QUALITY ENHANCEMENT OPTIONS FOR THE SCHOOLING SYSTEM

Learning from improvements in similar systems elsewhere

29 November 2009

Executive summary

The aim of the paper is twofold. Firstly, it uses results from international assessment programmes to examine systematically the degree to which countries, in particular developing ones, have seen improvements in learner performance and may therefore offer clues to how to bring about such improvements. Secondly, the paper zooms in on a few developing countries which have taken measures to improve learner performance where these measures seem interesting from a South African perspective.

Following an outline of the scope of the paper in section 1, section 2 provides an analysis of the data from the international assessment programmes. Specifically, data from the period 1999 to 2007 from the PISA, PIRLS and TIMSS programmes are analysed. It is found that different programmes can display contradictory trends, and that for a number of technical reasons certain trends cannot be regarded as reliable. Six developing countries are found to have experienced substantial educational quality improvements in the 1999 to 2007 period: Ghana, Brazil, Philippines, Chile, Mexico and Indonesia. During this period, the little data on South Africa in the international programmes suggest that South Africa experienced a small decline in its performance (based on 1999 and 2003 TIMSS results).

In section 3, quality enhancement strategies from five countries are examined.

- **Brazil** is examined as it is in the list of countries with improvements, and displays policies that seem innovative. Brazil's twin programmes SAEB (which assesses learners in a representative sample of schools) and Prova Brasil (which assesses learners in non-sampled schools) are found to be a good example of a two-tier approach that maximises both analytical rigour (through the sample-based SAEB) and extent of coverage (through Prova Brasil) simultaneously. Brazil has moreover developed systemic targets based on a simple two-part composite indicator that combines learner performance scores and an enrolment indicator.

- **Chile**, like Brazil, is examined because it has displayed measureable learner performance improvements as well as innovative programme design. Chile's SNED teacher rewards programme is found to be an innovative way of providing non-perpetual financial incentives for all the teachers within a school, based on how well the school does in the national assessment programme. The socio-economic context of schools is taken into account by letting schools compete only against schools in the same socio-economic category. Chile's AVDI programme, which essentially pays teachers a non-perpetual bonus for passing an examination on their subject knowledge and general knowledge of pedagogics, is also discussed.

- Although **Uruguay** is not on the list of improving countries from section 2, this country's key quality enhancement programme is examined partly because it displays elements that seem especially relevant to South Africa, and partly because Uruguay's own national assessments display strong performance improvements at the primary school level. Above all, Uruguay has succeeded in involving teacher unions in innovative ways in the design and governance of its learner assessment programme, and hence, it is argued, improved teacher commitment to improving learning outcomes. Uruguay has moreover used the results from its assessments to inform the focus of in-service teacher training programmes.

- **Mexico** is examined because an in-depth and government-sponsored evaluation on a widely publicised teacher rewards programme, Carrera Magisterial, suggests that Mexico has made important mistakes in its policy design. The fact that learner results have improved is probably more despite the teacher reward programme than as a result of it. The Mexican programme is thus put forward as an example of what not to do.

- **Peru** is examined very briefly on the strength of an intervention that used the mass media to advocate the use of simple tests of the ability of learners to read texts out loud. This, it is argued, has deepened social awareness of educational quality.

The exclusive focus on Latin American countries is not deliberate, and was determined largely by the availability of sources. Specifically, it was difficult to obtain texts on the details of quality enhancement programmes in Ghana, Philippines and Indonesia, three countries from other regions that appeared on the section 2 list of countries with improvements.

Section 4 provides a discussion of what South Africa can learn from the six countries examined in depth. Particularly important lessons seem to be the following (the reader who is not able to read the entire paper is encouraged to at least read the whole of section 4):

- The governance of a national assessment system should be set up in such a way that, firstly, a range of stakeholders, in particular trade unions, take ownership of the system and, secondly, that technical expertise from a range of organisations is harnessed.

- The pressure to take sample-based assessment systems (such as South Africa's Systemic Evaluation) to scale, by assessing all learners in specific grades, should not result in a loss of reliability and breadth of scope of the data gathered. Brazil and Uruguay's two-tier approaches provide interesting ways of dealing with this challenge.

- In any assessment system, there should be an appropriate balance of data collection and data use. Brazil's approach of collecting data every second year appears to offer the breathing space needed for appropriate absorption of results, and action based on the results.

- An effective communication strategy that indicates to the different roleplayers, including parents, in sufficiently straightforward ways how the assessment programme works and what its benefits are, is vital.

Finally, section 5 offers some concluding remarks.

1 Introduction

This paper is part of a series of additional papers accompanying a UNICEF-funded school finance and management survey[1]. This additional paper has a particularly wide scope insofar as it goes beyond the narrower school funding concerns, and examines what policy solutions from countries other than South Africa have succeeded in improving the translation of education inputs, such as funding and teaching posts, into outputs, in particular learner results. Moreover, the present paper considers what South Africa could learn from these experiences. The brief for the study is as follows.

> *A few quality enhancement policy strategies*. There is much interest in improving the quality of South Africa's schooling, and it is fairly common to view better non-personnel resourcing as one solution. Much of the challenge lies in combining the various resourcing, pedagogical, labour relations and organisational initiatives into a coherent over-arching strategy that can easily be communicated to the nation, and can inspire the education sector. A number of countries offer valuable lessons in what to do, and what to avoid. Policy approaches from a few countries (in particular developing countries) that have made serious attempts to raise the quality of schooling would be studied with a view to arriving at a set of lessons learnt that are of relevance to South Africa.

In any cross-country learning exercise one needs to be cautious to bear in mind contextual information as policy solutions that work in one context, may not work in another. In this regard, South Africa's economic position as a middle income country is important. There are a number of reasons why solutions employed in low income or high income countries may not be workable in South Africa. Moreover, the fact that South Africa is a democracy, and has strong teacher unions, is a key contextual factor. The decentralised nature of South Africa's schooling system (arising out of provincial competencies but also powers granted to schools) means that national policies must be constructed very differently to those of a country that is centralised. Finally, South Africa's history of apartheid and high levels of prevailing inequality create fairly (but not entirely) unique circumstances. Having said this, cross-country learning is clearly an indispensable part of the education policy dialogue in any country. Educational improvement is too complex a process to ignore the successes and failures (often very costly ones) experienced in other countries. In this paper we aim to draw lessons from other countries whilst exercising due caution with respect to contextual factors.

The focus of the paper is largely on national policy solutions where these might have resulted in verifiable improvements in learner performance. The tendency is therefore to steer away from policy solutions in other countries that seem to promise much, but are not accompanied by a verifiable improvement in learner performance (a few exceptions are made to this rule, however, where it was thoughts this was justified). Furthermore, the point of departure is not the literature on school-level improvement. The point of departure is rather the literature dealing with the design of national policies aimed at bringing about school-level improvements.

The paper leans towards the practical, and the specifics of policy design, and is not a review of the more theoretical or synoptic literature on what improves educational quality in a country. To take just one example of such literature, Hanushek and Woessman (2007) identify three key conditions for quality improvements to occur in a schooling system. These are (1) choice and competition with respect to where learners attend school (presumably this would apply largely in urban areas), (2) a degree of autonomy for school managers to deal innovatively with local problems, and (3) accountability of schools to the administration and to parents with respect to what children learn. Arguably, South Africa's education policies

[1] The full title of the project, begun in 2008, is 'Evaluation of the implementation and impact of the national norms and standards for school funding and the development of a monitoring and evaluation framework and strategy'.

have gone further with respect to the first two conditions, than with respect to the third one. Even if these kinds of synthetic accounts of what works are not reviewed in this paper, what this more global literature is saying does inform the country-specific analyses.

Finally, an attempt is made to base the selection of good practice elsewhere, and determination of policy implications for South Africa, on what already exists in the education policy domain within South Africa. Clearly, re-aligning existing policies is in many ways less costly than replacing existing policies with completely different policies.

In section 2 data from three different international testing programmes are examined to identify which countries could offer lessons for South Africa. In section 3 policy solutions are discussed, country by country, with reference to five countries. The fact that all five countries are Latin American countries is not the result of a deliberate choice. It turned out that the literature on the programmes in these countries was more accessible (texts in English, Spanish or Portuguese were accessed) than for other developing countries, and Latin American countries also happen to display many parallels with South Africa, in particular as far as the prominence of teacher unions is concerned. Finally, in section 4 how South Africa could learn from the programmes and policies presented is discussed. The conclusion in section **Error! Reference source not found.** includes some practical pointers for the way forward.

2 The international improvement patterns

In this section countries are identified that have displayed noticeable improvements in either mathematics or language (two key learning areas in the South African policy debates) during the last decade. Data from the following five time series are considered:

Programme	*Subject*	*Level*	*Years*
PISA	Mathematics	Youths aged 15	2000, 2003, 2006
PISA	Language	Youths aged 15	2000, 2003, 2006
PIRLS	Language	Grade 4	2001, 2006
TIMSS	Mathematics	Grade 4	1999, 2003, 2007
TIMSS	Mathematics	Grade 8	2003, 2007

PISA, PIRLS and TIMSS are all international programmes that test samples of learners from several countries using either identical tests or strictly comparable tests, and according to rigorous sampling and data collection standards. South Africa has participated in PIRLS, in 2006, and TIMSS, in 1999 and 2003. Only TIMSS, in the case of South Africa, would offer a time series. The country average scores considered in the analysis of improvements are given in Table 1 in the Appendix.

The five international time series analysed represent the key existing ones. As will be shown, it is important look at a variety of time series, and not just one, as they can provide contradictory information. Specifically, one sees improvements in certain programmes (or subjects) occurring at the same time as deterioration in others. In some cases these contradictions seem impossible to explain, yet they are obviously important to bear in mind when determining what countries could point towards good education policy practices.

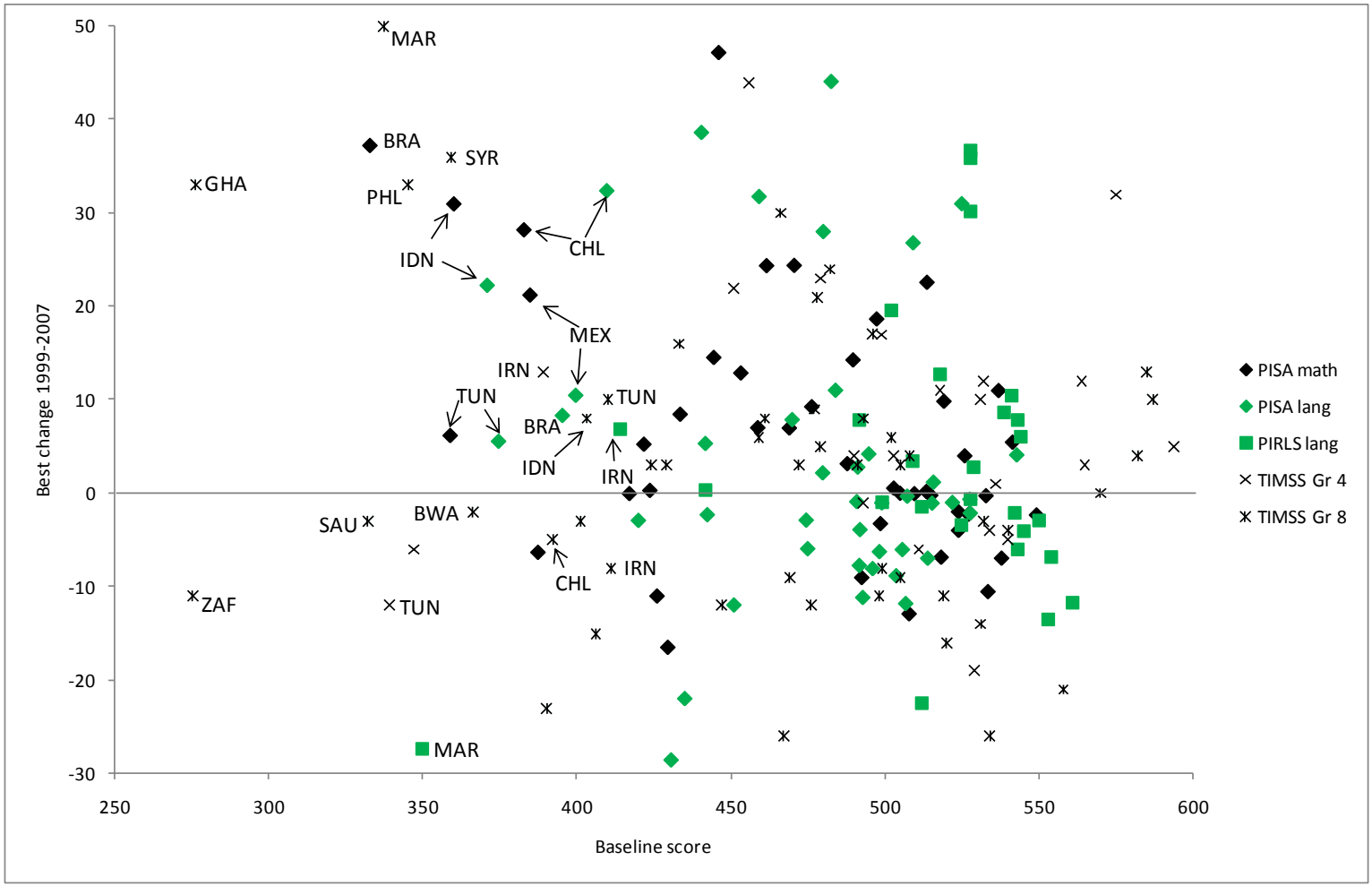
PISA and TIMSS science scores were not included in the analysis. Nor was the 1995 to 2000 SACMEQ series, given difficulties around obtaining 1995 data in this programme (however, the 2000 to 2007 SACMEQ series should provide vital information on improvements in the Southern and Eastern Africa region once the 2007 data become available).

Figure 1 below illustrates the improvement patterns. Each point represents the best improvement achieved by a country within a specific time series (or the least bad

deterioration, if there was no improvement). To take an example, Morocco (MAR) participated in TIMSS Grade 8 mathematics for all three years (1999, 2003 and 2007) and here the 1999 to 2003 improvement was the largest in terms of TIMSS score points. In 2007 Morocco did worse than it did in 2003 – see Table 1 for details. The horizontal axis in Figure 1 indicates the score before the improvement. In the case of Morocco (MAR) this would be the TIMSS Grade 8 score in 1999. The vertical axis indicates the size of the improvement, in the case of Morocco the difference between the 2003 and 1999 scores. Points below the zero improvement line indicate cases where the best (or only) change observed in one of the five time series was a deterioration. As an example, South Africa (ZAF) experienced a deterioration in its TIMSS Grade 8 mathematics average between 1999 and 2003 of 11 points (from 275 to 264). As the emphasis was on identifying improvements amongst countries that had relatively low scores, only countries on the left-hand side of the graph are labelled (using ISO country codes).

It should be emphasised that the scores across the five time series are only roughly comparable. In each time series, the published scores are set so that the unweighted mean across countries is 500, and the standard deviation 100, in some base year. However, as the countries included in the base year group vary from programme to programme, we cannot say, for instance, that a score of 500 in TIMSS Grade 8 mathematics indicates the same standard as a score of 500 in PISA mathematics. Yet the scales are sufficiently similar for our purpose, namely to isolate those countries with levels of performance relatively close to those of South Africa.

Figure 1: Improvements recorded in the international programmes



Sources: The international reports described on the following page.

Countries on the left of Figure 1 which were above the zero no-improvement line were considered to be potential indicators of good practice. One should bear in mind that changes of less than about 3 points tend to be statistically insignificant. What was also taken into account is the fact that the official reports indicated in some cases that there were reliability problems with a country's data, or that certain improvements were suspect due to comparability problems. The conclusion was that four countries which appear interesting in Figure 1 should be ignored, but that six of them seemed to merit further investigation. The four that were ignored are the following (all are from the Middle East and North Africa region):

- **Morocco (MAR)**. The apparent 1999 to 2003 improvement is suspect to due to problems with insufficient returns from schools[2]. Morocco's language performance in PIRLS was not good – there was a decline from 350 to 323 in the 2001 to 2006 period.

- **Syria (SYR)**. The outstanding improvement in this country with respect to TIMSS Grade 8 mathematics (359 to 395 between 2003 and 2007) could be unreliable given that Syria's 2003 results were not published in the official TIMSS 2003 report (though data for Syria do exist in the 2003 dataset).

- **Iran (IRN)**. Despite some improvement according to the graph with respect to PIRLS and TIMSS mathematics Grade 4 (though there appears to be a deterioration with respect to TIMSS mathematics Grade 8), the TIMSS 2007 report[3] and the PIRLS 2006 report[4] both question whether this improvement is real.

- **Tunisia (TUN)**. Though the graph reveals a moderate TIMSS Grade 8 mathematics improvement, the trend has been an unstable one, as indicated in Table 1, with the earliest 1999 level being the best in the series. The TIMSS mathematics Grade 4 trend has been downward. The PISA 2006 report does not regard the mathematics and language improvements as statistically significant[5].

The six countries that appear to merit further investigation are the following:

- **Ghana (GHA)**. The only time series we have for Ghana indicates a remarkable improvement, from 276 to 309 in TIMSS Grade 8 mathematics between 2003 and 2007. This is the largest mathematics Grade 8 improvement for all TIMSS countries in the period 2003 to 2007[6].

- **Brazil (BRA)**. PISA indicates that there has been a large improvement in mathematics (from 356 to 370 between 2003 and 2006). The PISA improvement in language is however questionable[7]. Here the mean score went from 395 to 403 and then back to 393 over the three years.

- **Philippines (PHL)**. Here the one available series, TIMSS Grade 8 mathematics, indicates a marked improvement, from 345 in 1999 to 378 in 2003 (it is only for these two years that we have data).

- **Chile (CHL)**. Despite no clear improvement in its TIMSS mathematics Grade 8 results, Chile displayed clear improvements in its PISA results with respect to language (410 to

[2] See Mullis, Martin, Gonzalez, and Chrostowski (2004: 35). Henceforth this report is referred to as the TIMSS 2003 report.
[3] Mullis, Martin and Foy (2008). Henceforth this is referred to as the TIMSS 2007 report.
[4] OECD (2007). Henceforth this is referred to as the PISA 2006 report.
[5] PISA 2006 report (301, 319).
[6] TIMSS 2007 report (48).
[7] PISA 2006 report (301).

442 between 2000 and 2006). The PISA mathematics improvement seen in the graph should be ignored as 2000 PISA mathematics results are said to be incomparable with 2003 or 2006 results[8]. This applies to all PISA countries.

- **Mexico (MEX)**. Here the PISA mathematics improvement from 385 in 2003 to 406 in 2006 is noteworthy and statistically significant, according to the PISA 2006 report. The PISA language improvement is however questionable as the mean score dropped between 2000 and 2003, and then increased between 2003 and 2006 (without reaching the original 2000 level).

- **Indonesia (IDN)**. The TIMSS Grade 8 mathematics improvement seen on the graph is difficult to interpret because Indonesia's score went up and then down across the three TIMSS years. However, the PISA improvements for mathematics and language are substantial and are considered statistically significant by in the PISA 2006 report (301, 319). In fact, the 2000 to 2006 improvement in mathematics is the largest of all the PISA countries with data for these years.

3 Interventions to improve quality

In this section policies and programmes that could underlie the quality improvements seen in the previous section are examined. The examination is relatively cursory, though where possible reference is made to rigorous in-depth studies of policy effects. One should not expect strong empirical proof of the linkages between certain policy interventions and quality improvements. Even in-depth studies rarely arrive at such proof. To a large degree, one must depend on informed judgement as to what the possible causal linkages are, and this obviously implies that any conclusions must be treated cautiously.

One needs to keep in mind that improvements are likely not to be the result of just one policy intervention, but rather of a combination of interventions in an environment in which a general political and social commitment towards quality education exists. When interventions and programmes are singled out for discussion in the sub-sections that follow, it is partly because it seems likely that they explain an important part (but not the whole) of the improvement observed earlier, but also partly because the interventions and programmes concerned seem of relevance to the South African situation.

The following sub-sections zoom in on key programmes or interventions in five countries: Brazil, Chile, Uruguay, Mexico and Peru. No sub-section offers a comprehensive overview of how educational quality is dealt with in the country concerned. The focus in each case is on one programme or intervention with a design that seemed particularly informative for the South African policy discourse. Uruguay and Peru are of course not in the list of improving countries in the previous section. However, Uruguay was included partly due to evidence within a national programme of substantial improvements, and Peru is included, but through a very brief discussion, as this country implemented an interesting reading campaign, which to some degree has been tried out in South Africa (the education economist involved in the Peru initiative also works intermittently in South Africa).

The use of Spanish and Portuguese texts obviously added value to the analysis, yet there are a number of texts in English relating to the programmes and interventions described.

[8] See PISA 2006 report (319).

3.1 Brazil: A two-tier approach to standardised assessment data

Brazil offers an example of how broad coverage and rigorous quality controls can be reconciled within a national assessment system at the basic schooling level, through what can be described as a two-tier approach. The two tiers are the following:

- **SAEB.** This programme, the full name of which can be translated as National Basic Education Evaluation System, is a sample-based testing programme aimed at detecting aggregate trends in the quality of learner performance over time. It is similar to South Africa's Systemic Evaluation programme.

- **Prova Brasil.** This programme, which can be translated as Assess Brazil, borrows tools from its sister programme SAEB and is designed to encompass all public schools across Brazil.

Both of the above programmes are run by INEP, or the National Institute of Educational Studies and Research. INEP is a government body attached to the Ministry of Education, with an area of responsibility delineated through legislation. The description of SAEB and Prova Brasil which follows is based on information obtained from INEP's website (http://www.inep.gov.br/), plus a number of non-INEP studies. Unfortunately, the information on Brazil was less available in English than the information on the other countries dealt with (only one of the Brazil studies is in English), so a special attempt was made to provide as comprehensive a picture as possible in the discussion that follows.

Brazil, a country plagued by social and income inequalities similar to those of South Africa, was until recently not known as a successful country from the perspective of basic education (Schwartzman (2003) provides some background information on Brazil's schooling system, in English). Recent developments, for instance with respect to SAEB and Prova Brasil, have placed Brazil at the forefront of developing countries when it comes to the design of programmes aimed at improving educational quality. Nonetheless, Brazil's challenges in this area remain great. Educational quality is not what it should be, and a complicated system of governance, whereby public schools can be run by municipalities, a state (of which there are 27 in Brazil if one includes the capital territory Brasilia), or even the federal government, makes change difficult to bring about.

First some background issues are discussed, before the details of the two programmes receive attention.

INEP, established in 1938, has a long history. In the late 1980s the institute began looking at the use of sample-based standardised assessments, following the introduction of such assessments in the United States in the 1960s (specifically, the National Assessment of Educational Progress, or NAEP) and in Europe in the 1970s. In Latin America, the first country to introduce these assessments was Cuba, in 1975. Other Latin American countries mostly followed in the 1990s (Horta Neto, 2006). In Brazil, assessment approaches piloted in the north east of the country, starting in 1980 and funded by the World Bank, influenced subsequent work considerably. It is out of this World Bank-funded work that Harbison and Hanushek's (1992) widely used book *Educational performance of the poor* emerged. SAEB was proposed in Brazil's 1993 ten year plan for education. This plan specified that within five years all states should have the necessary information, statistical, planning and evaluation capacity to run sample-based assessments and a proper school census, the two being inter-dependent insofar as sample-based assessments require a reliable census dataset from which to draw the sample. In 1990 SAEB was run across all states for the first time. In 1994 legislation was introduced specifying the following:

- Monitoring of educational quality through SAEB should be permanent and properly institutionalised.

- SAEB should be the outcome of a collaborative process involving not just the federal Ministry, but also other government organs, universities and research organisations.

- A technical committee should be formed to advise the Minister on how to take SAEB forward.

- SAEB results should be disseminated with a view to fomenting a culture of evaluation in the schooling system, and promoting society's ownership of the programme and its findings.

1995 saw major methodological improvements to SAEB. For instance, for the construction of the learner assessments item banks of questions were developed. Moreover, private schools were included for the first time. The increased complexity of the programme saw INEP outsource much of the work to two research organisations, both NGOs with a long history in schools-based assessments. Many states criticised, firstly, increased centralisation in the control of the programme and, secondly, the fact that non-government organisations were given too much control. The second criticism resulted in greater involvement by INEP in the programme after 1995, though it remained rather centrally controlled.

A national plan produced in 2001 said that within five years, all states should be assessing the performance of learners in all schools with an enrolment of at least 50 pupils. The plan furthermore indicated that the federal government should provide financial support for this. The ambitious goal was to a large degree realised four years later, in 2005, not through expanding SAEB, but through the introduction of the parallel programme Prova Brasil.

The SAEB and Prova Brasil duo has reached relative maturity since 2005. The programmes are run concurrently, every two years, the most recent runs being 2005, 2007 and 2009. Each run involves the following:

- **Preparation of tests and questionnaires.** Tests are prepared for two subjects, Portuguese language and mathematics, and for three grades, Grades 4, 8 and 11. The three grades are the final grades of three phases in the school curriculum. The tests include a range of questions in terms of the level of difficultly. A few questions that learners would typically not be able to answer correctly are included, in order to establish the degree to which learners perform beyond their grade level. The questions are all multiple choice questions, with four options per question for the Grades 4 and 8 tests, and five at the Grade 11 level. There are between 70 and 100 items or questions per test. The questions are designed to assess cognitive ability and skills rather than subject content. The item bank approach allows for tests to be designed which have different sets of questions, but are nevertheless equivalent with respect to the final score generated. Thus different schools may use different tests. Some questions are deliberately repeated across the years so that analysts can confirm the comparability of results. The item bank approach employed is similar to that used in the international Programme for International Student Assessment (PISA). Apart from tests, a background questionnaire for the learner, the teacher, the school principal and the school are prepared (the school principal questionnaire captures personal information relating to the school principal whilst the school questionnaire captures information about the school as a whole). The learner and school questionnaires are 2 pages each, the teacher questionnaire is 7 pages, and the principal questionnaire is 5 pages. The learner questionnaire includes questions on the month and year of birth of the learner, possessions in the home (in order to calculate an index of socio-economic advantage) and experiences of the learner in the classroom.

- **Selection of schools.** SAEB is run in around 6,000 schools across the country, where around 200,000 children are tested in the three grades. Thus an average of 33 learners per school are tested. Because SAEB aims to produce statistically reliable averages of learner performance at the level of each state (but not at a lower level), the percentage of learners tested in less populated states is higher than in more populated states, in other words there is over-sampling in smaller states. Mostly, each run of SAEB involves a new set of schools. However, a limited number of schools are retained over the years so facilitate the analysis of continuity in the assessment standards used. For instance, 480 schools took part in the 1999, 2001 and 2003 runs (De Paiva Franco and Menezes Filho, 2008: 4). SAEB is designed to represent all schools in the country, both public and private. Prova Brasil, on the other hand, excludes private and rural schools, and schools with fewer than 20 learners enrolled (the original 50 learner threshold was changed). Thus Prova Brasil's focus is currently on public urban schools that are not very small. Brazil uses an explicit classification of schools into rural and urban schools. Rural schools enrol around 15% of all learners. It is not made explicit why rural schools are excluded, and when these schools will be included within Prova Brasil, but one can assume that they are excluded due to logistical difficulties and cost considerations, and that the ultimate aim is for all public schools to be covered. Prova Brasil tests only Grades 4 and 8 for now. There were around 6.8 million learners in these two grades in 2007 (in public and private schools), and 4.1 million were tested, so the coverage ratio for Prova Brasil was around 60% in that year (where 100% would refer to the population represented by SAEB). In 2005 only 3.3 million learners were tested, so 2005 to 2007 saw a substantial increase. It should be noted that apart from SAEB and Prova Brasil certain states, such as São Paulo, run similar programmes themselves. There is deliberately no attempt by the national authorities to interfere with these state-run programmes in any way. For instance, there is no alignment of the school selection processes. This is to maintain the integrity of the national programmes.

- **School visits.** Schools are visited over a period of 15 days. This spreading out of the testing process is obviously facilitated by the use of the item bank approach, which allows different schools to write different but equivalent tests. Fieldworkers external to the national and state administrations are hired to conduct the visits. This obviously improves the integrity of results, but raises the costs considerably (and is undoubtedly a factor behind the exclusion of rural schools presently).

- **Compilation of data.** It was difficult to obtain information on this step in the cycle, but it is clear that a national dataset of SAEB tests results and background questionnaire responses is compiled, as well as a national dataset of Prova Brasil test results. It seems as if background questionnaire responses from Prova Brasil may not be entered into a database, though Prova Brasil schools do fill in these questionnaires.

- **Dissemination of results.** Results from Prova Brasil are posted to each participating school within half a year of the school visit. Each school receives a two-page summary plus a booklet explaining the method and aims of the programme. Moreover, the two-page summary for each school is made available on the INEP website – the 2007 summary for a specific school provided below, with superimposed explanatory labels in English, is from the website. School-specific results are not made available on the website if a school has queried the results. It was not possible to find out to what extent this has occurred. Apart from the average score per subject for each grade in the school, the two-page summary provides details on grade promotion rates, the IDEB index value combining learner performance and grade attainment (IDEB is explained below), and age-grade distortion statistics (this is essentially the percentage of learners who are more than one year above the grade-appropriate age). These statistics are all calculated for the school, the municipality, the state and the country. Moreover, the distribution of the school's scores across a range of performance levels is provided in the summary. It is not

clear whether public schools that participated in SAEB and not Prova Brasil (no school participates in both in any one year) also see their results distributed with the Prova Brasil results, but it seems to be implied that this is indeed what happens. Results for individual learners are never reported on. Moreover, there is no 'pass mark', so pass rates are not calculated anywhere in the two programmes.

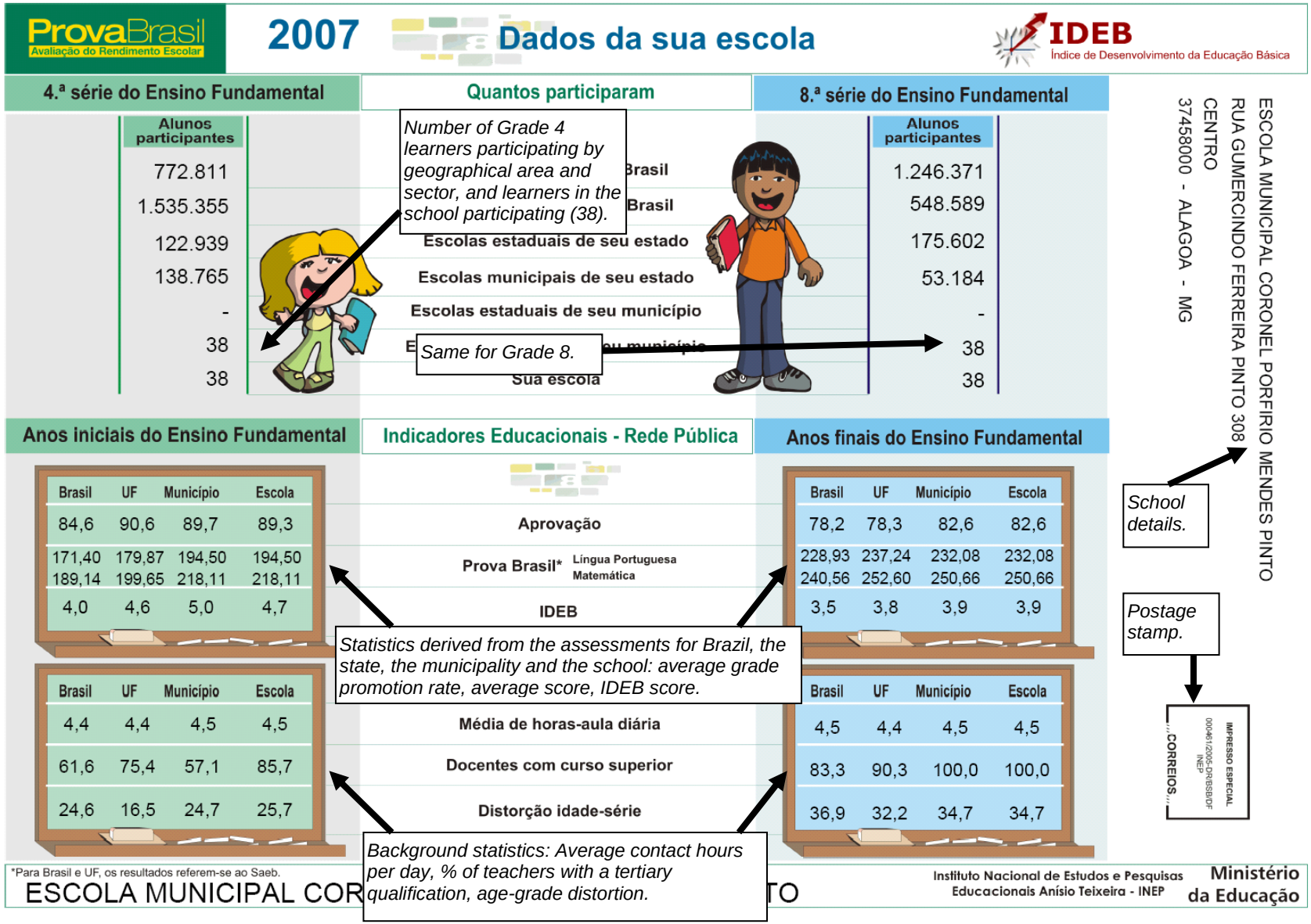
The Basic Education Development Index, or IDEB, is an important part of Brazil's monitoring of educational quality. It combines learner performance statistics from SAEB and Prova Brasil with grade attainment statistics. The index has been calculated since 2005, every second year, at the national, state, municipality and school level. For example, the IDEB index value for the Grades 1 to 4 phase for the country was 3.8 in 2005 and 4.2 in 2007, and a target of 6.0 has been set for the year 2021, implying that the country will have to improve the index value by an average of 0.26 every second year. The 2021 targets differ by, for instance, state and municipality, depending on what the historical values are. Where recent values are higher, the 2021 target values are higher. The target values generally lie in the range of 4.5 to 7.0. Whilst at the national and state level targets are calculated for three curriculum phases (given that SAEB tests at three grades), at the municipal and school levels targets are only set for the first two phases. The IDEB index is intended to prevent an overly narrow focus on just test scores, without some consideration of how many learners gain access to meaningful learning. Clearly, a situation where all the focus is on test scores could directly or indirectly cause worse performing learners to drop out of the system, just as a situation where all the focus is on enrolment and grade progression could result in inadequate attention to what learners learn. A balance between enrolment and performance must obviously be struck in evaluating the success of the schooling system. The IDEB index is calculated as follows:

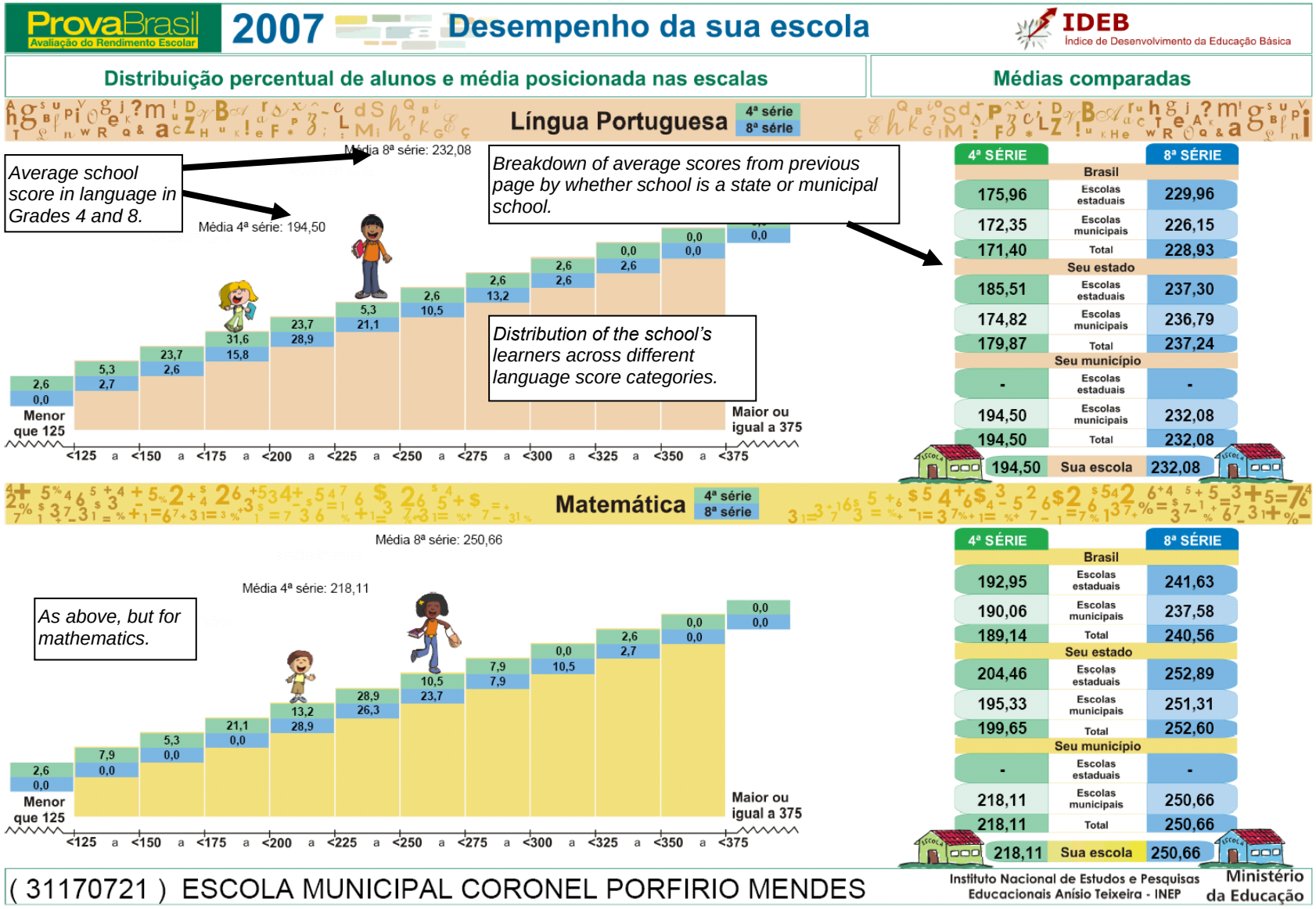
$$IDEB = N^{\alpha} P^{\beta}$$

N is the average score for a grade (for instance nationally or in a single school) and P is an indicator of grade attainment based on the grade promotion rates of all the grades in the curriculum phase. α and δ are weights determining how much emphasis to place on each of the two factors. P, the indicator of grade attainment, is calculated as follows:

$$P = \sum_{r=1}^{n} \frac{n}{p_r}$$

p is the promotion rate for a specific grade, in other words the percentage of learners who do not drop out or repeat, and n refers to the number of grades within the phase. Extensive discussion of this methodology and how it might influence schools can be found in an official INEP document (Fernandes, 2007). IDEB values are included in the two-page summary for a specific school appearing below.


ProvaBrasil
Avaliação do Rendimento Escolar
2007
Dados da sua escola
IDEB
Índice de Desenvolvimento da Educação Básica
4.ª série do Ensino Fundamental
Quantos participaram
8.ª série do Ensino Fundamental
Alunos participantes
772.811
1.535.355
122.939
138.765
-
38
38
Brasil
Brasil
Escolas estaduais de seu estado
Escolas municipais de seu estado
Escolas estaduais de seu município
Sua escola
Alunos participantes
1.246.371
548.589
175.602
53.184
-
38
38
Number of Grade 4 learners participating by geographical area and sector, and learners in the school participating (38).
Same for Grade 8.
ESCOLA MUNICIPAL CORONEL PORFIRIO MENDES PINTO
RUA GUMERCINDO FERREIRA PINTO 308
CENTRO
37458000 - ALAGOA - MG
School details.
Postage stamp.
IMPRESSO ESPECIAL
CORREIOS
Anos iniciais do Ensino Fundamental
Indicadores Educacionais - Rede Pública
Anos finais do Ensino Fundamental
Brasil UF Município Escola
84,6 90,6 89,7 89,3
Aprovação
78,2 78,3 82,6 82,6
171,40 179,87 194,50 194,50
189,14 199,65 218,11 218,11
Prova Brasil* Língua Portuguesa Matemática
228,93 237,24 232,08 232,08
240,56 252,60 250,66 250,66
4,0 4,6 5,0 4,7
IDEB
3,5 3,8 3,9 3,9
Statistics derived from the assessments for Brazil, the state, the municipality and the school: average grade promotion rate, average score, IDEB score.
Brasil UF Município Escola
4,4 4,4 4,5 4,5
Média de horas-aula diária
4,5 4,4 4,5 4,5
61,6 75,4 57,1 85,7
Docentes com curso superior
83,3 90,3 100,0 100,0
24,6 16,5 24,7 25,7
Distorção idade-série
36,9 32,2 34,7 34,7
Background statistics: Average contact hours per day, % of teachers with a tertiary qualification, age-grade distortion.
*Para Brasil e UF, os resultados referem-se ao Saeb.
ESCOLA MUNICIPAL COR TO
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP
Ministério da Educação


ProvaBrasil
Avaliação do Rendimento Escolar
2007
Desempenho da sua escola
IDEB
Índice de Desenvolvimento da Educação Básica
Distribuição percentual de alunos e média posicionada nas escalas
Médias comparadas
Língua Portuguesa
4ª série
8ª série
Média 8ª série: 232,08
Média 4ª série: 194,50
Average school score in language in Grades 4 and 8.
Breakdown of average scores from previous page by whether school is a state or municipal school.
Distribution of the school's learners across different language score categories.
Menor que 125
Maior ou igual a 375
<125 a <150 a <175 a <200 a <225 a <250 a <275 a <300 a <325 a <350 a <375
4ª SÉRIE
8ª SÉRIE
Brasil
175,96 Escolas estaduais 229,96
172,35 Escolas municipais 226,15
171,40 Total 228,93
Seu estado
185,51 Escolas estaduais 237,30
174,82 Escolas municipais 236,79
179,87 Total 237,24
Seu município
- Escolas estaduais -
194,50 Escolas municipais 232,08
194,50 Total 232,08
194,50 Sua escola 232,08
Matemática
4ª série
8ª série
Média 8ª série: 250,66
Média 4ª série: 218,11
As above, but for mathematics.
Menor que 125
Maior ou igual a 375
<125 a <150 a <175 a <200 a <225 a <250 a <275 a <300 a <325 a <350 a <375
4ª SÉRIE
8ª SÉRIE
Brasil
192,95 Escolas estaduais 241,63
190,06 Escolas municipais 237,58
189,14 Total 240,56
Seu estado
204,46 Escolas estaduais 252,89
195,33 Escolas municipais 251,31
199,65 Total 252,60
Seu município
- Escolas estaduais -
218,11 Escolas municipais 250,66
218,11 Total 250,66
218,11 Sua escola 250,66
(31170721) ESCOLA MUNICIPAL CORONEL PORFIRIO MENDES
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP
Ministério da Educação

The part of the INEP website devoted to SAEB and Prova Brasil is extensive. It includes around 50 web pages with separate web navigation routes for parents, school principals and teachers. A frequently asked questions (FAQ) section exists, and three promotional videos on the programmes can be viewed (but not downloaded). There are around 100 downloadable documents dealing largely with the pedagogical framework and providing examples of questions from the item bank. Background questionnaires can be downloaded, as well as the policy and legislative documents governing the programmes. To obtain school-specific Prova Brasil results, the user must select a municipality, and then select a school from a list (even schools which yield no results are included on the list).

A critical matter is of course to what extent educational quality is improving in Brazil. The SAEB results in fact indicate that the situation worsened between 1995 and 2003, but that the 2003 to 2005 trend is slightly positive (the PISA improvement for the 2003 to 2007 period discussed in section 2 seems to mirror this). The deterioration in the averages between 1995 to 2003 is generally attributed to the entry of large numbers of poorer learners into the schooling system in this period. The fact that despite the downward trend in the averages, results have been widely publicised can be seen as indicative of the absence of political manipulation of the process. It seems as if INEP's location as a professional body accountable to the Ministry, yet sufficiently protected by law from undue political pressure, is an effective solution.

Yet is must be noted that there are some conspicuous absences from the INEP website, which could be the result of political considerations. As of November 2009, the 2007 SAEB results were not available yet. Moreover, evaluation reports on how the new and relatively costly Prova Brasil has progressed since 2005 are not available. It would be important to know, for instance, the extent to which municipalities are opting out of the programme (by law they may do so) and to what extent schools have queried their results (something which prevents the publication of the school's results).

Prova Brasil and SAEB data have been used in a number of studies to identify what interventions seem most effective for improving learner performance. For instance, Prova Brasil data have been used to examine what practices in the municipalities are most associated with better performance relative to the socio-economic status of learners in municipally run schools (World Bank, 2008 – this is in English). It is worth noting that both the SAEB and Prova Brasil datasets have been made available to researchers outside INEP.

What is the benefit of running SAEB as well as Prova Brasil, in other words of maintaining a two-tier assessment system? The benefits of Prova Brasil clearly include the promotion of the accountability of the school to the authorities and to its community. The programme provides an empirical basis on which to decide how serious the need is for the school to change its teaching approach, for example. Similarly, Prova Brasil provides a basis for deciding which municipalities and states are in most urgent need of assistance. SAEB, by using a representative sample of schools, is able to produce reliable estimates of aggregate performance at the state and national levels, despite the coverage gaps in Prova Brasil. Moreover, SAEB is able to provide a reliable comparison of the public and private sectors. By dealing with a smaller number of schools, SAEB permits greater attention to the reliability of data, including data from the background questionnaires, than would be possible in the larger Prova Brasil. Arguably, even if Prova Brasil were extended so that it covered all schools and all three grades, there would still be a need for SAEB to serve as the quality control and in-depth research element in the larger programme.

3.2 Chile: Extensive use of financial incentives

Chile is often cited as a developing country that has introduced an exceptionally wide range of well-designed programmes to improve the education system. These programmes are well documented, both in a set of outstanding official websites (nearly all in Spanish) and in a

number of studies (many in English). Both these types of sources were consulted. The following are the studies used (all are in English): Delannoy (2000), McMeekin (2000), and Mizala and Romaguera (2005). Anand, Mizala and Repetto (2009), who examine private schooling for low-income learners in Chile (and whose study is therefore not directly relevant for what follows), was also consulted.

In fact, it is often asked why Chile's improvements in the international tests is not more unambiguous, given the presence of strong incentives for schools and teachers to perform (as was seen in section 2, PISA scores did improve, but not TIMSS scores).

Below, the focus is on two key financial incentive programmes in the Chilean schooling system: SNED (the full name in English is National Performance Evaluation System) and AVDI (Variable Assignment for Individual Performance). Each of these programmes rests on a sister monitoring programme: SNED rests on SIMCE (National Evaluation System), and AVDI rests on the Teacher Evaluation System.

SIMCE and SNED

SIMCE (website http://www.simce.cl), begun in 1996, is run every year and involves the testing of all learners in public and private schools in Grade 4, and in either Grade 8 or Grade 10. This means that just over half a million learners are tested in each year. Grades 8 and 10 alternate in a two year cycle. There are three tests, all of which draw explicitly from the official curriculum: language and communication, mathematics, natural and social sciences. They include a mix of multiple choice questions and open-ended questions. All learners in the same grade write the same tests, meaning there is no variation on the basis of a formal item bank of questions. The same test is written on the same day throughout the country. Fieldworkers external to the schools administer the tests. In each year some 12,000 fieldworkers are mobilised for this purpose. There are widely publicised rules around how the testing should occur, for instance school staff may not be present in the rooms in the school where learners sit for the tests. Universities assist in moderating the marking of the open-ended sections in the tests. Once the marking is complete, information sheets describing the results of individual schools, and how well they performance relative to other similar schools, are distributed to schools. Sheets for distribution to parents are also sent to schools. For the purposes of the comparison, socio-economic data collected through contextual questionnaires are used to divide schools into five socio-economic groups. Apart from questionnaires for teachers, principals and learners, a questionnaire for parents or guardians of tested learners exists. Copies of this questionnaire and envelopes are sent to schools before the test date, and parents are asked to complete the questionnaire, insert it into the sealed envelope, and to send it to the school on the test date so that it can be handed over to external fieldworker. These procedures are designed to protect the privacy of parents, an important matter given that certain responses in the questionnaire could imply criticism of the school's management. The categorisation of schools into socio-economic groups uses, apart from data gathered over time through the contextual questionnaires, other data on schools, for instance their degree of geographical remoteness. The categorisation follows a statistical procedure which aims to maximise differences between the five groups, and minimise differences within the groups. Thus, whilst the five groups are more or less of the same size, making the five groups equal is not considered a primary goal.

The SIMCE website makes school results available publicly, but with a strong emphasis on comparing schools to other schools within the same socio-economic group. A range of analysis and graphing tools that assist the understanding of the results are included on the website. Below, a translated output for a community that has five schools is given. The technical notes provide a sense of the complexity involved in presenting and interpreting results fairly, given the range of irregularities that could occur in the data collection stage. The SIMCE website allows researchers to download a database of results without restrictions, and

for researchers wishing to access the contextual data as well, a series of online forms must be filled in as part of an online approval process where MoE officials grant the permission.

Results of analysis

You have selected the following:

Variables	**Minimum, Maximum, Standard deviation, Average score, Difference**
Grade	**Fourth**
Years	**2006, 2007**
Subject	**Mathematics**
Filters applied	**None**

Compare Generate graphs Export to Excel

Generate PDF Help

2006 Grade 4 results:

	Mathematics			
School	Learners	Average score	Difference	Standard deviation
Escuela Básica Poeta Neruda de Isla Negra, El Quisco	33	221	↓ -24	41
Escuela Básica El Totoral, El Quisco	20	231	● 5	52
Escuela Particular San Miguel Arcángel, El Quisco	35	220	↓ -17	47
Colegio El Alba, El Quisco	17	253	● -13	37
Colegio El Quisco, El Quisco	32	244		41
TOTAL	137	232	● 1	45

2007 Grade 4 results:

	Mathematics			
School	Learners	Average score	Difference	Standard deviation
Escuela Básica Poeta Neruda de Isla Negra, El Quisco	32	249	↑ 28	37
Escuela Básica El Totoral, El Quisco	19	261	↑ 30	41
Escuela Particular San Miguel Arcángel, El Quisco	38	234	↑ 14	48
Colegio El Alba, El Quisco	12	261	● 8	36
Colegio El Quisco, El Quisco	28	203	↓ -41	49
TOTAL	129	238	● 6	48

Key:

↑	Score is significantly higher than scores in socio-economically comparable schools.
●	Score is not significantly different from scores in socio-economically comparable schools.
↓	Score is significantly lower than scores in socio-economically comparable schools.
	Where the score is blank this means that the school did not participate in the assessment or does not offer the subject.
-	No results are available as fewer than six learners participated in the assessment.
..	No results are available as a low rate of participation in the assessment made the average score unrepresentative.
*	No results are available as the average score was considered unrepresentative given that learners who did not participate in the assessment had separate school marks which were considerably lower than those of learners who participated.
**	No results are available as the assessments were not applied in the correct way.
***	No results are available because it was impossible to apply the assessment.
****	All learners or a substantial percentage of learners refused to participate in the assessment.
/	No results are available as there were problems in the application of the assessment.
(###)	The starting time of the assessment was significantly delayed for reasons beyond the control of the school.
~	No results are available for reasons beyond the control of the school.
%	Score must still be confirmed.
~~	For reasons beyond the control of the Ministry it was impossible to obtain the results.
+	The socio-economic classification is not valid as parent/guardian questionnaires were not filled in according to the guidelines.

SNED (website http://www.sned.cl), a programme which rewards financially all the teachers in relatively well performing publicly funded schools, is to a large extent (but not exclusively) based on the data generated by SIMCE. Every second year, a SNED score is calculated for every school in the country. 37% of this score is based on raw SIMCE results, whilst a further 28% is based on improvements in the school's SIMCE results over time, meaning that altogether 65% of the score is based on SIMCE data. The remaining 35% is based on school management and governance factors, for which data are largely derived from certain variables in the contextual questionnaires. Schools are then compared to other schools within the same geographical region and socio-economic group. The SNED comparison thus occurs within 80 groups (there are 16 regions and 5 socio-economic groups). Up to 35% of schools in each of the 80 SNED groups may qualify for the financial reward, though generally the percentage has been around 20%. The percentage may be lower than the maximum of 35% if there are not enough schools attaining absolute minimum thresholds. Schools qualifying for the reward receive additional funds, for two years only, which, on average, allow the school to pay all teachers an additional 'thirteenth cheque' in each of the two years. However, the funding is calculated in relation to the enrolment of each school, meaning that in general teachers with larger classes will receive more than teachers with smaller classes (the latter would to a large degree be the case in rural areas). The school is obliged to spend 90% of the SNED reward on additional pay for teachers, in proportion to the time each teacher spends teaching. The remaining 10% may be used for special projects, or to reward outstanding individual teachers. Between 1996 and 2008, SNED was run seven times, and in that period over 50% of schools have at some time benefitted from the programme. Spending on SNED comes to around 1.8% of overall public spending on schools. Does this mean that within each socio-economic group Chile spends more on better performing schools than worse performing schools? This important question does not seem to have been answered anywhere. Because the state spends money on programmes that focus more on poorly performing schools, for instance with respect to teacher development, the overall spending equity picture is complex. However, the overall picture created by the literature on Chile's schooling system is that public spending is highly equitable.

Mizala and Romaguera (2005) attempt to establish empirically whether SNED has succeeded in improving learner performance in Chile, obviously a vital question given the financial cost and effort that goes into SNED. It is extremely difficult to establish a causal link in a situation such as the one in Chile, because all publicly funded schools are participants and there is no control group that allows one to see what happens when the programme does not exist. Despite this, and using a reflexive model, Mizala and Romaguera (2005) find some evidence that SNED has made at least a small difference to the quality of education. Their approach should probably be supplemented with more qualitative research, based perhaps on instruments gauging the opinions of school principals, teachers and parents towards the programme. It was not possible to find such qualitative research.

The Teacher Evaluation System and AVDI

The second financial incentive that will be discussed here is the AVDI, which uses information emerging from the Teacher Evaluation System. The Teacher Evaluation System[9] is designed to evaluate, in a developmental manner, the professional capacity of the 70,000 teachers working within Chile's public schools. The evaluations are thorough, but resource intensive. Formally, the system is designed to evaluate each teacher every fourth year. However, between 2003, when the evaluations began, and 2009 only 53,000 of the 70,000 teachers had been through at least one formal evaluation. The roll-out of the evaluations has thus been slower than anticipated. The evaluation of each teacher requires the compilation of four different inputs. Written evaluation reports are needed from three different people: the

[9] A description of the programme in English is available at http://www.docentemas.cl/sistema_eng.php?PHPSESSID=c8ef3aaad1701db9208fb9cb361f3bae.

teacher herself (this is thus a self-evaluation), another teacher from another school, and the teacher's school principal. The fourth input is a portfolio of materials put together by the teacher. This portfolio must include a video of a lesson by the teacher. School principals and peer teachers are trained in how to perform the evaluation, and the education authorities provide the service of the video recording. An extensive computerised system captures all the four inputs for each teacher, which are then viewed by teams of accredited evaluators based at tertiary institutions, who grade the teacher as being excellent, competent, basic or unsatisfactory. The evaluators must moreover provide a comprehensive report on how they reached their conclusions, and what their recommended professional development activities are. The teacher receives the full report of the evaluators, and a summary is provided to the school principal. Teachers who are considered unsatisfactory are evaluated every year until they reach at least the 'basic' level. According to an agreement concluded with teacher unions, a teacher who remains unsatisfactory for three years running, despite receiving intensive professional support, is dismissed and paid a dismissal package.

Teachers who are considered competent or excellent may receive a financial incentive, lasting two to four years, if they pass an examination focussing on their knowledge of general pedagogics, and knowledge specific to the teacher's grade or subject. This incentive is referred to as the AVDI. Examinations are held annually, but a teacher may only sit for her specific examination once every third year. Teachers apply to sit for the examinations outside of the regular school channels, so participation is a relatively private matter. Whilst the names of those teachers who did pass the examination are widely publicised (for instance on the web), names of teachers who did not pass are not published. The AVDI website (http://www.avdi.mineduc.cl/) provides the names of around 3,800 teachers who passed the examination in 2008. One's AVDI amount is dependent on both one's examination score and on the grading one has received in the Teacher Evaluation System. This is why the years of eligibility range from two to four years. Moreover, the annual amount can vary considerably, from 5% to 25% of the value of the minimum salary for teachers. No cognisance is taken of the socio-economic status of one's school in the determination of the AVDI as it is assumed that although school communities are unequal, teachers have been through more or less the same quality of pre-service training (and the AVDI amount is not linked in any way to the performance of the teacher's learners). Partly to deal with the slowness of the roll-out of the Teacher Evaluation System, a second route to the examinations is through the Pedagogical Excellence Award, which is open even to teachers who have not been evaluated yet. However, this programme results in a smaller financial incentive than the AVDI.

3.3 Uruguay: A focus on involving teachers and their unions

This section is informed by two sources, both in English: Benviste (2000) and Ravela (2005 and 2006). Ravela was previously head of UMRE, the unit in the Ministry of Education responsible for the assessment of learners. The details found in these sources are barely featured on the Ministry of Education website, a fact that reflects the rather internal and not public nature of the quality enhancements programme described here. As will be seen, in many ways the programme is deliberately not very public.

Uruguay is a small country, with primary school enrolments just below 400,000 (85% of this is in public schools). The education system is highly centralised. The smallness of the country and relatively high levels of social cohesion (resulting, for instance, from comparatively low income inequalities in the Latin American context), can be considered enabling factors behind the high degree of agreement that has been reached between teacher unions and the education administration in the implementation of the quality enhancement programme.

The aspect of Uruguay's programme that is regarded as a critical success factor is the strong involvement of unions. As Benviste (2000) points out: 'This has been no slight

accomplishment in light of the fact that the evaluation of student performance may potentially exert a destabilizing role by highlighting deficiencies in educational service provision'.

The roll-out of a series of standardised learner assessments, beginning in the mid-1990s, occurred within a broader initiative to improve schooling. The four main objectives of this initiative were:

- The consolidation of social equity
- The appreciation of teacher professionalism and training
- The improvement of educational quality
- The strengthening of institutional management

The second objective was realised partly through improvements in teacher salaries, which had been declining in real terms. Some funding and advice for the initiative came from the World Bank and the Inter-American Development Bank. Presumably the financial assistance was not for teacher salaries, but for other elements of the initiative such as an extensive amount of preparatory research.

Below, details on the grade coverage, population coverage and external controls of Uruguay's standardised assessments for the period 1996 to 2005 are provided. The details do not reveal a neat pattern in terms of grade coverage and approach, which is probably a reflection of the fairly experimental and highly consultative nature of the assessments that were implemented. It is only twice, in 1996 and 1999, that every learner in a grade (Grades 6 and 9) was assessed through tests administered by external fieldworkers. However, information on developments after 2005 could not be found. Importantly, even when assessments were sample-based, assessment materials were sent to all public and private schools so that non-sampled schools could run the tests themselves and compare their performance to that of the sampled schools. There is thus a two-tier approach similar to the one in Brazil, but with less stringent controls (for instance no central marking of tests) for the non-sampled schools in most years.

Year	*Grade*	*Sample-based or everyone*	*Who administers the tests*
1996	Grade 6	Everyone	External fieldworkers
1998	Grade 3	Sample	External fieldworkers
1999	Grade 6	Sample	External fieldworkers
	Grade 9	Everyone	External fieldworkers
2001	Grade 4	Everyone	Teachers
	Grades 1 and 2	Sample	External fieldworkers
2002	Grade 6	Sample	External fieldworkers
2005	Grade 6	Sample	External fieldworkers

Teacher unions were at first opposed to having standardised assessments in schools. They suspected that the assessment results would be used in a punitive manner against teachers. However, a number of guarantees from the employer, and extensive consultation and teacher involvement in the design of the assessments eventually resulted in strong support from the unions. The details of the assessment cycle are generally as follows:

- **Test design.** At the early stages of the programme, local meetings of teachers were held where teachers themselves designed items for the tests after the purpose of the entire assessment programme had been explained to them. These local inputs were collected nationally, and informed the construction of national tests by assessment experts. The assessments employ a mix of multiple-choice questions and open-ended questions. A

formal item bank approach has not been followed, and in any year and grade, the tests for all learners would be exactly the same. Clear thresholds are established between four different performance levels, including a threshold distinguishing satisfactory from unsatisfactory performance.

- **Background questionnaires.** These are designed for learners, teachers and the school principal, and place much emphasis on establishing the socio-economic status of the learners in the school.

- **A two-tier application of the tests.** In most years, the two-tier approach mentioned earlier is employed whereby sampled schools are tested in a controlled manner, and assessment in the remaining schools is relatively uncontrolled. The reason for this combination of approaches is largely budgetary. It would be too costly to employ external fieldworkers in all schools. The assessment are first run in the sampled schools, and thereafter the assessment materials are distributed to all remaining schools.

- **Collection and dissemination of data.** The programme upholds a rule that results from centrally marked tests will be distributed to schools within 40 days after the tests have been written. All schools are sent tables similar to the one appearing below (Ravela, 2006). All cells in the table will already have been filled in for sampled schools, and for non-sampled schools the first column will be blank. Non-sampled schools will fill in the missing information and return this to the administration. They also submit their background questionnaires for central data capturing. Once the results from non-sampled schools have been collected, learner performance levels for districts can be calculated. The district-level results are widely distributed, at least amongst teachers and education managers, but the results of individual schools are sent only to the schools concerned. This is to prevent a 'name and shame' tendency, something which teacher unions were very interested in avoiding.

Reporting results to schools

Percentage of students:	In this school	In schools of the same social context	At the national level
Passing the cut-score in the whole test			
Passing the cut-score in "Understanding of mathematical concepts"			
Passing the cut-score in "Use of algorithms"			
Passing the cut-score in "Problem solving"			

Much emphasis is placed on how the assessments can be used to improve the quality of schooling. The following steps are taken:

- **Targeting and designing teacher in-service training.** The information from the background questionnaires is used to identify the most socio-economically disadvantaged schools, and these schools are favoured in the targeting of in-service teacher training. As part of this process, schools are divided into five socio-economic quintiles. It was not possible to establish to what extent the favouring of poorer schools occurs, specifically how much more likely disadvantaged schools are to be targeted than advantaged schools. Manuals prepared for the teacher training are directly based on the results from the assessments, with much attention going toward what learners got wrong, and what teaching approaches could remedy the situation. At least twenty different training manuals have been developed out of the assessment results. Training lasts for a year, and takes place every second Saturday. Participating teachers are paid a stipend for participating equal to 25% of their annual salary. This is why it is important for the targeting of schools to occur on the basis of socio-economic advantage, and not test

results, as the latter could create a perverse incentive for schools to perform poorly. Three teachers per school are selected, two by district officials (they are told to identify the most capable teachers teaching the grade in question) and one by the teacher union. Teachers participate in the training in groups of around 60. These arrangements clearly imply some risks, especially given how large the stipend is. The choices made by district officials could be influenced by nepotism, for instance. The sources do not discuss these risks.

- **Identification of best practice schools.** The learner performance and background data are used to identify schools that perform exceptionally well relative to their socio-economic status. These schools are then considered model schools whose staff can advise other schools on optimal teaching and learning practices.

- **Monitoring the effectiveness of the overall initiative.** The results are used to gauge the success of the wider school improvement initiative. At the national level, the improvements have apparently been remarkable, as indicated by the following graph (Ravela, 2005: 38). In just six years, the percentage of learners achieving a satisfactory score in mathematics increased from 35 to 48. The figure in the most advantaged quintile of schools is around 70. The improvement has been greatest in the poorest quintile, where the percentage of learner achieving a satisfactory score doubled. Similar improvements have been recorded in the language scores. The improvements are so large that one may ask how reliable the data are, and specifically how comparable results are over time. The sources used claim the results are comparable, though how this comparability is ensured was not explained.

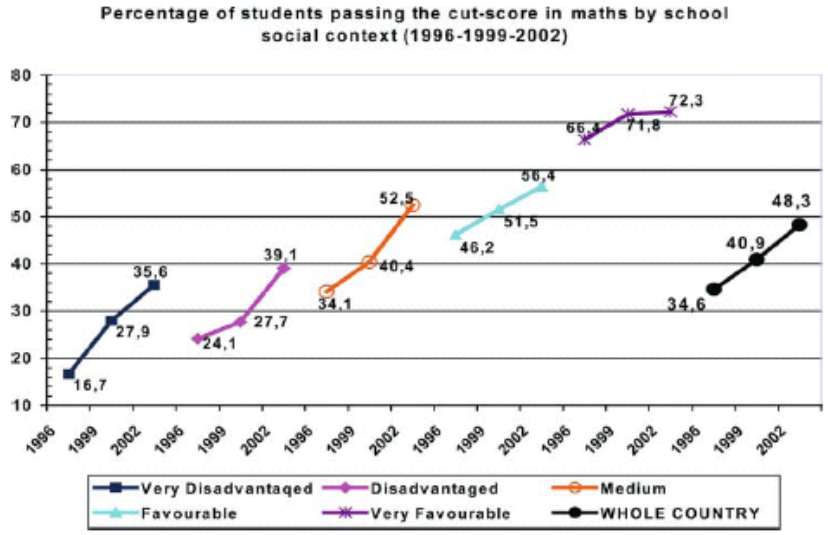


FIGURE 3. Percentage of pupils passing the cut-off score in mathematics by school socio-economic context, 1996, 1999 and 2002. *Source:* ANEP/GIE, 2003.

Ravela (2006: 2), in explaining how it was possible for Uruguay to achieve these dramatic improvements when a country like Chile, which also introduced an extensive system of learner assessments, did not, says:

> The main assumption underlying this approach has been that there will be no improvement in students learning if the state does not invest in teachers' knowledge and ability to teach... There is no direct [causal link] between implementing assessments and improving student achievement. You necessarily need to improve teaching, if you want to improve learning.

An important element of the programme at the national level is a 15 member advisory group representing teacher unions, education districts, teacher training institutions and private school organisations. When UMRE's assessment programme began, the advisory group met UMRE officials on a monthly basis. When the programme stabilised, the frequency was reduced to three times a year.

Surveys by UNESCO, have revealed considerable buy-in by teachers in the programme, despite initial misgivings. 70% of teachers support the national assessment programme, and 70% were familiar with the teaching manuals emerging from the programme. 55% of teachers have said that they changed their teaching methodologies as a direct result of their involvement in the programme.

To sum up, Ravela (2006) emphasises that the following are necessary features of a national assessment system:

- Teachers should be involved in the design of the assessments and the programme as a whole.
- The data obtained through the programme should feed into the teacher training methodologies.
- The development of training materials should be based on the data, and should involve teachers.
- Tests should be methodologically sound so that the comparability between schools, geographical areas and over time is credible.
- There should be no political interference in the programme. To illustrate this point, Ravela (2006) mentions that national results from the assessments are generally disseminated amongst teachers before the Minister of Education has seen them, and before they are released to the media.
- The programme should be predictable and follow a plan that has been agreed on by all stakeholders.

Ravela (2006) identifies as a key challenge the dissemination of results to parents. This, it is argued, could improve accountability of schools to their communities, yet has not been emphasised in the programme in the past as the focus has been very much on the use of the data for teacher training purposes.

3.4 Mexico: Widespread teacher incentive payments with little effect

The section on Mexico focuses on a widely known and rather well documented teacher incentive programme named Carrera Magisterial, henceforth abbreviated to CM (the translation to English would be more or less Teacher's Career Path). This programme is different from the other programmes described above in that analysis of the programme has led to conclusions that CM is not having the intended effect, and that the high cost of the programme makes it inefficient. There are therefore good reasons for regarding CM as indicative of what not to do. A study of the programme commissioned by the Ministry of Education (see Santibáñez, Martínez, Datar *et al*, 2007) was consulted, as well as a summary of the study (McEwan and Santibañez, 2005) (both texts are in English). The Mexican Ministry of Education website was offline whilst this section was being written, so it was not possible to obtain information from that source (http://www.sep.gob.mx/).

Historically, CM is significant, as the official study points out:

> By implementing Carrera Magistral, Mexico became one of the first countries in the world to link teacher salaries to performance in public schools. (Santibáñez, Martínez, Datar *et al*, 2007)

The beginnings of CM can be traced back to 1993. It was the outcome of close collaboration between the teacher union and government, and the governance of the programme occurs through a national committee with eight government and seven union members. Mexico is a federation with 32 states, each of which is responsible for providing school education. As will be seen, CM has both highly centralised elements and elements where decision-making is devolved to the states. 1998 saw major reforms to the programme, following a survey gauging teacher perceptions of CM.

The annual CM cycle, as teachers would experience it, is described below.

- **Application by eligible teachers.** First of all, eligible teachers who are not already in CM must decide if they would like to enrol for the programme. A teacher is eligible if he is employed on a permanent full-time basis in a public primary or secondary school. Moreover, the teacher must have been employed for a minimum number of years, which is two to six years, depending on, for instance, the teacher's qualifications level (more qualified teachers have a shorter waiting period). Principals in public schools are also eligible, but are subject to slightly different CM rules. To facilitate the discussion, only the CM rules applicable to teachers are described. Teachers who are already within the programme (once one is within the programme, one stays in the programme as long as one remains in a public school) must decide whether they would like to move up a level in the programme. CM has five levels, named A, B, C, D and E. One first enters level A, and then moves up the other levels as far as level E. Each subsequent level brings with it additional pay, which is never lost as long as one remains within the public schooling system. In 2001, 58% of public school teachers were already within CM, and of this 58% over half were on level A (figures for more recent years could not be found). Teachers who are already within the programme must wait a minimum of two to four years between each promotion up the CM ladder. The precise number of years depends partly on which state you are in. Poorer states tend to have shorter waiting periods.

- **Years of service and level of qualification.** Teachers applying to enter CM or to move up a level must provide proof of their years of service and their qualifications level. The evidence must be sent to the national CM authority. Teachers who are already within CM receive a form with information based on previous submissions of information already filled in. Years of service and level of qualifications are the first two of six factors that are used to calculate a score which is then used to determine whether the teacher can enter CM, or move up a level.

- **Teacher development courses.** Teachers who have submitted the application form must participate in teacher development courses offered by the authorities on a part-time basis over a six-month period. Most of the courses are run by the state, but there is also a short course offered by the national authorities. The courses vary only by the grade you teach and (for teachers in secondary schools) your subject specialisation. Within the same course there may for example be a mix of teachers trying to enter level A of the CM ladder, and teachers trying to be promoted from D to E (in other words supposedly the best teachers in the system). A score is attached to each teacher following the development course, and this score is submitted to the national CM office.

- **Subject content test.** Next teachers write sit for a subject content test offered by the national authorities. Again, this test only differs according to grade or subject taught.

- **Learner test.** A test is now run for the learners in one class currently taught by the applicant teacher. The testing procedure is said to include comprehensive security and control aspects, including the use of external fieldworkers to administer the learner tests. However, it was not possible to find much in the way of details on this part of the CM programme, clearly a very costly part. For instance, details on the length of the test, how the test is prepared, how the class is selected in the case of subject teachers, and the extent to which multiple-choice questions are used could not be found.

- **Professional evaluation.** Next, a score is attached to the applicant teacher's teaching ability by a panel of evaluators, which includes the school principal and fellow teachers. This score is sent to the central CM authority.

- **Determination of successful promotions.** At this point, the central CM office should have six scores for each applicant teacher, assuming all six steps have been followed (many teachers drop out already at the stage of the professional development courses). A final score is calculated per teacher using the following weightings for the six CM elements: (1) qualifications level (10%), (2) years of service (15%), (3) results from the professional development courses (28%), (4) results from the subject content test (17%), (5) results from the learner tests (10%) and (6) score given by the principal and peer teachers (20%). The teacher and student test scores used in the calculation are adjusted according to how well a teacher did in relation to other teachers in the same group, where group is defined by level of schooling, the state one works in and socio-economic stratum (there are five of these strata, e.g. 'marginal rural area' and 'medium development urban area'). This helps in ensuring that learner tests results from disadvantaged learners do not unduly disadvantage teachers. The CM office sends final CM scores to individual teachers and state governments, and also publishes results on the internet. It is the state that has the final say on which teachers are successful in moving up one level in the CM ladder, or entering level A of the programme, insofar as it is the states that determine threshold scores for successful entry into each of the five levels. This determination is based on the budget available. The federal government transfers CM funds to states, but states may vary, within limits, as to how much goes towards teachers, as opposed to principals and administrators, and how much goes towards movement up levels A to E, and how much goes towards teachers entering level A.

- **Substantial salary rewards.** Finally, a move onto or up the CM ladder results in a non-reversible improvement in one's salary. This improvement amounts to, for instance, 27% of the basic salary for teachers on level A, and 104% for level C and 215% on level E. The rewards are thus substantial.

The CM cycle involves a large number of public school teachers. Of approximately one million teachers, around 60% attempted in 2001 to enter CM or move up a level in the programme (2001 was the more recent year for which these figures were available). However, only around 0.6% of teachers (around 6,000 teachers) moved into level A in 2001, and only around 200,000 moved up a level within the programme. The fact that there are already so many teachers on one of the five levels within the programme (58% of all teachers, as mentioned above) is explained by the fact that when the programme began, it was much easier to enter CM than once the programme had matured.

The criticisms of the CM programme can be summed up as follows:

- **Virtually no discernable impact on learner performance.** The weightiest criticism is that though the programme costs are high (precise figures could not be obtained, but one should remember that teachers retain their CM benefits for their working lives), there are hardly any observable benefits in terms of improved learner performance. This conclusion is based on thorough analysis of the CM dataset (see Santibáñez, Martínez, Datar *et al*,

2007). In a regression analysis, it was found that the association between the learner test scores and the other five scores used by CM was non-existent or very small. It was only the score from the professional development course that was positively correlated with the learner scores, and this association was small. A higher teacher subject knowledge score, or a higher level of qualifications score, are not associated with better learner performance, at least not within a multivariate regression model. This is surprising, and suggests that educational quality in Mexico is not easily improved through typical interventions. The analysis also tried to establish whether there were at least temporary improvements in learner performance in cases where it was clear that even a relatively small improvement in the learner performance variable would make a difference between obtaining the CM promotion and not obtaining it. It seemed natural to assume that in such cases teachers would make a special effort to improve learner performance. A weak phenomenon of temporary improvements was found in secondary schools, but such temporary improvements would be followed either by a reversion to the original level of performance, or even a deterioration relative to the original level. These findings should not surprise one, given that teachers cannot lose a CM promotion once they have obtained one, no matter how poorly they perform.

- **Lack of rigour in test preparation.** The teacher and learner tests seem not to follow sufficiently rigorous preparation steps. They tend to be prepared by few people with little validation by a larger technical reference group. One manifestation of this is a lack of technical documentation explaining the rationale and procedures behind the tests.

- **Universally generous professional evaluations.** The evaluation of the teacher's professional capacity by the principal and peer teachers seems fairly meaningless in the sense that scores are consistently very high for all teachers.

- **Problem with the use of just one class in one year.** The use of data from just one year for the final scoring of each teacher seems inadequate, especially given that over the years historical data for each teacher is collected within the CM database. It is also questionable whether learner test scores from one class represent a fair reflection of the teacher's abilities, given that such scores are the result of the efforts put in by many teachers within one school.

Despite the above criticisms, as noted in section 2, PISA detected an improvement in Mexico's mathematics performance between 2003 and 2006. It is possible that if the assessment of CM had used data beyond 2002 (the most recent data used by the analysts) stronger evidence of possible linkages between CM and learner performance would have been found. However, the structural weaknesses of CM identified by the analysts are serious, to the extent that it seems improbable that post-2003 improvements would be clearly attributable to the CM programme. It seems more likely that improvements would have occurred despite the structural problems with the programme.

3.5 Peru: Advocating greater accountability to parents

Whilst no long-range programme is discussed with reference to Peru, a research-driven advocacy campaign that occurred in 2005 to 2006 seems to deserve mention. Peru has participated to a lesser degree than other Latin American countries such as Brazil and Chile in internationally benchmarked testing programmes, and has a less ambitious national testing programme. There appears to be no evidence of significant quality improvements in Peru's schooling in recent years. Nonetheless, the 2005-2006 campaign appears to be sufficiently unique to warrant a brief discussion of it.

Research undertaken in a small number of schools (fewer than 30) involved testing primary level learners using traditional pen and paper tests, but also recording the ability of learners to

read a simple text out loud. It was found that scores in the traditional tests and a count of words read out loud per minute correlated well, suggesting that reading out loud was a good proxy measure of a range of language skills. The researchers went further, and argued that it would be strategically appropriate to encourage teachers to focus on improving the ability of their learners to read out loud, given that this ability is so easily measured, and given that improvements in this ability are unlikely to occur without corresponding improvements in other aspects of language cognition.

The researchers identified several benchmarks against which teachers could measure themselves. For instance, one benchmark was that Grade 2 learners learning in Spanish should be able to read a text of 60 words in one minute.

The research led to the production of a number of videos aimed at parents, teachers and policymakers[10]. As an example, the video aimed at parents encouraged parents to test the words per minute ability of their children, and to talk to their children's teachers if they fell short of minimum standards. The analytical literature on the campaign emphasises the importance of balancing the various research and communication imperatives in the production of the videos.

> … the video team received detailed comments, with minute and second markers, from the researchers, over several drafts of the video. The video producers should not be expected to simply produce directly from the research results without ongoing, fine-grained, and intense interaction. (USAID, 2006: 3)

The videos were used in a range of public discussions on major television channels. One political party included within its education manifesto goals using the methodology in the videos.

An academic article (in English) describing the campaign is offered by Abadzi, Crouch, Echegaray *et al* (2005).

4 Lessons for South Africa

As has been emphasised in earlier sections, the optimal design of any policy aimed at enhancing educational quality depends strongly on contextual factors such as the role of teacher unions in the political economy of the country, the general level of social cohesion, and the degree of federalism or decentralisation in the schooling system. Perhaps the most important lessons that can be extracted from the above discussions for South Africa have to do with what policy questions to ask, what items to place on the agendas of consultative and decision-making meetings, and what issues to bear in mind when answering the key policy questions. Below, what appear to be key policy design questions emerging from the foregoing discussion are listed, and discussed, but in a way that seemed relevant for the South African context.

Several key questions emerge with regard to the optimal design and usage of a national assessment programme such as Brazil's SAEB or Chile's SIMCE.

- **How should the governance of the programme be structured?** In South Africa, the sample-based Systemic Evaluation, clearly an important point of departure for rolling out universal testing for Grades 3, 6 and 9, as envisaged by a number of semi-official

[10] The videos (in Spanish, but some with English sub-titles) can be accessed at: http://web.worldbank.org/WBSITE/EXTERNAL/BANCOMUNDIAL/EXTSPPAISES/LACINSPANISHEXT/PERUINSPANISHEXT/0,,contentMDK:20844760~pagePK:141137~piPK:141127~theSitePK:501764,00.html.

documents, is governed largely from within the education bureaucracy. The Uruguay and Brazil experiences point to two important governance considerations. Firstly, the Uruguay experience illustrates how formal involvement by teacher unions in the decision-making of the programme can greatly assist in making the programme acceptable to teachers and encouraging teachers to use the findings from the assessments in determining professional development approaches. In Uruguay, teacher unions are included in a standing advisory committee which, at the beginning of the programme, met as frequently as once a month. The emphasis was thus not on a once-off endorsement by the teacher unions, remembering for instance the teacher union leadership changes over time, but rather intensive and ongoing involvement through a formal structure. Moreover, Uruguayan teachers themselves made inputs into the process of designing test items, something which would have improved the buy-in from teachers. Secondly, in Brazil the national assessment programme is required by law to take into consideration advice from a technical committee which includes experts both from the education bureaucracy and other institutions such as universities. This measure recognises the fact that designing an optimal assessment programme is technically complex, and the fact that important expertise resides outside the bureaucracy.

- **Should national and provincial programmes be aligned?** In South Africa, provinces differ greatly in their application of provincial learner assessment programmes. Western Cape and Gauteng have both introduced their own universal testing programmes at the primary level, whilst most other provinces only tentatively looked at work in this area. Currently, it appears as if the Systemic Evaluation and the existing provincial programmes function completely independently of each other. However, if universal testing from a national level is introduced there could be strong arguments in favour of aligning the provincial programmes (where they exist) and the national programme, partly to avoid assessment fatigue within schools. In Brazil, the national universal assessment programme operates completely independently of similar state-level programmes, the argument being that any attempt to align programmes (for instance by using the same tests) could compromise the integrity of the national programme. The answer to the alignment question is not an easy one. Costs such as assessment fatigue in schools must be weighed up against costs such as loss in programme coherence and integrity. One of the advantages with a decentralised schooling system is that it allows the national level to act, in a sense, as the judge of education governance at lower levels, so that the accountability of provincial administrations is improved through practices such as inter-provincial comparisons. However, this advantage is only realised if the national level retains considerable control over monitoring systems. It should be noted that the federal government in Brazil has moved towards increasing central control over the national assessment programme, even though this has not always been popular amongst state governments. Of course national control over the national programme should not prevent provinces from running their own parallel programmes that fulfil specific provincial needs.

- **Should an item bank approach with differentiated tests be followed?** South Africa's Systemic Evaluation has involved the development of item banks of test questions. However, the approach of producing different but equivalent tests for different schools seems not to have been pursued. Brazil's assessments do pursue this approach, Chile's assessments appear not to, and Uruguay's do clearly not. Differentiated tests for the same grade undoubtedly do introduce additional costs and make the logistics more complex. However, they could also save costs and simplify the logistics. Specifically, the differentiated tests used in Brazil make it easier to spread the assessments out over many days. In Chile the same tests must be written on the same day throughout the country. The Brazilian approach can clearly make it easier in a context (like the South African one) where many schools are difficult to reach.

- **Should open-ended questions be included in the tests?** The Systemic Evaluation has included a mix of multiple-choice questions and open-ended questions (Department of Education, 2005: 8). Brazil's assessments employ only multiple-choice questions, whilst Chile and Uruguay employ a mix of the two approaches. In Chile universities are involved in moderating the marking of the open-ended questions. Clearly, the cost of a mixed approach is higher, and the additional cost obviously increases if universal testing of specific grades is introduced. Assessing the advantages and disadvantages of the two options falls beyond the scope of this paper, but expert pedagogical advice must be combined with cost considerations when decisions are taken. The Latin American evidence suggests that neither option should be ruled out as being unfeasible.

- **Should a two-tier approach be pursued?** This is a vital question. In South Africa, political pressure to universalise assessments is sometimes interpreted as a need to expand the existing Systemic Evaluation so that it covers all schools with a specific grade, and not just a sample of schools. The high unit cost of the Systemic Evaluation, caused for instance by the employment of external fieldworkers and the data entry of several background questionnaires, including parent questionnaires, from each school, makes it difficult or impossible to universalise the Systemic Evaluation without compromising on elements of the programme one may not want to lose, at least not for a core sample of schools. In Brazil and Uruguay the education administrations have responded to the pressure to universalise assessments by introducing a two-tier approach. On one tier, a closely controlled sample of schools is tested (and extensive background data are collected) whilst for remaining schools controls, but also costs, are reduced. In Brazil this cost reduction appears to occur mainly through the fact that background questionnaires are not fed into a central database in the case of non-sampled schools, through the exclusion of private schools in the non-sampled group (the sample includes private schools), and through the exclusion of Grade 8 testing in the non-sampled group. Uruguay goes a step further and does in general not employ external fieldworkers in non-sampled schools and allows these schools to mark their own tests. The controls for non-sampled schools are thus much thinner in Uruguay than in Brazil, but costs would also be reduced. A two-tier approach seems optimal, but precisely how the non-sampled or universal tier of schools is treated seems to be a matter for careful decision-making. Pedagogical, programme impact and cost considerations all need to be taken into account. From both a cost and programme integrity perspective, the question of whether to employ external fieldworkers in non-sampled schools is a key one. One compromise solution is for officials from district offices to act as external fieldworkers. However this could reduce the integrity of the programme unacceptably, especially if results are used to assess the performance of district offices, as is done in Brazil.

- **Should independent schools be included?** South Africa's Systemic Evaluation does not include independent schools. Brazil and Chile do include such schools, Brazil within the tier of sampled schools only, and Chile on a universal basis. There are strong arguments for including independent schools within the Systemic Evaluation. Not only is it important to know how well learners in independent schools perform, comparisons of independent and public schools, especially where both offer schooling mainly to poorer communities, can reveal important trends that can inform the design of policies governing both types of schools.

- **Should a phased approach to eventually covering all schools be followed?** A key question is whether the desired universalisation of assessments (for instance for three grades across all public schools in every year) should be reached incrementally, or in a 'big bang' manner. The tendency is probably to under-estimate the costs and complexities of expanding an assessment programme. Such under-estimation could conceivably have serious consequences if it leads to an over-ambitious implementation plan. Budgetary or human resources constraints may make the plan unattainable, and introduce a sense of

failure. If too much effort goes towards the volume of assessment data collected, the quality of the data may be compromised, and there may not be sufficient effort devoted towards the use of the data to improve performance. In South Africa, having universal testing in three grades every year has been proposed. Brazil's approach of testing learners every second year (there is apparently no commitment to annual testing in the near future) seems to allow for an opportunity for results to be absorbed and acted on in the system in non-test years. Moreover, if the aim is to gauge improvements in schools, and not to evaluate individual teachers or learners, then universal assessments every second year seem more than adequate. Chile runs its assessments every year, but it should be remembered that Chile has a more centralised governance system, and enjoys relatively high levels of capacity within its education bureaucracy. The present context in South Africa is probably more like that of Brazil than that of Chile. Another key phase-in question is whether certain types of schools, for instance rural schools in the case of Brazil, should be excluded from the universal testing programme initially due to capacity constraints. Such exclusion would probably not be acceptable politically in South Africa. Yet including rural schools right from the start of a universal assessment programme underlines the difficultly of running the programme annually, and the importance of an item bank approach that would allow testing to occur over several days or weeks.

- **How should the programme be communicated to the nation?** The Brazilian and Chilean programmes both feature sophisticated public communication strategies. The internet, videos, and television programmes are used to advocate the importance of the programme, to communicate what its findings mean for various stakeholders (including parents), and to assure the public that the intention of the programme is a constructive one and not one of 'naming and shaming'. South Africa's Systemic Evaluation, on the other hand, features very few of these communication elements. For instance, it is difficult to find information on when the programme will be run in future years, and official parent-friendly explanations of what the national and provincial average scores mean in terms of the challenges facing education do not seem to exist. If universal standardised assessments are to be introduced in South Africa, it would be important for a fully-fledged communication strategy to be launched. Arguably, an office should be dedicated just to the implementation of effective communication. The Peru advocacy campaign described in section 3.5 underlines how important changing public awareness of educational quality, through effective media strategies, can be.

- **Should there be a pass mark?** The Systemic Evaluation has a 'soft' pass mark which is more or less the dividing line between the categories 'not achieved' and 'partly achieved' (the other two categories used are 'achieved' and 'outstanding') (Department of Education, 2005: 73). Interestingly, Brazil's programme places no or very little emphasis on such categories. On the other hand, Uruguay places much emphasis on the percentage of learners reaching a pass mark (the national improvement described in section 3.3 is based on such a percentage). Clearly, there are advantages to each approach. Having a clear indication of what learners should be achieving is important for gauging the success of the system or individual institutions. However, too much emphasis on pass marks and thresholds can pervert the impact of the programme if, for instance, this makes schools with high pass rates complacent about their performance.

- **Should results for individual schools be disseminated widely?** This is a complex and relatively widely debated matter. On the one hand, it is argued that parents have a right to be able to compare the quality of education in their school to that in other schools, so that appropriate parent pressure can be put on schools to improve. On the other hand, it is argued that publicly available school scores can cause migration towards better performing schools, and hence instability with regard to the supply of teachers and other education resources. Eventually better performing schools can become swamped, and lose their performance advantage. This scenario is clearly possible in a context such as the

South African one where parents are not confined to strict school zones, large performance differences exist between schools (even within the group of historically disadvantaged schools) and it is relatively easy for principals to translate increased enrolments to more public resources (and a salary improvement for the principal). For many years, Grade 12 examination results (specifically the school pass rate) have been widely publicised in South Africa. It appears as if there has been no rigorous analysis of how this availability of information influences parent and student school choices. It is thus difficult to gauge to what extent the hypothetical problems associated with the publicising of school-level results would be felt in primary schools. Brazil and Chile both publish school-level results on the internet and presumably this information filters through to the media in general. Uruguay, on the other hand, distributes school performance results only to the staff of the schools concerned (not just principals, but also teachers). Parents therefore have no automatic access to this information. The former chief of the Uruguayan programme acknowledges that lack of parent access to results is a drawback with the programme. It seems beyond the scope of this paper to make any strong recommendations on what would be optimal for South Africa. It should be noted, however, that nothing in the Brazil or Chile sources suggested that widespread dissemination of school-level results caused problems in the management of the education system.

- **How should the assessment programme contribute towards quality improvements?** There is a saying that one does not make the cow fat by measuring it. Gathering of learner assessment data cannot, on its own, make a difference to educational performance. Brazil, Chile and Uruguay display different and interesting approaches to using the results from the assessment programme to bring about change. In Brazil, much of the emphasis has been on combining the assessment results with an indicator of enrolment (specifically a grade promotion rate) to calculate a composite indicator, the IDEB, at the national, state, municipality and school level. On the basis of baseline values of this composite indicator, target values (for the year 2021) have been calculated and widely publicised. The country has a target, as does each state, municipality and school. A higher baseline value translates into a higher target value. A crucial advantage with the Brazilian IDEB indicator is that it is designed to avoid perverse incentives whereby worse performing learners are overtly or subtly pushed out of the school, a real problem in Brazil given lingering problems with primary level dropping out. In Chile, the results from the assessment system are used to drive a competition every two years between schools in similar socio-economic contexts. Winning schools receive a reward in the form of a bonus for teachers, amounting to more or less a 'fourteenth cheque', which lasts for two to four years, depending partly on how well the school performed. Whilst the teacher reward programme seems popular, whether it actually causes quality improvements is still not clearly established. This is an important matter, given the financial cost of the rewards, and the effort that goes into the programme. Uruguay has used its assessment results mainly to inform teacher in-service training in explicit ways. For instance, state-run workshops for teachers (where teachers are paid a stipend for attending) focus largely on specific learning problems revealed through the assessments and on how better teaching can remedy this. What can South Africa learn from all this? There is already a strong commitment in South Africa towards establishing rigorous learner performance targets. Brazil's experiences suggest that a composite quality-quantity indicator may be optimal, and that target values should perhaps be differentiated according past trends (thus there would be pressure even on better performing schools to improve their results). Uruguay's practice of basing teacher training explicitly on what the assessments show seems easy to reproduce in South Africa, and of undeniable value. There is no reason why this approach should be used not just in in-service training, but to some extent even in pre-service training. Put differently, pre-service teacher trainees should be exposed to and learn to analyse specific learning problems that are experienced by South African learners. A Chilean-type teacher rewards system has in fact been the subject of discussion in South

Africa. However, this approach is clearly more debatable than the Brazilian indicator or Uruguay's teacher training. Designing a rewards system that would work in the South African context would be a complex matter. Moreover, even with an optimal design, there tends to be no assurance (at least not yet) that these kinds of rewards exert a positive impact on learner performance.

Moving beyond the matter of standardised learner assessment programmes, the Peru campaign discussed earlier shows how parallel to traditional pen and paper testing, very simple word count exercises where learners read out loud can popularise the setting of quality standards in education. These exercises are so simple that parents can initiate them, which in turn can strengthen parent pressure on schools to perform.

A teacher rewards programme from Mexico was discussed in section 3.4 largely to illustrate how a poorly designed quality improvement programme can bring about virtually no benefits at high costs. It is instructive to compare the Mexican and the Chilean teacher rewards programmes. Whilst the Chilean programme recognises that teaching is a team effort (all the teachers in a school are rewarded on the basis of the learners' performance scores), the Mexican programme attempts to reward individual teachers based on the scores of the learners taught by the specific teacher. The literature on teacher incentives indicates rather strongly that such individually based reward programmes are unlikely to be well targeted. In the Chilean programme teachers enjoy the reward for a limited number of years, and have to ensure that the school performs exceptionally well (within its socio-economic category) in future if the rewards are to recur. In Mexico, on the other hand, once an individual teacher has secured the financial bonus (which is much larger in Mexico in annual terms than in Chile) she retains it for her entire working life. The evidence from Mexico is that this approach does not bring about sustained improvements in learner performance. In South Africa, it has been suggested that non-reversible movements up the salary scales for individual teachers be linked somehow to learner performance. The Mexican experience suggests that this would be difficult to achieve technically, and would not bring about the desired performance gains.

5 Concluding remarks

It is problematic when education policies are transposed from one country to another without due consideration to how contexts differ between countries. Equally, it is problematic when education policies in one country are designed without due consideration of important options explored and lessons learnt in other countries. This paper has tried to chart a path between these two extremes.

Because this paper has tried to describe several different government programmes, each programme has been described to only a limited degree of depth. More desktop analysis, however, would probably not take one much further as there is a limit to how well even the most effective governments document critically the design of their policies and programmes. In a way, the paper points towards countries and programmes that could be investigated further through other means such as e-mail enquiries, where links already exist, or through the establishment of new cross-country links with relevant government officials and researchers.

Finally, it is acknowledged that the bias towards studying developing countries in this paper has not been fully defended. There are undoubtedly important lessons to be learnt from developed countries too. However, given that policy analysts in South Africa have arguably focussed the bulk of their attention on developed countries (especially Anglophone ones), it seemed opportune to produce a paper with a different bias.

References

Abadzi, H., Crouch, L., Echegaray, M., Pasco, C. & Sampe, J. (2005). Monitoring basic skills acquisition through rapid learning assessments: A case study from Peru. *Prospects*, 35(2): 137-156.

Anand, P., Mizala, A. & Repetto, A. (2009). Using school scholarships to estimate the effect of private education on the academic achievement of low-income students in Chile. *Economics of Education Review*, 28: 370-381.

Benviste, L. (2000). Student assessment as a political construction: The case of Uruguay. *Education Policy Analysis Archives*, 8(32). Available from: <http://epaa.asu.edu/epaa/v8n32.html> [Accessed November 2009].

De Paiva Franco, A.M. & Menezes Filho, N.A. (2008). *Uma análise de rankings de escolas brasileiras com dados do SAEB*. São Paulo: Universidade de São Paulo. Available from: <http://ideas.repec.org/p/anp/en2008/200807211446380.html> [Accessed November 2009].

Delannoy, F. (2000). *Education reforms in Chile, 1980-1998: A lesson in pragmatism.* Washington: World Bank. Available from: <http://www1.worldbank.org/education/globaleducationreform/pdf/delannoy.pdf> [Accessed January 2006].

Fernandes, R. (2007). *Índice de Desenvolvimento da Educação Básica (Ideb)*. Brasília: Ministério da Educação. Available from: <http://www.publicacoes.inep.gov.br/detalhes.asp?pub=4121#> [Accessed November 2009].

Harbison, R.W. & Hanushek, E.A. (1992). *Educational performance of the poor: Lessons from rural Northeast Brazil*. Oxford: Oxford University Press.

Hanushek, E.A. & Woessman, L. (2007). *The role of school improvement in economic development.* Washington: National Bureau of Economic Research. Available from: <http://papers.nber.org/papers/w12832.pdf?new_window=1> [Accessed June 2007].

Horta Neto, J.L. (2006). *Avaliação externa: A utilização dos resultados do Saeb 2003 na gestão do sistema público de ensino fundamental no Distrito Federal*. Brasilia: Universidade de Brasília.

Kennedy, A.M., Mullis, I.V.S., Martin, M.O. & Trong, K.L. (2007). *PIRLS 2006 Encyclopedia.* Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed January 2008].

McEwan, P.J. & Santibañez, L. (2005). Teacher and principal incentives in Mexico. In Emiliana Vegas (ed.), *Incentives to improve teaching: Lessons from Latin America*. Washington: World Bank: 213-253.

McMeekin, R.W. (2000). *Implementing school-based merit awards: Chile's experience.* Washington: World Bank. Available from: <http://www1.worldbank.org/education/globaleducationreform/pdf/mcmeekin.pdf> [Accessed January 2006].

Mizala, A. & Romaguera, P. (2005). Teachers' salary structures and incentives in Chile. In Emiliana Vegas (ed.), *Incentives to improve teaching: Lessons from Latin America*. Washington: World Bank: 103-150.

Mullis, I.V.S., Martin, M.O. & Foy, P. (2008). *TIMSS 2007 international mathematics report.* Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed May 2009].

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J. & Chrostowski, S.J. (2004). *TIMSS 2003 international mathematics report*. Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed January 2008].

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J. & Gregory, K.D. (2000). *TIMSS 1999 international mathematics report.* Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed August 2005].

Mullis, I.V.S., Martin, M.O., Kennedy, A.M. & Foy, P. (2007). *PIRLS 2006 international report*. Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed April 2007].

Mullis, I.V.S., Martin, M.O., Olson, J.F. & Berger, D.R. (2008). *TIMSS 2007 Encyclopedia*. Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed May 2009].

OECD (2007). PISA 2006: *Science competencies for tomorrow's world*. Paris Available from: <http://www.oecd.org/dataoecd/30/17/39703267.pdf> [Accessed May 2008].

Ravela, P. (2005). A formative approach to national assessments: The case of Uruguay. *Prospects*, 35(1): 21-43.

Ravela, P. (2006). *Using national assessments to improve teaching... and learning: The experience of UMRE in Uruguay*. Washington: World Bank. Available from: <http://info.worldbank.org/etools/docs/library/235974/D2_cartagena_2006_PRavela_fv_11sep06.pdf> [Accessed November 2009].

Santibáñez, L., Martínez, J.F., Datar, A. & McEwan, P.J. (2007). *Breaking ground: Analysis of the assessment system and impact of Mexico's teacher incentive program "Carrera Magistral"*. Santa Monica: Rand Corporation. Available from: <http://www.rand.org/pubs/technical_reports/2007/RAND_TR458.pdf> [Accessed May 2009].

Schwartzman, S. (2003). *The challenges of education in Brazil. Oxford: University of Oxford Centre for Brazilian Studies*. Available from: <http://www.international.ac.uk/resources/The%20Challenges%20of%20Education%20in%20Brazil.pdf> [Accessed May 2009].

South Africa: Department of Education (2005). *Grade 6 Systemic Evaluation: National*. Pretoria. Available from: <http://www.hsrc.ac.za/research/output/outputDocuments/3580_Grade6National.pdf> [Accessed February 2006].

USAID (2006). *Success in implementing education policy dialogue in Peru*. Washington Available from: <https://www.eddataglobal.org/documents/index.cfm?fuseaction=showdir&pubtype=5&statusID=3&showtypes=0> [Accessed May 2009].

World Bank (2007). *Towards high-quality education in Peru: Standards, accountability and capacity building*. Washington.

World Bank. (2008). Prova Brasil: Building a framework to promote measureable progress in learning outcomes. *En breve*, 121. Available from: <http://siteresources.worldbank.org/INTENBREVE/Newsletters/21809358/Mar08_121_Prova_EN_FINAL.pdf> [Accessed November 2009].

Appendix

Countries are sorted in descending order according to their highest 'Largest change' value.

Table 1: Mean score improvement details

	ISO code	*Programme*	*1999-2001*	*2003*	*2006-2007*	*Largest change*
Morocco	MAR	TIMSS Gr 8	337	387	381	50
Morocco	MAR	TIMSS Gr 4		347	341	-6
Morocco	MAR	PIRLS lang	350		323	-27
Luxembourg	LUX	PISA math	446	493	490	47
Luxembourg	LUX	PISA lang	440	479	479	39
Liechtenstein	LIE	PISA lang	483	527	510	44
Liechtenstein	LIE	PISA math	514	536	525	23
Armenia	ARM	TIMSS Gr 4		456	500	44
Armenia	ARM	TIMSS Gr 8		478	499	21
Brazil	BRA	PISA math	333	356	370	37
Brazil	BRA	PISA lang	395	404	393	8
Russia	RUS	PIRLS lang	528		565	37
Russia	RUS	TIMSS Gr 4		532	544	12
Russia	RUS	PISA math	478	469	476	7
Russia	RUS	TIMSS Gr 8	526	508	512	4
Russia	RUS	PISA lang	462	442	440	-2
Syria	SYR	TIMSS Gr 8		359	395	36
Hong Kong	HKG	PIRLS lang	528		564	36
Hong Kong	HKG	TIMSS Gr 4		575	607	32
Hong Kong	HKG	PISA lang	525	509	536	27
Hong Kong	HKG	TIMSS Gr 8	582	586	572	4
Hong Kong	HKG	PISA math	560	549	547	-2
Ghana	GHA	TIMSS Gr 8		276	309	33
Philippines	PHL	TIMSS Gr 8	345	378		33
Chile	CHL	PISA lang	410		442	32
Chile	CHL	PISA math	383		411	28
Chile	CHL	TIMSS Gr 8	392	387		-5
Latvia	LVA	PISA lang	459	491	479	32
Latvia	LVA	PISA math	462	483	486	24
Latvia	LVA	TIMSS Gr 8	505	508		3
Latvia	LVA	TIMSS Gr 4		536	537	1
Latvia	LVA	PIRLS lang	545		541	-4
Korea	KOR	PISA lang	525	535	556	31
Korea	KOR	TIMSS Gr 8	587	589	597	10
Korea	KOR	PISA math	548	542	547	5
Indonesia	IDN	PISA math	366	360	391	31
Indonesia	IDN	PISA lang	371	381	393	22
Indonesia	IDN	TIMSS Gr 8	403	411	397	8
Singapore	SGP	PIRLS lang	528		558	30
Singapore	SGP	TIMSS Gr 4		594	599	5
Singapore	SGP	TIMSS Gr 8	604	605	593	1
Israel	ISR	TIMSS Gr 8	466	496	463	30
Israel	ISR	PISA math	434		442	8
Israel	ISR	PIRLS lang	509		512	3
Israel	ISR	PISA lang	451		439	-12
Poland	POL	PISA lang	480	496	508	28
Poland	POL	PISA math	471	490	495	24
Lithuania	LTU	TIMSS Gr 8	482	502	506	24
Lithuania	LTU	TIMSS Gr 4		534	530	-4
Lithuania	LTU	PIRLS lang	543		537	-6
Slovenia	SVN	TIMSS Gr 4		479	502	23
Slovenia	SVN	PIRLS lang	502		522	20
Slovenia	SVN	TIMSS Gr 8	530	493	501	8
Norway	NOR	TIMSS Gr 4		451	473	22
Norway	NOR	TIMSS Gr 8		461	469	8
Norway	NOR	PIRLS lang	499		498	-1
Norway	NOR	PISA math	499	495	490	-3
Norway	NOR	PISA lang	506	500	484	-6

	ISO code	*Programme*	*1999-2001*	*2003*	*2006-2007*	*Largest change*
Mexico	MEX	PISA math	387	385	406	21
Mexico	MEX	PISA lang	422	400	410	10
Czech Republic	CZE	PISA math	497	516	510	19
Czech Republic	CZE	PISA lang	492	488	483	-4
Czech Republic	CZE	TIMSS Gr 8	520		504	-16
Australia	AUS	TIMSS Gr 4		499	516	17
Australia	AUS	PISA lang	528	526	513	-2
Australia	AUS	PISA math	534	524	520	-4
Australia	AUS	TIMSS Gr 8	525	505	496	-9
England	ENG	TIMSS Gr 8	496	498	513	17
England	ENG	TIMSS Gr 4		531	541	10
England	ENG	PIRLS lang	553		539	-14
Lebanon	LBN	TIMSS Gr 8		433	449	16
Greece	GRC	PISA math	447	445	459	14
Greece	GRC	PISA lang	474	472	460	-3
Germany	DEU	PISA math	490	503	504	14
Germany	DEU	PISA lang	484	492	495	11
Germany	DEU	PIRLS lang	539		548	9
Iran	IRN	TIMSS Gr 4		389	402	13
Taiwan	TWN	TIMSS Gr 8	585	585	598	13
Taiwan	TWN	TIMSS Gr 4		564	576	12
Iran	IRN	PIRLS lang	414		421	7
Iran	IRN	TIMSS Gr 8	422	411	403	-8
Slovakia	SVK	PIRLS lang	518		531	13
Slovakia	SVK	TIMSS Gr 8	534	508		-26
Portugal	PRT	PISA math	453	466	466	13
Portugal	PRT	PISA lang	470	478	472	8
United States	USA	TIMSS Gr 4		518	529	11
United States	USA	TIMSS Gr 8	502	504	508	6
United States	USA	PIRLS lang	542		540	-2
United States	USA	PISA lang	504	495		-9
United States	USA	PISA math	493	483	474	-9
Finland	FIN	PISA math	537	544	548	11
Finland	FIN	PISA lang	547	543	547	4
Italy	ITA	PIRLS lang	541		551	10
Italy	ITA	PISA math	459	466	462	7
Italy	ITA	TIMSS Gr 8	479	484	480	5
Italy	ITA	TIMSS Gr 4		503	507	4
Italy	ITA	PISA lang	487	475	469	-6
Tunisia	TUN	TIMSS Gr 8	448	410	420	10
Tunisia	TUN	PISA math		359	365	6
Tunisia	TUN	PISA lang		374	380	6
Tunisia	TUN	TIMSS Gr 4		339	327	-12
Belgium	BEL	PISA math	519	529	520	10
Belgium	BEL	PISA lang	507	507	501	0
Belgium	BEL	TIMSS Gr 8	558	537		-21
Spain	ESP	PISA math	476	486	480	9
Spain	ESP	PISA lang	493	482	461	-11
Serbia	SRB	TIMSS Gr 8		477	486	9
Hungary	HUN	PIRLS lang	543		551	8
Hungary	HUN	PISA math	488	490	491	3
Hungary	HUN	PISA lang	480	482	482	2
Hungary	HUN	TIMSS Gr 8	532	529	517	-3
Hungary	HUN	TIMSS Gr 4		529	510	-19
Moldova	MDA	PIRLS lang	492		500	8
Moldova	MDA	TIMSS Gr 8	469	460		-9
Canada	CAN	PIRLS lang	544		550	6
Canada	CAN	PISA math	533	533	527	0
Canada	CAN	PISA lang	534	528	527	-1
Canada	CAN	TIMSS Gr 4		511	505	-6
Canada	CAN	TIMSS Gr 8	531		517	-14
Cyprus	CYP	TIMSS Gr 8	476	459	465	6
Turkey	TUR	PISA lang		442	447	5
Turkey	TUR	TIMSS Gr 8	429		432	3
Turkey	TUR	PISA math		424	424	0

	ISO code	*Programme*	*1999-2001*	*2003*	*2006-2007*	*Largest change*
Uruguay	URY	PISA math		422	427	5
Uruguay	URY	PISA lang		435	413	-22
Switzerland	CHE	PISA lang	495	499	499	4
Switzerland	CHE	PISA math	528	526	530	4
Scotland	SCO	TIMSS Gr 4		490	494	4
Scotland	SCO	PIRLS lang	528		527	-1
Scotland	SCO	TIMSS Gr 8		498	487	-11
Japan	JPN	T4		565	568	3
Japan	JPN	T8	579	570	570	0
Japan	JPN	Pr	522	499	498	-1
Japan	JPN	Pm	557	534	523	-11
Jordan	JOR	T8	428	424	427	3
New Zealand	NZL	T8	491	494		3
New Zealand	NZL	R	529		532	3
New Zealand	NZL	Pr	529	522	521	-1
New Zealand	NZL	T4		493	492	-1
New Zealand	NZL	Pm	538	524	522	-2
Romania	ROU	T8	472	475	461	3
Romania	ROU	Pm	426		415	-11
Romania	ROU	R	512		489	-23
Romania	ROU	Pr	427		396	-31
Denmark	DNK	Pr	497	491	494	3
Denmark	DNK	Pm	514	514	513	0
Ireland	IRL	Pr	527	516	517	1
Ireland	IRL	Pm	503	503	501	0
Macedonia	MKD	R	442		442	0
Macedonia	MKD	T8	447	435		-12
Austria	AUT	Pm	515	505	505	0
Austria	AUT	Pr	508	491	490	-1
Thailand	THA	Pm	433	417	417	0
Thailand	THA	Pr	431	420	417	-3
Thailand	THA	T8	467		441	-26
Sweden	SWE	Pm	510	510	502	0
Sweden	SWE	Pr	515	514	507	-1
Sweden	SWE	T8		499	491	-8
Sweden	SWE	R	561		549	-12
Iceland	ISL	Pm	515	515	506	0
Iceland	ISL	R	512		511	-1
Iceland	ISL	Pr	507	492	484	-8
Botswana	BWA	T8		366	364	-2
Macao	MAC	Pm		527	525	-2
Macao	MAC	Pr		498	492	-6
Bulgaria	BGR	R	550		547	-3
Bulgaria	BGR	T8	511	476	464	-12
Bulgaria	BGR	Pm	430		413	-17
Bulgaria	BGR	Pr	431		402	-29
Bahrain	BHR	T8		401	398	-3
Saudi Arabia	SAU	T8		332	329	-3
France	FRA	R	525		522	-3
France	FRA	Pm	518	511	496	-7
France	FRA	Pr	504	496	488	-8
Netherlands	NLD	T8	540	536		-4
Netherlands	NLD	T4		540	535	-5
Netherlands	NLD	R	554		547	-7
Netherlands	NLD	Pr	532	514	507	-7
Netherlands	NLD	Pm	563	538	531	-7
Argentina	ARG	Pm	387		381	-6
Argentina	ARG	Pr	417		374	-43
Malaysia	MYS	T8	519	508	474	-11
South Africa	ZAF	T8	275	264		-11
United Kingdom	GBR	Pr	524	507	495	-12
United Kingdom	GBR	Pm	530	508	495	-13
Egypt	EGY	T8		406	391	-15
Palestine	PSE	T8		390	367	-23

	ISO code	*Programme*	*1999-2001*	*2003*	*2006-2007*	*Largest change*
Kuwait	KWT	R	396		330	-66