and **M. Springer** (2007), *Teacher Attitudes on Pay for Performance: A Pilot Study*, Working Paper 2007-06, National Center on Performance Incentives.
- Johnson, J.** (2005), *State Financial Incentive Policies for Recruiting and Retaining Effective New Teachers in Hard-to-Staff Schools*, Education Commission of the States Clearinghouse on Teacher Incentives, Denver, Colorado.
- Koppich, J.** (2008), *Toward a More Comprehensive Model of Teacher Pay*, National Center on Performance Incentives, Working Paper 2008-6.
- Lavy, V.** (2002), "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement", *Journal of Political Economy*, Vol. 110(6), pp. 1286-1317.
- Lavy, V.** (2004), *Performance Pay and Teachers' Effort, Productivity and Grading Ethics*, NBER Working Paper 10622, National Bureau of Economic Research.
- Lawler, E.E. III** (1981), *Pay and Organizational Effectiveness*, McGraw-Hill, New York.
- Lazear, E.** (1996), "Performance Pay and Productivity", NBER Working Paper No. W5672, National Bureau of Economic Research, Cambridge, MA, SSRN: <http://ssrn.com/abstract=225573>, accessed April 2010.
- Lewis, J.** and **M. Springer** (2009), *Effective Technical Assistance Principles: Lessons from Three Performance Pay Programs*, Center for American Progress, Washington, DC.
- Loeb, S., L. Darling-Hammond** and **J. Luczak** (forthcoming), "How Teaching Conditions Predict Teacher Turnover in California Schools", *Peabody Journal of Education*.
- Malen, B.** (1999), "On Rewards, Punishments, and Possibilities: Teacher Compensation as an Instrument for Education Reform", *Journal of Personnel Evaluation in Education*, Vol. 12, No. 4, pp. 387-394.
- McEwan, P.J.** (2008), "Quantitative Research Methods in Education Finance and Policy", in H.F. Ladd and E.B. Fiske (eds.), *Handbook of Research in Education Finance and Policy*, Routledge, New York, pp. 87-104.
- McKinsey & Company** (2007), *How the World's Best-Performing School Systems Come Out on Top*, Identifying Teacher Quality Project, Washington, DC.
- Milanowski, A.T.** (2007), "Performance Pay System Preferences of Students Preparing to be Teachers", *Education Finance and Policy*, Vol. 2(2), pp. 111-132.
- Milgrom, P.** and **J. Roberts** (1992), *Economics, Organization & Management*, Prentice Hall, Englewood Cliffs, CA.

- Muralidharan, K.** and **V. Sundararaman** (2008), *Teacher Incentives in Developing Countries: Experimental Evidence from India*, Working Paper 2008-13, National Center on Performance Incentives.
- Muralidharan, K.** and **V. Sundararaman** (2009), *Teacher Performance Pay: Experimental Evidence from India*, Working Paper 15323, National Bureau of Economic Research, retrieved 15 December 2009, from www.nber.org/papers/w15323.pdf.
- Odden, A.** and **C. Kelley** (1997), *Paying Teachers for What They Know and Do: New and Smarter Compensation Strategies to Improve Schools*, Corwin Press, Thousand Oaks, CA.
- Organisation for Economic Co-operation and Development (OECD)** (2005), *Teachers Matter: Attracting, Developing and Retaining Effective Teachers*, OECD Publishing, Paris.
- OECD** (2007), *Education at a Glance 2007: OECD Indicators*, OECD Publishing, Paris.
- OECD** (2009a), *Evaluating and Rewarding the Quality of Teachers: International Practices*, OECD Publishing, Paris.
- OECD** (2009b), *Creating Effective Teaching and Learning Environments: First Results from TALIS*, OECD Publishing, Paris.
- OECD** (2010a), *Education at a Glance 2010: OECD Indicators*, OECD Publishing, Paris, accessed at www.oecd.org/document/52/0,3343,en_2649_39263238_45897844_1_1_1_1,00.html.
- OECD** (2010b), *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas*, OECD Publishing, Paris.
- OECD Steering Group on Evaluation and Teacher Incentive Policies** (2009), *Preliminary Advice on the Design of State-Level Pilots of Teacher Performance and Incentive Mechanisms*, OECD Publishing, Paris.
- Patton, M.Q.** (2002), *Qualitative Research and Evaluation Methods*, Third Edition, SAGE Publications, Thousand Oaks, CA.
- Pfeffer, J.** and **N. Langston** (1993), "The Effect of Wage Dispersion on Satisfaction, Productivity, and Working Collaboratively: Evidence from College and University Faculty", in *Administrative Science Quarterly*, Vol. 38, pp. 382-407.
- Podgursky, M.** and **M. Springer** (2007), "Teacher Performance Pay: A Review", *Journal of Policy Analysis and Management*, Vol. 26, No. 4, pp. 909-949.
- Raue, K., K. MacAllum** and **L. Ristow** (2008), *Building Systems to Recognize Teachers of Excellence: Lessons from the Ohio Teacher Incentive Fund*, Westat, Rockville, MD.
- Raue, K., K. MacAllum, A. Winkler** and **L. Ristow** (2008), *Different Designs, Common Paths: A First Look at the Ohio Teacher Incentive Fund*, Issue Paper, Westat's Education Studies Group.
- Rogers, H.** (2009), "Piloting and Evaluating Teacher Incentive Programs", PowerPoint presented at the OECD Mexico Joint Workshop on Education Quality Standards and Assessment, Mexico City, July.
- Rogers, F.H.** and **E. Vegas** (2009), *No More Cutting Class? Reducing Teacher Absence and Providing Incentives for Performance*, World Bank Policy Research Working Paper No. 4847, World Bank, Washington, DC.
- Rose, S.** (2010), *Pay for Performance Proposals in Race to the Top Round II Applications*, Briefing Memo, Education Commission of the States.
- Rossi, P.H., M.W. Lipsey** and **H.E. Freeman** (2004), *Evaluation: A Systematic Approach*, Seventh Edition, SAGE Publications, Thousand Oaks, CA.
- Roza, M.** and **R. Miller** (2009), *Separation of Degrees: State-by-State Analysis of Teacher Compensation for Master's Degrees*, Center on Reinventing Public Education Rapid Response Brief, Seattle, WA.
- Sanders, W.L.** and **J.C. Rivers** (1996), *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement*, Value-Added Research and Assessment Center, University of Tennessee, Knoxville, Tenn.

- Sclafani, S.** and **M. Tucker** (2006), "Teacher and Principal Compensation: An International Review", Center for American Progress, Washington, DC, www.americanprogress.org/issues/2006/10/teacher_compensation.html/pdf/education_report.pdf, accessed May 2009.
- Shavelson, R.J.** and **L. Towne** (eds.) (2002), *Scientific Research in Education*, National Academy Press, Washington, DC.
- Skinner, B.** (1981), "Selection by Consequences", *Science*, Vol. 213, pp. 501-514.
- Slotnick, W.J.** (2009), *It's More than Money: Making Performance-Based Compensation Work*, Center for American Progress, Washington, DC.
- Snowden, J.** (2007), *The Future of Teacher Compensation: Déjà Vu or Something New?*, Center for American Progress, Washington, DC.
- Springer, M.** (2010), *Performance Incentives: Their Growing Impact on K-12 Public Education*, Brookings Institution Press, Washington, DC.
- Springer, M.** and **M. Winters** (2009), *New York City's School-Wide Bonus Pay Program: Early Evidence from a Randomized Trial*, Working Paper 2009-02, National Center on Performance Incentives.
- Springer, M., J. Lewis, M. Podgursky, M. Ehlert, L. Taylor, O. Lopez and A. Peng** (2009), *Governor's Educator Excellence Grant (GEEG) Program Year Two Evaluation Report*, National Center on Performance Incentives, Nashville, TN.
- Springer, M., J. Lewis, M. Podgursky, M. Ehlert, L. Taylor, O. Lopez and A. Peng** (2009), *Governor's Educator Excellence Grant (GEEG) Program Year Three Evaluation Report*, National Center on Performance Incentives, Nashville, TN.
- Springer, M., M. Podgursky, J. Lewis, M. Ehlert, T. Gronberg, L. Hamilton, D. Jansen, O. Lopez, A. Peng, B. Stecher and L. Taylor** (2009), *Texas Educator Excellence Grant (TEEG) Program: Year Two Evaluation Report*, National Center on Performance Incentives, Nashville, TN.
- Taylor, L., M. Springer and M. Ehlert** (2009), "Characteristics and Determinants of Teacher-Designed Pay for Performance Plans: Evidence From Texas' Governor's Educator Excellence Grant Program", in M.G. Springer (ed.) (2010), *Performance Incentives: Their Growing Impact on American K-12 Education*, Brookings Institution Press, Washington, DC.
- Tukey, J.W.** (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.
- United States Department of Education** (2010), *Application for the Teacher Incentive Fund*, CDFA No. 84.385, United States Department of Education, Washington, DC.
- Vegas, E.** and **I. Umansky** (2005), *Improving Learning through Effective Incentives: What Can We Learn From Education Reforms in Latin America?*, World Bank, Washington, DC.
- Watson, J.G., S.B. Kraemer and C.A. Thorn** (2009), *Data Quality Essentials: Guide to Implementation, Resources for Applied Practice*, Center for Educator Compensation Reform.
- Zenger, T.** and **C. Marshall** (1995), "Does Size Matter in Group Rewards? Factors Affecting the Incentive Intensity and Performance of Group-Based Pay Plans", paper presented to the annual meeting of the Academy of Management, Vancouver, BC.

Conclusion

This report presents the main findings and recommendations of the Steering Group on Evaluation and Teacher Incentive Policies resulting from the Co-operation Agreement between the government of Mexico, represented by SEP (*Secretaría de Educación Pública*), and the OECD. The report forms part of a larger body of work: three related publications,¹ expert papers, workshops and technical meetings with officials from SEP and relevant stakeholders. Much of the work was conducted jointly with SEP officials and on several occasions, the Education Minister was personally involved in meetings and workshops.

Rather than taking a thematic or country-specific approach, the structure and content of the report reflects the interconnected nature and systemic perspective inherent in large-scale reform efforts. The Steering Group focused on developing effective and timely policies for in-service teacher incentives, which naturally led them to focus on teacher evaluation frameworks and the development of teaching standards. Drawing on policy analysis and reform processes from several countries, the Steering Group developed a unifying framework to analyse and consider the various policy priorities and issues. Chapter 2 of the report focuses on how national governments, in this case SEP, can continue reform efforts in a systematic, methodical manner as a follow-up to the OECD's key recommendations, through the *Public Policy Framework for Implementing Education Reform*.

With regard to the Public Policy Framework, a distinction is drawn between the phases of policy design, planning, implementation and evaluation. The report stresses the importance of local knowledge to complement OECD research, evidence and international practices. This implies gathering the necessary information at the local, national and international levels to make informed decisions about policy design, taking local constraints and opportunities into consideration. The aim of the Policy Framework is to guide the analytical process to ensure a systematic and thorough consideration of the policy options by the relevant actors. This will be crucial in adapting best practices to make them appropriate to local contexts. The biggest challenge is to bring together different types of knowledge – scientific, technical, professional, administrative, political and pedagogical – at the appropriate times. It is clear that the reform process in Mexico will require repeated and sustained efforts.

While accountability is at the centre of the policy debate on large-scale reforms in many countries, the concept of holding all stakeholders directly responsible for outcomes is seldom raised in Mexico. Discussion in this area revolves around the relationships between different actors at different levels in order to determine who is accountable to whom and for what. In Mexico, accountability, and the standards that accompany it, could form the basic building blocks of sustained reform efforts to improve student achievement, school performance and teacher practices.

The chapter on accountability reviews some of the mechanisms available to Mexico and other countries attempting system-level reforms to improve student achievement. The chapter suggests that accountability based on establishing clear expectations as reflected in clear and ambitious standards can be an effective instrument for public accountability. The report suggests that, ideally, multiple, cross-referenced, valid and reliable measures of performance should also be used, including student assessment data. The report therefore provides an in-depth review of the strengths, weaknesses and opportunities for further development of Mexico's ENLACE² assessment in Chapter 4.

Throughout this publication, a continuous thread emerges: student learning and growth over time are key criteria against which all actors in the education system should be held accountable. Given the importance of teachers and schools, a basic unit of accountability (although not the only one) should be the school.

Every student should have access to high quality relevant learning opportunities. It is clear that the policy driver for improvement is student learning which requires commitment from individual teachers and shared social responsibility from stakeholders.

Measuring student learning is not an easy task and the report recognises the need to develop more instruments, to compile more evidence, and to validate the results of different approaches. Chapter 4 suggests that to complement standardised or school-based assessments, other types of evaluations should also be considered. One such option is reviewed in the chapter on assessing the value-added of schools, a topic in which countries have shown increased interest in recent years. Proper modelling using value-added methods can produce measures of student growth that are more precise and ultimately fairer than those produced through comparisons of raw score averages. The challenges, however, are considerable, and the chapter outlines the key issues associated with value-added modelling.

The general view reflected in the report is that the education system as a whole (authorities, schools, teachers, parents and students) can be held accountable for learning. Given the importance of teachers, Chapter 6 presents guidelines for the development of a comprehensive, fair and transparent in-service teacher evaluation system. This proposal rests solidly on the development of clear standards of good teaching practices.³ Although Mexico has previous experience in teacher assessment, there is growing awareness of the need to further develop these experiences into a sound teacher evaluation system. The challenge lies in making the transition from the previous system to a new effective one.

Closely linked to such an evaluation process for teachers is the issue of teacher incentives. The chapter on incentives for in-service teachers reviews several international examples, identifying best practices in certain areas, including the benefits of pilots and how to conduct proper monitoring and evaluation of incentive programmes and policies (see Appendix 7A).

The analysis, findings and recommendations provided in this report should foster the processes required for country-specific implementation involving additional analysis and consensus-building at the federal, state and local levels. The combination of international practices and local and specific policy constraints in Mexico has made the work underlying this report both challenging and rewarding. The Co-operation Agreement has benefited from the OECD's stock of knowledge, and certain strategies for effective engagement are beginning to become clear. It is these kinds of insights, coupled with a comparative international perspective that the OECD will continue to put at the service of Mexico and other member and partner countries.

NOTES

1. The two sister publications are *Improving Schools: Strategies for Action in Mexico* (OECD, 2010) and *Evaluating and Rewarding the Quality of Teachers: International Practices* (OECD, 2009). Valuable material on value-added modelling was updated and translated into Spanish for the Co-operation Agreement: *La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas* (OECD, 2010). Working papers from experts commissioned by the OECD are available at www.oecd.org/edu/calidadeducativa.org.

2. *Evaluación Nacional del Logro Académico en Centros Escolares*.

3. The sister publication, *Improving Schools: Strategies for Action in Mexico* (OECD, 2010), presents an analysis on teacher selection, and initial and continuous training for Mexico.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Commission takes part in the work of the OECD.

OECD Publishing disseminates widely the results of the Organisation's statistics gathering and research on economic, social and environmental issues, as well as the conventions, guidelines and standards agreed by its members.

Establishing a Framework for Evaluation and Teacher Incentives

CONSIDERATIONS FOR MEXICO

In this era of knowledge-based economies and changing demographics, all educational systems must improve their learning outcomes and often also deliver more with less.

In order to assist Mexico and other countries in addressing this challenge, this report provides advice for designing, planning, implementing and evaluating policies and practices on educational assessment, standards and evaluation, drawing on the world's best available expertise. Considering that the quality of educational outcomes cannot exceed the quality of its teachers, the report puts particular emphasis on evaluating and recognising teachers.

Effective implementation of educational reforms can, however, prove challenging. Merely knowing what policy levers to apply is not enough. Governments also need to determine the “how” of effective policy design and implementation. The report therefore also provides advice for policy makers to analyse and adapt best practices to make them appropriate in local contexts.

This summary report presents the main findings and policy recommendations developed by the OECD Steering Group on Evaluation and Teacher Incentive Policies, consisting of international experts.

Further reading:

Improving Schools: Strategies for Action in Mexico (OECD, 2010)

Evaluating and Rewarding the Quality of Teachers: International Practices (OECD, 2009)

La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas (OECD, 2010)

This publication is a product of the co-operation agreement established between the government of Mexico and the OECD, which aims to improve the quality of education in Mexico.

Please cite this publication as:

OECD (2011), *Establishing a Framework for Evaluation and Teacher Incentives: Considerations for Mexico*, OECD Publishing.

<http://dx.doi.org/10.1787/9789264094406-en>

This work is published on the *OECD iLibrary*, which gathers all OECD books, periodicals and statistical databases. Visit www.oecd-ilibrary.org and do not hesitate to contact us for more information.

2011

OECD publishing
www.oecd.org/publishing

ISBN 978-92-64-09439-0
87 2010 03 1P

