Background paper to the study

National Testing of Pupils in Europe: Objectives, Organisation and Use of Results EACEA; Eurydice

THEORETICAL AND REAL EFFECTS OF STANDARDISED ASSESSMENT

Nathalie Mons

Senior lecturer in sciences of education at the University Pierre-Mendès-France (Grenoble 2, France, visiting fellow at the London Institute of Education (United-Kingdom)

August 2009

Eurydice Network

CONTENTS

INTRODUCTION

Standardised assessment is not a new phenomenon within the education systems of developed countries. Nevertheless, in the first decade of the 21st century, this tool, which combines the design, administration and marking of harmonised examinations, has prompted a great deal of debate both scientific and in the media (Haney, 2000; Hanushek and Raymond, 2003; House of Commons, 2007). If there is some debate about the instrument, it is because far from being a simple and neutral way of assessing pupils' skills and knowledge, it has become a key instrument for policy reform of education systems. **Whereas in the past, standardised assessment focused on measuring pupil attainment, nowadays its scope is much wider as it links pedagogy – its traditional stamping ground – and the policy for which it is now a pilot tool** (Behrens, 2006).

Standardised assessment is now at the intersection of new trends that have been shaping educational policies in OECD countries since the 1980s (Mons, 2007). In order to gain a clear understanding, standardised assessment needs to be viewed in relation to four recent developments in our education systems: a) **the emphasis on measuring quantitatively learning outcomes and the priority given to learning objectives over broader socialisation objectives** (Osborn, 2006) in conjunction with the notion of skills as defined in the economic theory of human capital; b) the development of **a new social supervision of teachers and schools by education administrations** in the broadest sense of the term (local districts, towns, decentralised authorities or regions, depending on the country), often in the context of decentralisation and school autonomy reforms (Maroy, 2008); c) **shifts in the balance of power** between central or federal authorities and local managers who thus have far less autonomy (Broadfoot, 2000) and lastly, d) **developments in schools' accountability to the general public**, and more specifically to parents, in the context of the new relationship between policy, the state and the administration on the one hand and civil society on the other. These new relationships have been shaped by the emergence of 'public democracy' in which the legitimate authorities do not have a monopoly on defining the common good (Manin, 1996).

Many stakeholders within the education system are resistant to these new concepts, with teachers being opposed to the new culture of measuring achievement, local authorities against interference by central or federal policy makers, and embattled schools defending their own interests against parental interference. As these factors have been converging, standardised assessment has become a political tool and therefore much questioned.

Building on the rhetoric of school effectiveness – the development of tests must facilitate improvements in the performance of education systems as a whole and student attainment in particular – this policy has to be assessed in the field in order to shed some scientific light on this lively public debate. **This report therefore aims to highlight the effects of standardised assessment.** The **first section of the report will consider the theoretical impacts of standardised assessment both as a monitoring tool in education systems and as a teaching tool used to improve individual pupils' performance.** What theoretical and conceptual constructs – both political and educational, since the tool has this dual role – have been formulated to explain how standardised assessment can improve the performance of education systems?

Moving beyond the various theoretical constructs, **the second part of our report will provide an empirical analysis of the *real* impact of standardised assessment in terms of effectiveness, educational equality and efficiency.** Because, in addition to the average improvement in pupil attainment – included in the concept of effectiveness – the impact this tool has on pupils from disadvantaged backgrounds (concerning social or ethnic factors) and those with disabilities should be examined also. Any evaluation of these policies must also consider cost-effectiveness. How much spending is required to implement the testing, and with what results? In order to shed some light on these questions, we will present a wide-ranging review of the scientific literature based on national case studies and international comparisons. **By examining this highly controversial scientific field, we will demonstrate that it is very hard to ascertain how standardised assessment influences both effectiveness and educational equality because there is no clear empirical consensus on the benefits of these reforms.**

This conclusion, or the lack of convergence between scientific studies, leads to question what processes can be identified regarding the reactions of local education stakeholders that could explain why standardised assessment appears to produce such a diversity of results ([1])? **In the third part of our report, we will therefore describe the mechanisms associated with the introduction of testing and the responses of various groups: teachers, both individually and as a body, the education system managers (head teachers, local and regional education officers), parents and the pupils themselves. We will examine how these groups react to the introduction of standardised assessment** and whether their responses vary according to the nature of the reforms undertaken.

([1]) In the first section, on the review of empirical research, we will show that divergences in the findings of research into the effects of the policy are also partly the result of applying different research methods.

I. PREDICTED EFFECTS: THEORETICAL POLITICAL AND PEDAGOGICAL FRAMEWORKS OF STANDARDISED ASSESSMENT

'The use of pupil performance in tests within accountability systems is not new. Examples of payment for results such as the flurry of performance contracting in the 1960s can be found cropping up and fading away over many decades. What is somewhat different about the current emphasis on performance-based accountability is its pervasiveness. As Elmore, Abelmann, and Fuhrman note, "What is new is an increasing emphasis on pupil performance as the touchstone for state governance" (1996, p. 65)' (Linn, 2000).

Standardised assessment is more than a pedagogical tool for measuring pupil attainment: it has acquired a new political status and become a mainstay of education system management. This dual political and educational role is underpinned by theoretical macro- and micro-backgrounds. **We will look first at the macro-political conceptual model that is based both on *New Public Management* and the school of thought of *Policy Evaluation* in the field of general public policy and, specific to the education sector, the economics of education and the trend of *school effectiveness*.** The amalgamation of these four theoretical and pragmatic schools of thought has generated specific macro-policy targets for standardised assessment. **Secondly, we present at micro-policy level of the school and classroom how testing is expected to function as a tool for guiding teachers' and pupils' activities.**

It is useful to consider the theory behind standardised assessment policies: these tools are often presented as a common sense approach, as measuring pupil attainment can be used to help them make progress. However, in the first section of this report, we will show that the testing approach is not based on common sense at all, but rather is tied to specific schools of thought associated with a particular intellectual, social and economic context, namely questioning state intervention and highlighting a 'crisis in state education', that may be genuine or 'manufactured' (Berliner and Biddle, 1995).

A. The political theory behind standardised assessment: a monitoring tool for education systems

The hypothetical functions and effects of standardised assessment reflect both the general trend towards public sector reform and specific schools of thought that have emerged within the education sector itself.

1. Standardised assessment and public sector reform: New Public Management and Policy Evaluation

New Public Management (NPM) is a school of thought that emerged in the United States and became established from the 1970s as a new way of analysing the way the public sector should operate in the context of challenging state action. At that particular time, various economic factors (stagflation, economic globalisation…) were undermining the Keynesian macroeconomic understanding of state intervention. Budgetary policy in the ministries of economy and finance was increasingly shaped by monetarist theories (Jobert 1994 and Siné 2006, quoted in Pons, 2008). The main aim was to streamline public spending by scaling back public sector activities.

The education sector did not emerge unscathed from the crisis, as resources came under pressure and the public sector's image was battered by the publication of reports exposing its failings (such as the 1983 report entitled *A Nation at risk* in the United States and the *Black papers* in the UK in the 1980s).

In response to this crisis situation, the New Public Management devised a series of guidelines which could be used to reform the administration by improving cost-effectiveness. In education, the guidelines included four key NPM principles that established a clear role for standardised assessment within the new organisational structures:

- quantifying public sector output;
- measuring output using scientifically proven customised tools;
- making public bodies with greater autonomy of action accountable to the system managers (management information model) and/or to citizens (democratic evaluation model);
- managing public organisations on the basis of output rather than solely through procedural checks based on resources (input).

Standardised assessment in education has a range of objectives and predicted effects. Firstly, the assessment must measure pupil attainment as an indication of the "quality" of the education service's output. It also serves as a link between those responsible for providing the service (teachers, schools and in some cases the local education authorities) and the administrators, often at national level, who define the service. Standardised assessment is therefore a management tool that both influences the actions of the implementing agents and provides information about their performance to their superiors (Woessmann, 2007). Lastly, in the context of greater accountability to civil society, standardised assessment is expected to provide information for people outside the school in general and for parents in particular.

In more general terms, standardised assessment also had a place in the Policy Evaluation approach (Spenlehauer 2003, quoted in Mons and Pons, 2009). Following the model used for scientific experiments, this pragmatic school of thought divided policy activities into three phases. Firstly, decision-makers have to define clearly the reference points for their policies. They can then select the measurement methods and relevant indicators in a second phase, and finally calculate the effects of a given policy by comparing the results. This three-step process involving planning, measuring and evaluating should underpin all policy evaluation.

Standardised assessment fits in neatly with the vision of sequential and rational public action – creating and implementing public policies by following a series of logical steps – since public policy targets can be achieved if the right resources are combined with clearly defined objectives. This school of thought reflects a top-down vision of public action in which local authorities implement public policies in accordance with the relevant legal framework (regulation or legislation). Standardised assessment is a key tool in the final phase of public policy evaluation and so there needs to be scope for comparing the initial aims and the end result.

Some researchers have raised concerns about the New Public Management and Policy Evaluation models. They question the strong focus on a results-based public policy approach and the emphasis on evaluating public service output, including within the education sector. As we will see later, studies of the process for implementing standardised assessment, in particular the behaviour of local stakeholders (teachers, head teachers, etc.), have shown that this theoretical top-down model is not perfect. Local stakeholders end up developing strategies to adapt to the institutional requirements that do not necessarily coincide with the anticipated results (for example, excluding struggling pupils from tests in order to boost the school's average marks artificially, or intensive training for tests whilst neglecting in-depth teaching).

In addition to public sector reform, standardised assessment policies are supported by two movements within the education sector itself.

2. The economic argument for standardised assessment and the pragmatic school effectiveness model

In accordance with the theory of human capital, economists no longer see education as a cost but rather as an investment that will produce additional capital. The combination of individual skills will reflect one component of a country's economic power.

The link between economic progress and pupil attainment has been highlighted by the American researcher Barro (1999) and others. Although some questions about the connection between education and economic growth remain unanswered, Barro has shown that using academic attainment to measure human capital is a better way to explain developments in national wealth than measuring school careers.

Standardised assessment is designed to address fundamental questions about the reality of the learning process. 'In the logic of an economy focused on "innovation" or "knowledge", we are increasingly told that we need to invest in human resources, which translates into demands to raise "general education levels" ' (Maroy, 2005).

But according to economists, the cost of better quality education must be contained, particularly as the generalisation of education with almost-universal access to upper secondary education in OECD countries is putting pressure on national budgets. In view of these budgetary constraints, economists emphasise their ability to provide solutions that will 'produce' more at a lower cost – better education with stable or lower budget resources – by defining institutional parameters to incorporate into policy to optimise the system. To this end, economic analysts use the production function model, which is borrowed from the business world, whereby companies combine production factors (inputs) in order to create a product (the output). When applied to education, the model translates as follows: the school combines different inputs (teaching materials, teachers with specific characteristics such as number of years of training, experience...) to produce education that can be measured in terms of quantity (years at school) or quality (pupil attainment). The economist therefore requires information about the inputs and the results of the education system to make the statistical models work and demonstrate which institutional structures are effective. Information about pupils' learning outcomes is provided by standardised assessments. Economists therefore advocate a new results-based approach to evaluating education systems (Woessmann, 2007).

The strategies and organisational structure chosen within a given school are also a key factor in student performance. Hence the familiar slogan 'school matters' adopted by the champions of school effectiveness. The quality of the leadership, disciplines in the school, pupil-teacher relations...are fundamental to academic success. Empirical studies by supporters of this approach have shown that schools with identical student populations in terms of the parents' socio-professional status can have very different attainment levels. This is why new standardised assessment policies are associated with the concept of making schools accountable both internally (to the relevant authorities) and externally (to civil society).

Those who advocate this economist approach to education do, however, stress its limitations. It is difficult to apply the conclusions of studies based on production functions. 'This model of technical change where science defines the way forward, and the "implementers" have merely to follow suit, struggles to translate to education. Of course, decision-makers can draw on research when deciding to eliminate a particular practice that has been proven to be disadvantageous (such as repeating a year or differentiated classes). But other than that, we know that directives handed down from on high tend to have limited impact, given the considerable autonomy that teachers have in the classroom … We also know that the effectiveness of an educational practice is often linked to context (effectiveness within a particular target group or particular education level) which means that assessments are necessarily tied to the conditions in which they were carried out' (Duru-Bellat and Jarousse, 2001).

So this review of the theoretical policy models appears to indicate that standardised assessment has a vital dual role: it is a management tool linking a number of educational policies which embody the far-reaching reforms undertaken in the vast majority of OECD countries (Mons, 2007) and also a measuring device that serves to evaluate those very reforms.

Effectively, in addition to its traditional function of measuring pupil attainment, standardised assessment is now seen as a multi-purpose tool. It is **an information-gathering device** generating comparable quantitative data that supports internal school accountability policies (supervision of schools by the education authorities) and external accountability to civil society (information about pupil attainment, particularly in relation to school choice policy). **As a tool for coordinating the actions of local participants at micro-policy level and decisions taken at macro-policy level**, standardised assessment can also be used to impose more effective learning principles in traditionally and newly-decentralised systems, rather than the content standards more commonly used. The development of testing therefore creates a new balance of power between schools, local education authorities and the decision-makers at central or federal level (Broadfoot, 2000). Consequently standardised assessment serves both as a **tool for determining cognitive content** (clear definition of compulsory priority teaching content) and **as a form of social control over the agents**. Lastly, the new standardised assessment policies, which nowadays tend to be based on school accountability (as opposed to the higher authorities responsible for education), are becoming established as a **means of transferring responsibility**: whereas in the past the pupil was seen as the primary agent in school achievement and policymakers had to justify any reforms, the emphasis is now on making schools accountable, in return for which they are given greater autonomy. **We can see from this multiplicity of roles that standardised assessment is expected to produce a variety of results in**

conjunction with policies on decentralisation, school autonomy and the freedom to choose a school.

In addition to the anticipated policy benefits, **standardised assessment is also expected to have an impact in the purely educational field. The spotlight has shifted away from managing education systems and is now focused on the school and the classroom.**

B. Overview of educational theory associated with standardised assessment: models are still being defined

Whilst other education policies such as school autonomy, decentralisation and school choice policy have prompted extensive, albeit much disputed, theoretical research, little work has been done on the educational theory underpinning standardised assessment, which focuses on the mechanisms within the school and the classroom. In terms of learning outcomes, the introduction of testing is designed to improve pupil attainment. Yet many authors (Linn, 2001; Nichols, 2007…) point out that theories seldom explain exactly which processes in the standardised assessment model are intended to boost pupil attainment.

A number of fairly vague arguments have been put forward by the advocates of standardised assessment. The explanations given vary and even contradict one another on certain points, depending on the accountability model: 'hard accountability' based on high-stakes testing as found in American and English schools or the 'soft accountability' model favoured in mainland Europe.

In the 'hard accountability' model (Goodwin, Eglert et al., 2002; Raymond and Hanushek, 2003; McDonnell, 2005; Haertel and Herman, 2005; Phelps, 2005; Woessmann, 2007) a series of penalties and rewards are tied to the test results which can have serious implications for schools, teachers and pupils (school funding, qualifications or repeating a year). **This model is based on the existence of the mechanisms described below, which are triggered by the tests and work together to improve the effectiveness of the education system:**

- The workload for **pupils** is expected to increase, resulting in better attainment through the combination of more clearly-defined academic targets and priority teaching content, and the psychological pressure generated by the test itself, which will determine the pupils' school career. In the hard accountability model, the possibility of repeating a year and of obtaining qualifications is directly linked to test results. Testing influences student behaviour by increasing the motivation to study, at least the extrinsic motivation ([2]).
- **Teachers** are held accountable for the pupils' achievements by their line managers and by civil society through the publication of the school's results, and will therefore work harder to ensure that pupils succeed, improving their professional skills through training and discussions with colleagues. Test results will be reviewed in each establishment. Student performance will be a boost for teaching staff in relatively high-scoring schools, whilst average or poor results will galvanise less-motivated teaching teams.

([2]) Unlike intrinsic motivation which comes from within the subject (e.g. pupils work because they like the subject), extrinsic motivation is generated by external factors (pupils will work for the test to obtain a qualification).

- Standardised assessment enables the **local administrative and political staff responsible for managing schools** to identify local problems and implement effective reward and penalty systems if a particular school is unable to improve its results within an agreed timeframe (these could involve a change in leadership, reorganisation or closure of schools).

In addition to improved effectiveness, some authors stress the potential benefits of standardised assessment for tackling educational inequalities (Grissmer et al., 2000; Hong and Youngs, 2008). By establishing common standards for all pupils, assessments force teachers to form common expectations irrespective of the pupil's individual situation (disability, ethnic minority, disadvantaged social background) and of the school environment in which they are progressing (schools with an advantaged or disadvantaged pupil population). This will serve to standardise education (exposure to same lesson content, same number of hours taught, identical learning targets) which will tend to limit social and overall educational inequalities and help to improve performances among pupils from disadvantaged social groups.

Furthermore, requiring schools to provide statistical reports broken down by social and ethnic groups will allow parents and decision-makers to identify whether certain schools are failing to address these issues. Teaching staff at schools in disadvantaged areas where pupils achieve relatively poor results under the new common standards will be motivated to improve. Local administrations will provide extra support for these institutions in the form of additional resources. Applying penalties to schools that are unable to improve their performance in the long term, with school closure as the ultimate sanction, will make it possible to eradicate educational establishments that seem doomed to failure for various reasons (mediocre teaching staff or ghetto schools).

In order to raise the stakes for struggling schools, a number of political stakeholders have rejected the notion of including information about pupils' social and ethnic backgrounds by calculating contextual added-value indicators ([3]) when publishing school statistical results. They argue that learning targets should be the same across the board.

In contrast to the high-stakes testing model used in the United Kingdom (England) and the United States, **the continental European model is associated with a philosophy of accountability which could be termed 'soft'** and is based on a number of premises that deviate from the arguments set out above. Here, we present the two main models, one of which has fed into the discussions on standards and assessment in Austria, Germany and Switzerland as described in the Klieme report (2004). The second model, known as the 'mirror effect' was developed in France by Thélot.

In his expert report 'On the development of national educational standards', Dr Eckhard Klieme, a researcher at the German Institute for International Education Research (DIPF), and his team recommend establishing 'output standards' (Klieme, 2004, p. 48). These results-based standards define the teaching objectives to be achieved and are validated using standardised assessment as opposed to content standards – input standards – which focus on the teaching content (similar to, but less detailed than, the concepts of curriculum, teaching programme and course of study). Klieme and his team claimed that schemes combining standards and testing had a key role to play: 'they highlight, in a clear and concise manner, the important learning features of our school

([3]) Statistics that calculate school performance by taking into account the socio-economic background of the pupils.

system. This guidance can be useful for pupils and parents alike, but it also serves to enhance teachers' skills and to improve quality within the school. The standards are embodied in the testing process, and are used to monitor education and evaluate schools … The aim of the tests is to analyse the effects (both primary and secondary) of the teaching methods and thereby to promote professional, rational action. Appraisals in the form of tests are therefore only useful if they contribute to improving the professional skills of the teaching staff, the quality of the school and the teaching it provides' (Klieme et al., 2004, p. 46).

The report goes on to explain how output standards can benefit each participant directly involved in the education process. For **pupils and their parents**, combining output standards with tests provides more information about priority subjects and is an ideal tool for facilitating dialogue with the teaching staff. However, unlike hard accountability models, in which both the school and the pupils are held accountable in tests that have serious implications for their school careers, the Klieme report explicitly rejects the use of results-based standards for determining individual academic progress. The information provided from these tests is not intended to be used to take decisions on whether pupils move up a school year or obtain a qualification.

For teachers, results-based standards are intended to guide their teaching ('Standards provide a frame of reference for teachers' (p. 49)) and indicate where responsibility lies (‘they emphasise the fact that teachers and pupils are responsible for learning outcomes' (p. 48)). However, in contrast to the hard accountability model, in which teaching staff are made accountable to the general public through the publication of school results, the Klieme approach recommends that 'feedback [should be] aimed at the teaching staff and school bodies, not the general public … We see measuring pupil attainment as an opportunity for schools to assess the results of their work and to respond in a professional manner' (p. 53).

In France, the “mirror effect” theory devised by Thélot (1994, 1998 and 2003, quoted in Pons, 2008), the former director of planning and statistics in the Education Ministry and ex-president of the HCEE ([4])**, also differs from the hard accountability approach. According to Thélot, standardised assessment should have a 'mirror effect'** (1998, quoted in Pons, 2008). This means that assessments should confront players in the education system with the results of their actions but need not necessarily provide them with explanations. Teaching staff need feedback on their methods to be able to make improvements if they have not achieved the intended aims. 'The mirror effect must be achieved, providing results without necessarily providing explanations, as these are not always available' (Thélot, 1998). The 'mirror effect' model is based on purely symbolic sanctions.

Both the hard and soft accountability approaches described above are based on a set of assumptions which some authors (Linn, 2000; Nichols, 2007) claim have either yet to be proven, or in some cases even run counter to existing empirical research. The underlying assumptions are:

- Tests are a good way of measuring the quality of the teaching provided by schools and pupils' actual skills and knowledge;

([4]) *Haut Conseil à l'Évaluation de l'Ecole* – High Council for Assessment in Education. This multi-stakeholder body, which was recently dissolved, served to monitor the progress of assessments at all levels of the French education system.

- The measurement, even if there is no weighting for contextual added-value, is not affected by differences in pupils' motivation, language skills, social status or ethnicity;
- Teachers and staff in schools are motivated by a system of rewards and penalties, and by external perceptions of their work, both parental and public; they seek to improve their teaching and have personal and shared resources for doing so;
- Test results help teaching staff to improve their teaching methods;
- Test results permit administrative educational staff to improve the management of the institutions in their charge;
- Schools can be held directly accountable for pupil attainment;
- Parents understand the significance of the tests and are able to interpret their child's results and those of the school as a whole. In systems where parents are free to choose their child's school, these indicators will be used to apply for the highest-achieving schools, thereby promoting healthy competition between schools, which in turn will tend to improve the performance of the education system as a whole.

As we will see subsequently, **a large body of research has shown that some of these assumptions, particularly those relating to teacher and parent behaviour, do not reflect the realities on the ground.**

Broadly speaking, as with the policy model described above, it appears either that the education theory underpinning standardised assessment requires further investigation or should be re-examined with regard to the aspects which run counter to the findings of certain empirical research.

In the first section of this report, we examined the processes through which standardised assessment is supposed to influence the regulation of education systems. In the second section, we will set out the actual empirical effects associated with these policies. In particular, we will examine whether testing measures are linked to higher average student performance and a reduction in social and overall educational inequalities. In our review of empirical literature, we thereby intend to give an overview of how standardised assessment influences effectiveness and educational equality. In the third section, we will also look at the processes brought into play by testing from the perspectives of teachers, pupils, education managers and parents.

In the two sections based on empirical data, we will refer to a wide range of quantitative and qualitative research from a wealth of sources (scientific articles, but also inspection reports, parliamentary hearings and public opinion polls designed to analyse behaviour patterns and stakeholder perception). Our review will not constitute a meta-analysis – using a common basis and clearly-defined criteria to compare a selection of studies and results from a range of research. Instead, we have chosen to present a broad selection of studies that have used a variety of methodologies. The other particularity of this literature review is the fact that it focuses on developed countries in Europe and North America.

II. THE REAL EFFECTS OF STANDARDISED ASSESSMENT ON THE EFFECTIVENESS OF EDUCATION SYSTEMS

Most of the literature analysing testing methods in terms of academic effectiveness and equality comes from North America, partly as a consequence of the vigorous public debate surrounding the reforms across the Atlantic. **Before looking at the studies that have examined standardised assessment policies and how they impact upon education system effectiveness, we felt it would be useful to describe two American case studies: the reforms in Texas and Chicago.** These experiences have generated a wealth of research which, despite using the same data, often came to contradictory conclusions. Both cases are doubly valuable as examples of empirical research into the impact of testing. Firstly, they demonstrate clearly the absence of empirical consensus on the impact of the measures, with conclusions varying considerably according to school year and subject. Secondly, the studies reveal a series of methodological pitfalls often encountered when analysing the results of testing and which render the evaluation invalid from a scientific point of view.

In the early 1990s, the state of Texas decided to make standardised assessment compulsory for pupils at the end of the fourth, eighth and tenth school year (Haney, 2000). **The Texas Assessment of Academic Skills (TAAS) was designed to make both schools and pupils accountable.** Schools were ranked in various categories according to their pupils' performance in the assessment: “exemplary”, “recognised”, “acceptable”, and “unacceptable”. These categories are then linked to rewards in the form of additional funding, or penalties which can extend to closing the school. Test results also have a significant impact on the pupil's school career, as they determine whether pupils have to repeat a year, as well as counting towards the high school diploma.

Early research into how this assessment programme impacted upon pupil attainment showed positive effects. In particular, the work by Grissmer and Flanagan (1998, 2001) showed substantial improvements in TAAS scores during the 1990s, both in terms of average scores and for different ethnic groups (white, Hispanic and African-American). For example, Hong and Youngs (2008) found that the percentage of pupils achieving the minimum standard for year 10 increased between 1994 and 2000. This was particularly noticeable among African-Black Americans, with the percentage of young pupils passing the test rising from 28 % in 1994 to 78 % in 2002.

The initial excitement was soon dampened by a series of studies that re-examined the effects of the reform using not only the TAAS data but also national test scores. While there appeared to be a significant improvement in test results for the local TAAS, if the results achieved by Texan schoolchildren in the federal examination – the NAEP ([5]) – were taken into account and a longer timeframe used, the progress was either noticeably less or insignificant, depending on the subject (Treisman and Fuller, 2001). **These findings therefore highlighted a few basic methodological rules that had to be applied in order to ensure a genuinely scientific evaluation of the standardised assessment programmes. Firstly, the effects of assessment should not be evaluated on the basis of local test results: local assessments cannot be used both as a management tool and a measure of their own effectiveness.** The divergence between the results in external and local examinations has been underscored by a wide range of studies (Nichols, 2007): local test results, particularly when associated with high stakes, tend to show substantial

([5]) National Assessment of Educational Progress.

improvements in the early years, largely as a result of intensive training for the test (teaching to the test) before hitting a ceiling – a subject we will come back to later. **Any evaluation of a testing model using only local assessment results will actually be a statistical demonstration of the phenomenon of teaching to the test, rather than an indication of whether this policy is effective. Another fundamental issue identified by this early research was that the effects of standardised assessment must be evaluated over the long term, as there are likely to be artificial effects in the early years.**

Research into the Texan example did not stop there. **The new method's remarkable impact on academic results among ethnic minorities prompted some authors to investigate how the test itself was implemented. They found that part of the 'Texan myth' (Haney, 2000) was explained by the fact that pupils with learning difficulties were excluded** (Haney, 2000; McNeil, 2005). In the American federal test (NAEP), states are specifically authorised to exclude pupils with learning disabilities who have an Individualized Education Plan (IEP) together with pupils of foreign origin whose English is poor. From 1992 to 1996, the exclusion rate in Texas rose from 8 % to 11 % in year four and from 7 % to 8 % in year eight, whilst the national figures fell from 8 % to 6 % and 7 % to 5 % respectively.

Qualitative analyses, such as the Booher-Jennings study (2005) also revealed changes in teachers' behaviour following the test's introduction: teachers tended to classify pupils into three groups – safe cases, suitable for treatment and hopeless cases – and focused primarily on the middle group, since any improvement in their results would boost the school's results in the short term. Meanwhile the weakest pupils received less attention. **On balance, the Texan experience raised questions about whether the benefits of standardised assessment were genuine, whether they were sustainable over time, the potential unintended consequences and the methodological shortcomings in some research designed to examine the impact of testing.**

A second case study, in Chicago, serves to highlight inconsistencies in results and problems arising from test content. In spring 1995, the Chicago district – the third largest district in the United States and home to a socially and ethnically underprivileged population – decided to put an end to automatic promotion. It was therefore decided that all pupils would be required to achieve a minimum standard in the Iowa Test of Basic Skills (ITBS) in their third, sixth and eighth school year in order to move up to the next school year (Roderick, Jacob and Bryck, 2002). Pupils who did not pass would be given extra support (summer camps, etc.) or would repeat the year if they were unable to improve their score despite the assistance provided. At the end of year eight, any pupils who failed the test twice were sent to 'transition centres'. **This policy had an immediate and significant impact: in the first two years, one-third of pupils in years three, six and eight repeated the year. After the initial shockwave, there was a significant improvement in results** (Roderick and Nagaoka, 2005). There was a marked drop in the number of pupils repeating the relevant school years. For example, in 1995, 37 % of pupils repeated year six, compared with 14 % in 1999. Jacob (2002) also found that there was a significant improvement in average ITBS results for all three school years between 1990 and 2000.

More detailed studies revealed once again that the advances were not as substantial as initial assessments appeared to indicate. Roderick, Jacob and Bryck (2002) demonstrated that progress varied considerably depending on the school year, the subjects tested and the

student population. For example, although the introduction of the test appeared to have benefits for pupils struggling with reading, in mathematics, the strongest pupils had an advantage. Jacob (2002) also showed that there was a very weak correlation between pupils' scores in the local mathematics test, the ITBS, and the Illinois Goals Assessment Programme, a test which focused more on thought processes. Good results in mathematics therefore seemed to be largely due to success in basic arithmetic and mental arithmetic exercises for which it was easier to train pupils. **As in Texas, qualitative analyses revealed that curricula had narrowed and little support was provided to pupils who were really struggling** (Lipman, 2004; Anagnostopoulos, 2006).

These two American case studies have been subjected to a great deal of scrutiny and reveal the difficulties inherent in evaluating testing: studies based on local tests are not reliable; the dates used for the study are a key factor; examining pupils' results in terms of absolute values or progress can alter the findings of the research. Generally speaking, the effects of standardised assessment appear to vary according to the subject and age group being tested and, as we will see in the meta-analysis below, there is no discernible pattern to these variations.

To extend our presentation beyond these two cases, we will now review empirical research on the subject, looking at national and regional models together with international comparisons.

As in the two case studies, research at national and regional levels ([6]) **has failed to establish a consensus.** Most of the literature comes from North America. **Some of these studies focused particularly on minimum competency tests.** Within this group, some research highlighted significant improvements in pupil attainment linked to the testing approach. This was the case in a pioneering study in the United States (Fredericksen, 1994) which looked at the NAEP examination for mathematics and showed that assessments based on minimum standards were associated with improved state averages in mathematics in the long term. However, Jacob (2001) reviewed the approach and found that data from another American national examination, the NELS ([7]), indicated that this policy was not related to higher mathematics and reading scores in year 12. Bishop et al. (2001) produced more conflicting conclusions. Minimum competency tests produce results that bear little relation to pupils' attainment unless the examination is closely linked to the school curriculum.

Other national studies looked at broader accountability issues rather than minimum competency tests. In the 1990s, many countries moved away from assessments based on minimum standards in favour of proficiency testing, which looks at a wider range of learning. Once again, research and even re-evaluations of similar data produced very varied results. Amrein and Berliner (2002) produced a chronological overview of the effects of new accountability policies developed in the United States by various states in the 1990s. They sought to establish links between the introduction of testing and potential improvements in pupil attainment as measured by the nationwide NAEP. The research revealed inconsistencies in the results for mathematics and reading in years four and eight: in some cases, the reforms were linked to performance improvements, whilst others resulted in a decline. Rosenshine (2003) examined the same data using a new methodological

([6]) In view of the methodological issues highlighted previously, we have only considered analyses based on external examinations rather than on local testing which is part of accountability policies.

([7]) National Educational Longitudinal Study.

approach and showed that state NAEP averages rose more significantly in states with high-stakes testing. Amrein-Beardsley and Berliner (2003) used the same methodology and data as Rosenshine but included the NAEP exclusion rates for pupils with learning difficulties. Like Rosenshine, they revealed that states using high-stakes testing saw results improve in year 4. However, if NAEP exclusion rates were used as a control variable, the effect was not significant statistically.

Other American studies went on to examine the development of accountability testing in conjunction with the No Child Left Behind Act of 2002 [8]. Rather than accepting the existence of two alternatives (tests or no tests), a number of studies created continuous variables in order to demonstrate the impact of test-based accountability systems (more or less serious rewards and penalties). Carnoy and Loeb (2002) obtained conflicting results using this approach. They showed that significant improvements in the American federal NAEP mathematics examinations between 1996 and 2000 in grade eight pupils from different dominant and minority ethnic groups (white, Hispanic and African-American) were linked to greater accountability. However, the grade four results were not so closely related to the degree of accountability among pupils from ethnic minorities. Among white pupils, there was no correlation at all, and nor could the researchers find any statistical correlation between the accountability system and changes in the student retention rates in grade nine or in the high school completion rates. In another paper based on an index of accountability, Hanushek and Raymond (2005) conversely demonstrated that apparent improvements in NAEP performance between grades four and eight were in fact closely related to the reforms and the length of application.

Nichols, Glass and Berliner (2006) picked up the idea of a scale of accountability by building a model using the 'Assessment Pressure Rating' which drew on a broad range of policy reports from 25 American states. Again, the findings were ambiguous. The study revealed a strong correlation between the assessment pressure rating and NAEP mathematics scores in grade four. The greater the assessment pressure, the better the results. Nevertheless, although a number of correlations were identified in relation to the year eight mathematics examinations, the findings were inconclusive: some correlations were positive whilst others were negative. The link between the assessment pressure rating and reading scores in grades four and eight was also very weak. Any significant correlations were negative, which suggested that standardised assessment methods could have a negative impact on pupil reading attainment, particularly in grade four. Overall, **there did not appear to be any consistently strong or regular links between effectiveness and the testing programmes.**

International studies on the subject reveal similar inconsistencies. Some international research highlights the benefits of testing, such as several studies by Woessmann (summarised in Woessmann, 2007). Building on a series of international standardised assessments for secondary pupils (TIMSS [9] 1995, TIMSS Repeat 1999, PISA [10] 2000) and using a multi-level statistical

---

([8]) Adopted in January 2002 with cross-party support from the Democrats and Republicans, the No Child Left Behind Act (NCLB) established a series of high-level objectives for American states with responsibility for education within the federal system. The objectives related to pupil attainment, the final high school diploma grade (end of upper secondary) and the quality of the teachers recruited. Individual state's efforts and performance are monitored by establishing standardised assessments for mathematics and reading in grades three and eight plus a minimum of three science examinations during a student's school career. Schools that fail to improve pupil attainment year-on-year receive targeted technical support and are permanently closed if the teaching staff are unable to reverse the trend after receiving assistance.

([9]) Trends in International Mathematics and Science Study.

([10]) Programme for International Student Assessment.

treatment that could take into account information about individual pupils – such as social background – the German economist revealed that the best secondary school performances were linked to the existence of an external final examination. Using data from PIRLS ([11]), Woessmann was able to show that testing could also prove beneficial at primary level. The benefits of accountability were enhanced when the testing was developed in conjunction with school autonomy.

In contrast, Mons (2007) used some of the same data (PISA 2000), but looked at national data to use control variables reflecting the national context (e.g. level of economic development). Mons showed that there was no link between the presence of accountability mechanisms and performance indicators. Standardised assessment policies in compulsory education were examined using several variables created by the author. One variable made it possible to identify the existence of centralised national examinations, either those for which certificates were awarded (mainly at the end of lower secondary education) or standardised assessments (for which a large group was required). This standardised centralised assessment contrasts with local assessments set by local authorities or individual schools. The research indicated that, when controlling for level of economic development, the existence of examinations or centralised tests bore no relation to average pupil attainment, nor to the percentage of pupils in the highest-scoring group (level five in the PISA study), nor to the percentage of children with learning difficulties (PISA level one). However, if the model did not include per capita GDP as a control variable, Woessmann's conclusions (2007) about the link between accountability and student performance resurfaced. So it would appear that the conclusions of the international research described above can be ascribed in part to the lack of control variables reflecting the country's economic development. Ultimately, standardised assessment has largely been developed in the richest OECD countries. Good effectiveness scores could therefore be attributed to the high level of economic development in certain countries which in turn allows them to organise expensive national testing schemes, rather than to the examinations themselves.

Moreover, the most recent PISA report (2007) failed to establish any clear links between testing measures and effectiveness. 'How do accountability policies and practices relate to the performances of countries? This is difficult to answer, most notably because these policies and practices are often closely interrelated with other school policies and practices' (OECD, 2007, p. 243). In response to this question, the OECD has developed multi-level models that incorporate pupils' socio-economic background as well as different features of the accountability models (links to standards, whether results are published, whether information is used to assess teachers, etc.). The results have been mixed. Standards-based evaluations tend to be associated with good national science scores in PISA 2006, but the correlation evaporates if information about demographic and socio-economic background is included. Only organisations publishing details of school performance are associated with higher results. 'For the other aspects of accountability policies, as measured by PISA, the relationships with performance are weaker and are not statistically significant' (OECD, 2007, p. 243).

Generally, in the literature on national situations or containing international comparisons, the relationship between effectiveness and testing appears to be unpredictable: there is no automatic, one-to-one correlation.

([11]) Progress in International Reading Literacy Study.

Is standardised assessment more associated with reducing educational inequalities among different social and ethnic groups as some of its supporters claim? **Once again, a wealth of empirical literature, most of it based on national North American systems, has failed to reach a consensus.** As we explained earlier, a large amount of research sprouted from the cases in Texas and Chicago and highlighted the contradictory results of testing for socially-disadvantaged pupils. The aforementioned study by Carnoy and Loeb (2002) revealed that testing in certain subjects could have benefits for young Americans from ethnic minorities. The conclusions of research by Hanushek and Raymond (2005) were contradictory. Although the research did find an improvement in average scores, it showed that accountability did not usually reduce divergences in performance among white and African-American schoolchildren. Accountability policies appeared to have more benefits for the Hispanic population. Meanwhile, research by Lee and Wong (2004) and Nichols et al. (2006) indicated that accountability reforms did not have any significant benefits for pupils from ethnic minorities. A report by the OECD (2007) based on PISA 2006 also failed to identify any links between the different forms of accountability covered by the investigation and educational inequalities related to social factors.

Only one accountability model – external national examinations at secondary level – appears to be associated with fewer inequalities. This conclusion is reiterated in a whole range of national studies (Harris and Herrington, 2006). An international comparison looked at the relationship between this accountability model and social inequalities at school (Mons, 2007), in which the author created a variable relating to external national examinations. As stated earlier, if this variable was not linked to performance indicators, centralised examination systems appeared to be associated with a lower level of educational inequality related to social factors. Standardised assessment could therefore potentially limit social reproduction phenomena within schools, not least because it harmonised the academic requirements for teachers, thereby preventing schools with pupils from disadvantaged areas from deviating from the standard curriculum. The standardised assessment approach implicitly assumes that targets are identical for all pupils capable of achieving a given level of education, as they all sit the same examination.

Bishop (2006) went further by defining the criteria that he felt would make external national examinations more effective. We can assume that some of these criteria would also contribute to reducing social inequalities. **Bishop (2006) argues that national examination certificates can have benefits if they are directly related to the curriculum and external standards, if they 'measure the full range and signal multiple levels of attainment' and lastly, if they provide broad coverage for a given age group. Harris and Herrington (2006) also stressed that external examinations could be beneficial not because they put pupils and teachers under psychological pressure but because they constituted an opportunity to increase the time that pupils spent on teaching content, both through additional teaching time and the emphasis on content.**

In view of the broad spectrum of research into the effects of standardised assessment and the resulting contradictory conclusions, Lee (2008) produced **a meta-analysis to compare and summarise the findings of these investigations in an attempt to identify some broad principles.** To this end, Lee selected 14 studies that met a number of specific criteria, many of which are cited in this report. To limit the teaching to test effects, the research had to refer to results for mathematics

and/or reading for non-local tests that could be compared with national results and were low-stakes tests (such as for NAEP). Lee's findings were disappointing. The average effect size from all estimates in the models used for these studies was certainly positive at the outset. In other words, the meta-analysis of these studies seemed to indicate that testing improved student performance. However, aside from the fact that the average actually masked considerable discrepancies in the research findings, the average effect disappeared as soon as adjustments were made to allow for duplication (several studies used the same data). Similarly, Lee tried to demonstrate that the effects of testing could be associated with a particular discipline, a school year or a period during which a policy was implemented. For example, some studies have shown that accountability has a greater impact on mathematics than on reading, on primary (as opposed to secondary) school pupils and over a long period. Once again, the meta-analysis did not reveal any significant impact. Lee's final question was whether standardised assessment reduced educational inequalities. Yet again, the findings were inconclusive.

While analysing effectiveness and educational equality, we can also examine the **efficiency of standardised assessment policies**, given that these reforms have been developed in accordance with economic rhetoric. The prospect was to improve pupil attainment at a lower price (Linn, 2000). Surprisingly, there has been very little empirical research regarding this issue (Behrens, 2006). Only the American economist Hoxby (2004) has tried to assess the financial cost of certain American schemes to show the low level of investment in these policies. Elsewhere, a few vaguely-related indicators are the best thing available. For example, the cost of developing, managing, marking and reporting on the Florida test programme is estimated at USD 42 million per annum (Florida Department of Education, 2003, quoted in Jones, 2007). In England, the most recent House of Commons report on testing (2007) also provides some information about the national cost of accountability schemes: each pupil sits an average of 70 tests during their school career, requiring the services of 54 000 examiners and moderators each year. 68 % of primary schools have extra dedicated staff for the tests.

Overall, whether we look at national or international research, research focusing on the relationship between testing and effectiveness or concentrating on the links to educational inequalities, there appears to be no empirical consensus on the benefits of standardised assessment. Nichols (2007) and Lee (2008) suggest that the divergences in the findings of different investigations should prompt further, more detailed research into the models used. We could contend that it is the diversity of the policies that has resulted in such divergent conclusions. Thus, for certain institutional structures, testing could prove beneficial whilst in other systems the unintended effects might be more apparent. In essence, it all depends on the implementation processes used, which in turn are related to the actions of the different stakeholders involved in the reform. We will examine this component in the third section of our report.

III. STANDARDISED ASSESSMENT AND EDUCATIONAL PROCESSES: HOW TEACHERS, MIDDLE MANAGEMENT, PUPILS AND PARENTS RESPOND TO TESTING

A great deal of European and North American literature has been produced on how tests affect the behaviour of the different groups within education systems. Jones (2007) pointed out that whilst the benefits of standardised assessment have been clearly identified, there is now a wealth of empirical literature from studies which have shown that some external accountability models, particularly those with high-stakes testing, can generate unintended outcomes. We will now review research into the processes associated with standardised assessment, making a slightly artificial distinction between the reactions of teachers, education system managers, pupils, and parents to these tools.

A. Teachers: a tendency to resist the culture of quantitative standardised assessment

Although empirical literature indicates that testing can have positive effects for teachers, it also highlights marked changes in teaching practices (emphasis on teaching to the test, narrowing of curricula…) that call into question the benefits of the model and help to explain the profession's resistance to the culture of standardised assessment in many countries.

Standardised assessment programmes can benefit teaching activities. Studies in several countries have shown that teachers support the principle of performance standards in education. A number of trade union organisations have come out in favour of the reforms. For example, in the United States, the American Federation of Teachers (AFT, 2001, quoted in Behrens, 2006) declares that the 'AFT was an early advocate for standards-based education. In 1992, in response to national concerns that pupils in the United States were not learning enough to compete in a global economy and that there was an intolerable gap between the achievement of whites and blacks, the then AFT president urged states to learn from other high-achieving countries and set clear and rigorous academic standards for all pupils … Standards-based reform as articulated by the AFT is an ordered process that includes well-developed standards and a curriculum to support their implementation; professional development for teachers; new assessments aligned to the standards; and fair incentives and sufficient resources to help pupils make the grade' (p.9).

Surveys of teachers also revealed that teaching staff in some countries support performance-based assessments. According to Johnson and Duffett (2003), despite their reservations about the implementation methods (which we will come to later), 80 % of teachers in the United States felt standards were a useful guide for identifying teaching requirements and improving student performance. 87 % of teachers were also of the opinion that pupils should sit tests in order to move up a school year, and that pupils who failed the test should be sent to remedial summer camps or repeat the year if their results did not improve after receiving remedial assistance. According to Jones (2007), surveys conducted in several American states (including Florida and Ohio) also showed that teachers took a positive view of testing: they saw tests as a means of improving the structure of teaching content for each school year, providing a real-life context to standards that may otherwise not be implemented, allowing teachers to identify pupils' weak points, and promoting a results-based culture among teachers.

In Sweden, a survey by the National Agency for Education (2004) also revealed that the majority of teachers were in favour of the national tests in their country. The vast majority of teachers claimed that standardised assessment provided clear guidelines on teaching content, helped to highlight pupils' strengths and weaknesses and did not limit the scope of their teaching. The tool was also appreciated because it provided a national framework for teaching content in a system that is now highly decentralised and in which local authorities have a significant role in determining teaching activities, which could potentially give rise to inequalities between regions and schools. A study by the Norwegian researchers Helgøy and Homme (2007) comparing the situation in Norway and Sweden found similar results for Sweden. However, the study revealed that Norwegian teachers were not as convinced, which would seem to indicate that teaching professionals' views on testing are determined by contextual factors.

In the United Kingdom (England), as in the United States and Sweden, despite current vocal objections to the SATs tests, teachers' representatives do not dispute the principle of standardised assessment (House of Commons, 2007). Testing that focuses on a culture of evaluation and the development of learning objectives also seems to have a clear influence on teaching practices. For example, in a survey conducted as part of a comparative study of French and English primary schools, Broadfoot, Osborn, Sharpe and Planel (2001) found that teachers in the United Kingdom (England) ranked 'assessment skills, 'subject knowledge' and 'clear aims' at the top of their education priorities ([12]). The researchers felt that this shift in teachers' priorities was linked to the introduction of the National Curriculum in 1988 and the development of the standardised assessments that were introduced in the early 1990s. Hargreaves (2002) argued that SATs tests now provide teachers with a clear picture of the requirements at different 'key stages' in a student's school career.

The guidance provided by standardised assessment in France has also been analysed: 'One way of influencing teaching programmes is to work on national tests and examinations. For example, after a few years of conducting diagnostic tests in geometry at the start of year six, this neglected area of mathematics has been revitalised' (IGEN-IGAENR, 2005). The harmonising effect of standardised tests appears to be even stronger in France where there are no stakes associated with the diagnostic tests. More generally 'there has been an increase in subject-specific monitoring in recent years through new forms of testing … Naturally, these tests require an adjustment period … But the effects are soon apparent' (IGEN-IGAENR, 2005).

Demailly (2001) also put forward a positive view of the impact of standardised assessment on teaching: 'the formative benefits of assessment are not insignificant, nor is the decompartmentalisation of professional cultures and the emergence of cooperation in interprofessional interactions. Discussing the standards, identifying relevant indicators to describe how testing works, or sometimes painful practice reviews are all opportunities to clarify and enhance certain professional skills'.

On the whole, teachers seem to respond positively to standards and their associated tests and, in certain institutional structures, the actual effects of these measures on teaching activities. Essentially, the reforms provide clear guidelines for implementing curricula, preventing the emergence of pronounced inequalities when developing a local syllabus, focusing on the

([12]) The lowest priorities were: 'relations with children' and 'maintaining order'.

actual performance of pupils, particularly those from disadvantaged backgrounds, and promoting teamwork built around an analysis of the test results.

Nevertheless, this openness to the *principle* of standardised assessment in some countries does not, however, prevent the teaching profession from criticising specific programmes that are seen to be negative because the skills tested are too archaic, because they fail to take into account the social characteristics of the student population or to make the link between student performance and teachers' rewards. These reservations are particularly prevalent in the United States, as evidenced by a recent survey in which 70 % of teachers say that there are too many tests [(13)] and in the United Kingdom (England), where teachers' unions have recently called for a boycott of the national tests, In France too, the teachers' trade unions have spoken vehemently against the new tests that were introduced in 2009. According to a recent inspection report, only 70% of schools have agreed to transmit the results of the test in year 5 to the Ministry, even though the test is compulsory, and 85% of schools for the test in year 2.

Demailly (2001) observes that three conditions need to be met when defining and implementing tests in order for the assessments to be viewed positively: 1) the evaluation must be developed through participation, with extensive teacher involvement, 2) the assessment objectives must be democratic (as opposed to authoritarian) and 3) the project's sponsors must be able to persuade teachers and display a firm resolve.

As these conditions were not present in all countries and test programmes, a very rich and extensive body of literature has examined different national contexts, focusing on **the unintended consequences of standardised assessment for teaching activities. This research has shown clearly that tests can result in the erosion of teaching standards, particularly when associated with high stakes, and in certain circumstances, to a sense of deprofessionalisation, leading to poor motivation among teachers.**

Indeed, as we have indicated in the cases of Texas and Chicago, **certain standardised assessment programmes primarily result in the much-studied phenomenon of teaching to the test** (Gordon and Reese, 1997; Jones and Egley, 2004; Bélair, 2005; Jones, 2007). Given the imperatives of the results, some teachers now devote a large amount of teaching time to coaching, using exercises similar to those that will appear in the tests. This new behaviour pattern has been the focus of studies in the United States (Jones, 2007) and the United Kingdom (England), both of which use high-stakes testing. According to the parliamentary report on 'Testing and assessment' (House of Commons, 2007), a study by the Royal Society in 2003 indicated that there were large variations in the amount of time devoted to tests in the United Kingdom (England) – where there is extensive testing – and Scotland, which has developed a more flexible approach. For example, secondary school teachers in England spent twice as much time on evaluation activities as their Scottish counterparts. The same report calculated that in the spring term, 70 % of primary schools spent three hours a week on teaching to the Key Stage Two test taken in year six.

In addition to intensive teaching to the test, **standardised assessment models can also result in what has been summed up as curriculum narrowing** (Behrens, 2006; Jones, 2007; House of

(13) Opinion poll on education carried out by the American Public Agenda organisation as part of its Reality Check 2006 series. The 2006 series is available online at: http://www.publicagenda.com/files/pdf/rc0603.pdf

Commons, 2007). This subset of teaching practices can take several forms. Firstly, it can entail **a reduction in the range of subjects taught.** As the tests generally concentrate on a limited number of subjects, teachers, particularly in the primary sector, tend to dedicate less teaching time and accord less importance to subjects that are not tested. Consequently, some empirical studies in the United Kingdom (England) and the United States have shown a considerable reduction in the time spent on social sciences, arts and sport because of standardised tests (Jones, Jones and Hargrove, 2003; House of Commons, 2007). The narrowing of the curriculum may also result in **teachers concentrating on the skills tested which tend to be fairly basic, whereas more complex skills such as problem-solving are rarely assessed.** In addition to their impact on teaching content, **standardised assessments tend to make teachers focus on pure learning objectives (Osborn, 2006; Jones, 2007) thereby detracting from other skills learned at school (social skills, developing creativity, independence, and citizenship)**.

In addition to shaping content and learning objectives, **standardised assessment can lead to changes in pedagogical methods themselves**. **Testing covers a wide spectrum of knowledge that pupils are now expected to assimilate in a limited time, which can prompt teachers to focus on teaching methods that build on rapid rote-learning rather than a more time-consuming active exploration of a subject** (Gordon and Reese, 1997). Perception of what teaching entails will also evolve. For example, a comparative study of the United Kingdom (England), Denmark and France (Osborn, 2006) showed that with the new focus on standardised assessment, teachers in England now see transmitting knowledge and skills as a priority that takes precedence over pastoral care.

High-stakes testing, and in particular the associated publication of results, also has implications for the way teachers view pupils, how attention is focused on certain pupils, and the nature of the pupils attending the school. As we saw earlier from the experiences in Texas and Chicago, teachers can end up pigeonholing their pupils (brilliant, able to pass with assistance, permanently failing). This classification can then cause pupils with serious learning difficulties to be isolated because they will be unable to pass the test straight away, even with support, and therefore will prevent the school from improving its performance. In the United Kingdom (England), Levacic (2001) showed that the most widely-publicised school performance indicator – GCSE-1 ([14]) – tends to influence teachers' actions in the light of fierce competition between schools. Van Zanten (1999) came to the same conclusion in a comparison of the French and English systems, and established a link between the choice of school, standardised assessment and mechanisms for selecting pupils: 'whilst head teachers and teachers tend at all times to look for "good customers", this trend becomes far more pronounced in the context of competition. In this environment, those schools that are able to be selective become even more so in order to choose pupils who will enhance their image by providing added-value: pupils whose academic attainments will contribute to the image of an effective school that ranks highly in national assessments … but also pupils whose manners, language and conduct

([14]) The General Certificate in Secondary Education (GCSE) is a nationwide external examination that English schoolchildren take at the age of 15-16, at the end of their eleventh school year. It marks the end of compulsory education. Most pupils choose to take examinations in eight to ten subjects. The national results are published in a breakdown by school. They are interpreted using two indicators. The GCSE-1 is the most widely-publicised in the form of league tables: the tables show the number of pupils who achieved grades A-C (top three grades) in at least five subjects. GCSE-2 reflects each school's overall performance but receives less coverage as it gives no indication of a school's ability to produce 'good pupils': the tables show the number of pupils who achieved grades A-G in at least five subjects. G is the lowest pass grade.

will serve to indicate the school's social calibre' (p. 145). Once good pupils have been selected, they receive special treatment. In some English schools, funding and teaching will be diverted, often against the teachers' wishes, towards activities targeting the most gifted pupils (van Zanten, 1999). It was also stressed that average pupils whose academic progress would improve school performance were also targeted by specific educational activities.

In addition to the new emphasis on certain student groups, **external hard accountability models can also push some teaching staff to use underhand methods to boost the school's results.** Referring to the Canadian state of Ontario, Bélair (2005) asserts that 'in certain cases cited by teachers, schools even went so far as to identify pupils on special programmes in order to reduce the number of pupils taking the test and increase pass rates'. As we have seen, the same thing occurred in Texas, where some pupils with serious learning difficulties were excluded from the federal NAEP. In 2006 in the Netherlands, a country with high-stakes testing, inspectors investigated exclusion practices in certain schools in response to rumours about the tests at the end of primary school. It emerged that in some cases, pupils who were most likely to be directed to the least prestigious school track – *Leerwegondersteunend onderwijs* (learning support) ([15]) – did not sit the test. Once again, the teachers' actions were designed to make the school's performance appear better than it really was.

These new trends, which run counter to teachers' professional standards, and in particular to their perceived educational role, resulted in a **loss of motivation among teaching staff** (DeBard and Kubow, 2002; Center on Education Policy, 2006; Jones, 2007). In some cases, **teachers developed a negative image of the profession, their job satisfaction decreased and a new kind of stress was perceived.** For example, a survey in North Carolina revealed that 84 % of teachers felt that their job had become more stressful since the introduction of high-stakes testing (quoted in Hargrove et al., 2004). Standardised assessment is also believed to be one reason for teachers leaving the profession. For example, in the study by Hoffman, Assaf and Paris (2001), 85 % of American teachers claimed that the best teachers were leaving the profession because of high-stakes testing. Other studies in the United States have shown that a significant proportion of teachers who want to remain in the profession have asked not to teach the school years that are subject to testing (Tobin and Ave, 2006, quoted in Jones, 2007).

This loss of motivation seems to be particularly marked in schools with disadvantaged pupils and relatively poor overall results, because the pupils' social background is not taken into account (Jones, 2007). Trade unions, particularly in the United States and the United Kingdom (England), have repeatedly called for added-value indicators to be defined to measure a school's actual educational output (Behrens, 2006). Teachers also claim that these schools have problems recruiting good quality teaching staff. Jones (2007) emphasised the need for research to measure the reality of recruitment problems and attrition rates as well as studies on the image of teachers.

Some researchers have attributed the loss of motivation among teachers to a deep-seated sense of deprofessionalisation. This is the theory advanced by the Belgian researchers Maroy and Cattonar (2002), who stressed that corporatist arguments alone did not explain why teachers were opposed to certain testing models. The theory was supported by Osborn in the United Kingdom (England) (2006) and Dupriez (2005) who compared the English and Belgian situations. Up to now,

([15]) The report is available in Dutch at http://www.owinsp.nl/nl/home/naslag/Alle_publicaties/Eindtoets_po

teachers have been an integral part of what organisational sociologists call professional bureaucracies (Bidwell, 1965, quoted in Maroy and Cattonar, 2002). This is a hybrid organisational model combining rigid, impersonal but rational bureaucratic rules (as defined by Weber (1922)) with a freedom of action based on solid recognition for highly-qualified professionals. By virtue of this recognition, sociologists consider teaching to be one of 'the professions' in the English usage of the term, akin to the French concept of *profession libérale*. Indeed, although teachers are bound by very strict rules (fixed school structure, national curricula in many countries, presence of a supervisor monitoring and assessing their work), they also have a great deal of freedom in terms of how they teach and in day-to-day activities in the classroom because their professional skills are recognised. It is this freedom and the resulting semi-professional status of teachers that is called into question by the development of standardised tests. The teacher is no longer seen as the only person able to make what are often final judgments with significant consequences about 'their' pupils' attainments.

According to Maroy and Cattonar (2002) 'the recent reforms affect the curriculum (more centralised and redefined with a skills-based approach) and the evaluation methods (with the arrival of a series of standardised tests), they are more prescriptive and no longer fully under the teachers' control, becoming instead the product of external actors, limiting the traditional sphere of activity associated with the group. We could call this a kind of "dequalification" (Lessard, 1999)'.

On the whole, teachers appear to have mixed views on standardised assessment. They accept the general principle but criticise high-stakes mechanisms that have too much influence on educational approaches.

B. Education system supervisors: getting to grips with the tool

To date, little has been written about how managers supervising the education service view standardised assessment, but surveys in several countries indicate **that this professional group has fewer doubts than teachers about the principles and the mechanics of testing. Nonetheless, questions do emerge in discussions on the subject.**

For example, in the United States, Farkas, Johnson et al. (2003) revealed that the vast majority of managers, particularly head teachers and superintendents ([16])**, saw standardised assessment as a good thing:** 'Only handfuls think it is just a fad, and many indicate they have been focusing on student achievement, teacher quality and accountability for quite some time. Large majorities say their districts are working to reduce the achievement gap between minority and white pupils, improve the language skills of non-English pupils … Superintendents in urban districts seem to be especially responsive to implementing standards' (p. 48). A recent survey by the New York-based Public Agenda organisation (2006) ([17]) revealed that 90 % of superintendents and 85 % of head teachers felt that student's standardised assessment results could be used to improve teaching. However, this did not preclude negative reactions to the measures in place in the United States under

([16]) Superintendents usually oversee a district (the local authority responsible for providing education services in conjunction with the national administration under the federal system in the United States). Most superintendents are responsible for schools in the district, teacher selection and recruitment policies, for setting the operating budget and for defining and monitoring school policies in the broadest sense of the term.

([17]) Public Agenda conducts opinion polls on education, primarily through its Reality Check publications. The survey described here is part of the Reality Check 2006 series and is available online at http://www.publicagenda.com/files/pdf/rc0603.pdf

the No Child Left Behind Act of 2002: less than half of managers saw the legislation as a way of raising standards.

In the United Kingdom (England), the parliamentary hearings conducted by the Children, Schools and Family Committee for its report on 'Testing and assessment' (House of Commons, 2007) indicated that managers supported testing. In one hearing, the General Secretary of the National Association of Head Teachers stated: 'Nobody in our association wants to return to the 1970s when we did not know what the school up the road was doing' (p. 12). At another hearing, a representative from Hampshire County Council declared: 'Schools readily acknowledge the need to monitor pupil progress, provide regular information to parents and use assessment information evaluatively for school improvement' (p. 12).

In France, perception and the use of standardised assessment appears to depend on the managerial staff questioned (Mons and Pons, 2006). The general inspection agency (IGEN-IGAENR, 2005) noted that head teachers and educational administrators (decentralised administration) did not make much use of the standardised assessment results, whereas middle managers with closer links to teachers (primary and secondary school inspectors) tended to use results more frequently, particularly from diagnostic tests. In some cases, inspectors were in favour of additional testing. These views sometimes shaped the underlying message being sent to teachers. 'Some management groups dealing directly with teachers … see monitoring the results targets and the associated culture of blame as a useful tool for exercising power over the staff under them and a means of validating their own professional role, even though they are not willing to be subject to the same evaluation themselves' (Demailly, 2001). However, use of these tools remains limited. The inspectors reported: We were surprised by the limited use of the initial entrance tests in year six [(18)], which secondary schools are more likely to use when setting up remedial assistance rather than as a means of identifying potential problems' (IGEN-IGAENR, 2005).

In Belgium, Maroy and Cattonar noted that the development of standardised assessment as part of the broader centralised skills standards had led to the creation of a new educational technostructure in the traditionally decentralised country. Education scientists and former teachers were mandated with defining core skills and devising a series of standardised tests. 'We could … call it a knowledge elite, an intellectual or techno-pedagogical elite within the profession, which is undoubtedly not new but has expanded considerably over the last decade … Teachers are ambivalent about the elite, perceiving them both as potential assistants when performing their tasks but also as agents of standardisation and the formalisation of teaching practices' (Maroy and Cattonar, 2002). In parallel with the emergence of this educational elite, current reforms in Belgium are increasing the powers of the more traditional administrative elite. Taken as a whole, these trends will lead to a redefinition of the context and relationship networks within which teachers operate. 'We suspect that the division of labour between these categories will become more pronounced and that increasingly teachers will find themselves in a position of dependence, either technical and professional dependence on the techno-pedagogical elite, or administrative and managerial dependence in relation to head teachers and education administrators' (Maroy and Cattonar, 2002).

([18]) First year of French secondary school or sixth year of compulsory education.

Paradoxically, whilst standardised assessment policies are breaking away from old bureaucratic systems, it appears that they can lead to the development or regeneration of middle management supervision and therefore to greater effective control over teaching methods and practices as well as academic performance.

If education professionals have mixed views on testing, how do pupils, who are most closely affected by these educational reforms, see matters?

C. Pupils and the burden of testing

As we saw in the first section on the theoretical effects of standardised assessment, **testing is supposed to improve student's motivation, at least their extrinsic** ([19]) **motivation to study. The few studies on the subject have generated contradictory findings.** In a survey conducted in one school district in Ohio, 83 % of primary school pupils and 45 % of secondary pupils maintained that they had worked harder because of the tests (DeBard and Kubow, 2002). Conversely, teachers claimed the tests either challenged their intrinsic motivation or at least failed to increase it. Several researchers have shown that testing either reduced the "love of learning", which is one element of intrinsic motivation, or had no significant effect (Jones et al., 1999; Rapp, 2002; Yarbrough, cited in Jones, 2007). In certain situations, standardised assessment appears to make pupils unresponsive because it gives rise to educational practices that are less stimulating for pupils (memory exercises rather than active learning by doing).

Although there are few studies on student motivation, much research has focused on the new feeling of stress caused by testing reforms. Pupils and teachers alike have testified to experiencing anxiety, irritation, tears or pain related to testing within certain high-stakes testing systems (Jones et al., 1999; Hoffman, Assaf and Paris, 2001; DeBard and Kubow, 2002; Gregory and Clarke, 2003; House of Commons, 2007; Jones, 2007). In Sweden, while most teachers supported the principle of standardised assessment and claimed that the system did not restrict how they taught, one-fifth of teachers reported that their pupils suffered from stress linked to national tests (National Agency for Education, 2004).

In addition to stress, **testing can have a negative impact on pupils' school careers.** With some pupils forced to repeat a year and others stigmatised because of learning difficulties, assessments can lead to an increase in school drop-out rates in the longer term (Haney, 2000; Jacob, 2001; Amrein and Berliner, 2003). As discussed earlier, in some cases pupils with serious learning difficulties who are unlikely to pass the tests are given little teacher attention, resulting in a loss of motivation which in turn leads to pupils leaving school earlier.

These trends are even more pronounced among pupils from disadvantaged backgrounds (Lipman, 2004; Jones, 2007; Hong and Youngs, 2008). With limited resources to prepare for tests, branded as failures in schools with mediocre performance results, these pupils also experience more intensive coaching for tests and a stronger narrowing of the curriculum than in schools with privileged pupils. The focus on standardised assessment detracts from environmental studies and general culture, which is particularly damaging for pupils whose families are less able to provide support.

([19]) See footnote 2 for details of the difference between intrinsic motivation which derives from the subject and extrinsic motivation which is determined by external inputs.

On the whole, although tests can produce educational coaching benefits in certain circumstances, in the case of high-stakes testing, the pressure on pupils and the number of tests – for example, English pupils have to sit 70 tests in their school career – have a negative impact on pupils' attitudes to school.

The last stakeholders that we will examine are the parents, who seem to support the idea of testing although surveys clearly show that they feel schools have a broader remit than the learning objectives covered by the test.

D. Parents see testing as positive but expect more from schools

In most countries, there has been little research into parents' views on standardised assessment.

In the United States, a lively debate has grown around the No Child Left Behind Act (NCLB), particularly following its recent amendment, prompting numerous surveys that have produced fairly consistent results. First and foremost, it would appear that **there is massive parental support for the notion of standards and testing.** According to a survey conducted in the United States (Johnson and Duffett, 2003), 82 % of parents think that clear guidelines on teaching content facilitate improvements in pupil attainment. The report by the American MrRel Research Institute (Goodwin, 2003) based on 60 interviews with focus groups found that parents were in favour of standards being linked to tests and considered that the absence of any evaluation rendered the standards useless because they would not be applied. Their support largely stemmed from the fact that the tool provided them with information about their children's scholastic achievements in an environment which they felt was not very open to external dialogue. However, some parents mentioned that their offspring had suffered psychologically because of testing. Many parents – 80 % – also expected their children to achieve the standards set out in the NCLB Act for mathematics and reading by 2013-14, in contrast to just one-fifth of teachers (according to a survey by AP-AOL Learning Services in April 2006) ([20]).

According to the poll by the MrRel Institute (Goodwin, 2003), **this massive support for standardised assessment does not preclude the possibility of a move away from testing. Parents maintained that schools should not be evaluated solely on the basis of test results.** This point came up time and again in a variety of surveys conducted in the United States, which also showed that parents were not convinced that schools should be held solely responsible for pupil attainment, as the family's social background was considered to be a crucial factor in academic success.

Research has also shown that parents see testing as part of the accountability mechanisms for administrative authorities rather than for parents, and are concerned that the measures can effectively restrict dialogue with civil society rather than reinforcing it. Lastly, a number of surveys (including Public Agenda, 2006) have also revealed that **parents see achieving the learning targets measured by the tests as just one of the school's missions. When asked about their main concerns relating to school, families ranked poor academic attainment at the bottom of the list, far below safety, discipline, respect for teachers and teaching values.**

([20]) The AP-AOL Learning Services poll of 1 085 parents and 810 teachers of children in kindergarten through 12th grade was conducted online between 13 and 23 January by Knowledge Networks after respondents were initially contacted by telephone.

In the province of British Columbia in Canada, parental surveys have sparked further debate in the wake of an argument between the provincial Ministry of Education and teachers' unions over standardised assessment – more specifically the Foundation Skills Assessment (FSA) and the Fraser Institute ranking. In a poll conducted in April 2008, the Fraser Institute, a neo-liberal think tank supporting free choice of schools and high-stakes testing, found that 83 % of parents expressed a broad support for testing and 66 % were not opposed to tests being used to rank schools. However, the teachers' union claimed that the wording of the questions was too general and did not relate to the specific measures (Vancouver Sun, 17 April 2008) ([21]).

Surveys of parents in the United Kingdom (England) revealed similarities with other countries (House of Commons, 2007). Although parents did consult the league tables, their choice of school was only partly based on the statistics. Parents also felt that the indicators were unclear.

*

* *

Overall, the reactions of the different stakeholders – teachers, education managers, pupils and parents – within the education system appear to be strongly influenced by the context in which the standardised assessment measures are developed. Unlike other policies where rejection or support relates to the principle (creation of specialist departments within comprehensive-style schools, decentralisation, school autonomy…), this instrument does not automatically produce positive or negative reactions, undoubtedly because it is associated with the traditional school routines of examination and grading. At first sight, assessment appears to be a neutral tool and seems to remain neutral provided that it has limited consequences. It is mainly high-stakes testing that provokes hostile reactions among stakeholders and can produce unintended consequences that can be detrimental to the learning process.

In conclusion, from this review of the wealth of literature that has grown up around the hotly-contested issue of standardised assessment, we can make a number of observations:

- At present there is still no firm theoretical basis to describe the effects of standardised assessment on student performance in terms of effectiveness and educational inequalities. The processes involved and the mechanisms through which the assessments are supposed to influence student attainment need to be studied in more detail and assessed empirically;
- Empirical research into the effects of testing on the performance of education systems has not yet produced a consensus, as the studies often obtain conflicting results. Reforms based on standardised assessment are backed by strong political rhetoric, but for the most part the effects of the instruments still appear to be random. The cost of these policies has not been examined in any great detail, despite the fact that economic theory underpins the reforms and focuses, quite rightly, on making the best possible use of public resources, particularly budgetary funds.

([21]) See http://communities.canada.com/vancouversun/blogs/reportcard/archive/2008/04/17/parents-support-standardized-tests-poll-shows.aspx

- Some variations in the results of testing can certainly be ascribed to the fact that the tool can be used in very different accountability models. **Further research is required on this point. However, empirical data thus far has already highlighted certain key parameters that policy-makers need to consider when establishing or reforming standardised assessment models** (for a more comprehensive overview of this subject, see Mons and Pons, 2006):

    - In general terms, serious questions need to be asked at the outset about the link between standardised assessment and other reforms, particularly where failing to establish a connection could limit the potential benefits of testing. In particular, a link should be envisaged between standardised assessment measures on the one hand and content standards or curricula, ongoing training for teachers, funding for struggling schools and the development of support plans for schools in general once they have evaluated their test results on the other hand.

    - Questions also need to be raised about the impact and nature of the stakes associated with the test, both for schools and teachers. What penalties and rewards could be associated with the test results? How many tests are needed to obtain the least distorted picture possible of the realities of the teaching provided in the schools, as a faithful representation can never be achieved using solely quantitative measures? Having just one test with high stakes for pupils and schools alike appears to generate unintended consequences. Establishing a single test to achieve several objectives (managing academic progress (moving up a year, certification), supervision of schools, assessing the education system as a whole) results in serious dysfunctions.

    - Careful thought must also be given as to how to involve teachers in accountability models. The more input teachers have in the design, management and analysis of the results, the more committed they become to the process and accept the testing culture more. Careful consideration is therefore required concerning how to involve teachers in the testing itself (design, management) and, at a more general level, in developing the internal assessment model of schools, which must be linked to quantitative external accountability.

    - **Lastly, the way the results are published needs to be considered:** should schools be ranked? What information should be provided for parents and on what basis: national or regional results, school results (whether anonymous or not) or just their children's results? Which indicators should be highlighted (raw data or added-value indicators that take into account the background of different school populations)?

It is clear from this long list of questions that there is no entirely beneficial simple recipe for a standardised assessment model. **The questions raised here indicate the shortcomings of many standardised assessment models that have been highlighted by public policy research** (Duran and Monnier, 1992). Put very simply, there is a choice between **management-based assessment** (a technical measure designed primarily to enable the administrative authorities to monitor their agents) or a **democratic evaluation** in which the model used and the interpretation of the test results are largely determined by policy-makers (as opposed to the administration) and civil society. A **new model of 'professional evaluation' (formative assessment that focuses on input from education**

professionals who are the primary users) could certainly enhance the traditional set-up. Whilst models can be combined when creating standardised tests, one approach needs to take precedence to define the general policy thinking behind the project which will reflect the answers to the questions above (what is at stake, which results are published, what kind of teacher input?).

REFERENCES

Amrein A. L., Berliner D. C. (2002), 'High-Stakes Testing, Uncertainty and Student Learning', *Education Policy Analysis Archives*, vol 10-18

Amrein-Beardsley A. A., Berliner D.C. (2003), Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Responses to Rosenshine

Education Policy Analysis Archives, vol 11-25

Anagnostopoulos D. (2006), "'Real students' and 'true demotes': Ending social promotion and the moral ordering of urban high school", American Educational Research Journal, vol 43-1, p. 5-42

Behrens M. (Ed.) (2006), *Analyse de la littérature critique sur le développement, l'usage et l'implémentation de standards dans un système éducatif*, Neuchâtel: IRDP

Belair L.M. (2005), Les dérives de l'obligation de résultats ou l'art de surfer sans planche, in C. LESSARD et P. MEIRIEU (Eds), *L'obligation de résultats en éducation*, Bruxelles: De Boeck Université

Berliner D.C., Biddle, B.J. (1995), *The Manufactured Crisis: Myths, Fraud, and the Attack on America's Public Schools*, Reading, MA: Addison-Wesley

Bishop J. H., Mane F., Bishop M., Moriatry J. (2001), The role of end-of-course exams and minimum competency exams in standards-based reforms, in D. Ravitch D. (Eds) *Brookings Papers on education Policy 2001*, p. 267-330, Washington D.C.: Brookings Institution

Bishop J. H. (2006), Drinking from the fountain of knowledge: Student incentive to study and learn. In A. Hanushek et F. Welsh (Eds), *Handbook of the economics of education*, p. 909-944, Amsterdam: North-Holland

Booher-Jennings J. (2005), "Below the bubble: 'Educational triage' and the Texas accountability System", *American Educational Research Journal*, vol 42-2, p. 231-268.

Broadfoot P. (2000), «Un nouveau mode de régulation dans un système décentralisé: l'État évaluateur», *Revue Française de Pédagogie*, n° 130, p. 43-55

Broadfoot P., Osborn M., Sharpe K., Planel C. (2001), Pupil assessment and classroom culture: a comparative study of the language of assessment in England and France, in D. Scott (Ed) *International Perspectives in Curriculum Series Vol. 1: Assessment and the Curriculum*, Greenwood Publishing Group

Carnoy M., Loeb S. (2002) 'Does external accountability affect student outcomes? A cross-State Analysis', *Educational Evaluation and Policy Analysis*, vol 24-4, p. 305-331

Center on Education Policy (2006), *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Mars 2006, disponible://www.cep-dc.org/nclb/Year4/Press

DeBard R., Kubow P. K. (2002), From compliance to commitment: The need for constituent discourse in implementing testing policy. *Educational Policy*, vol 16-3, p. 387-405

Demailly L. (2001), Enjeux de l'évaluation et régulation des systèmes scolaires, in L. Demailly (Ed), *Evaluer les politiques éducatives*, Bruxelles: Editions De Boeck Université

Duran P., Monnier E. (1992), «Le développement de l'évaluation en France. Nécessité techniques et exigences politiques», *Revue Française de Science Politique*, vol. 42, 2, p.235-262

Duru-Bellat M., Jarousse J.-P. (2001), «Portée et limites d'une évaluation des politiques et des pratiques éducatives par les résultats», *Éducation et Société*, n°8, p. 97-134

Farkas S., Johnson J., Duffett, A. (2003), *Rolling Up Their Sleeves. Superintendents and Principals Talk about What is Needed to Fix Public Schools,* New York, NY: Public Agenda

Fredericksen N. (1994), *The influence of minimum competency tests on teaching and learning*. Princeton, NJ: Educational Testing Service

Goodwin B., Englert K., Cicchinelli L. F. (2002), *Comprehensive Accountability Systems. A Framework for Evaluation*, Mid-continent Research for Education and Learning

Gordon S. P., Reese M. (1997), 'High-stakes testing: Worth the price?' *Journal of School Leadership*, vol 7, p. 345-368

Gregory K, Clarke M. (2003), 'High-Stakes Assessment in England and Singapore', *Theory into Practice*, vol 42-1, p. 66-74

Grissmer, D., Flanagan, A. (1998), *Exploring rapid achievement gains in North Carolina and Texas*. Washington, DC: National Education Goals Panel. (ERIC Document Reproduction No. 425204.)

Grissmer, D., Flanagan, A. (2001), *Searching for indirect evidence for the effects of statewide reforms*. In D. Ravitch (Ed.), Brookings papers on education policy, p. 181-207, Washington, DC: The Brookings Institution

Grissmer, D.W., Flanagan, A., Kawata, J., Williamson, S. (2000), *Improving student achievement: What state NAEP scores tell us*. Santa Monica, CA: The Rand Corporation

Haertel E. H., Herman J. L. (2005), A historical perspective on validity arguments for accountability testing. In E. H Haertel., J. L. Herman (Eds) *Use and misuses of data for educational accountability and improvement. The 104th Yearbook of the National Society for the Study of Education, Part II*, p. 1-34, Malden MA: Blackwell

Haney W. (2000) 'The myth of the Texas miracle in education', *Education Policy Analysis Archives*, vol 24-1, Disponible sur le site internet: http://epaa.asu.edu/epaa/v8n41/

Hanushek E., Raymond M. (2004) 'Does School Accountability Lead to Improved Performance? and Student Performance', *Journal of Policy Analysis and Management*, vol 24-2, p. 297-327

Hanushek E., Raymond M. (2003), 'High-Stakes Research: Accountability works after all', *Education Next*, Version électronique, disponible à l'adresse Internet suivante: http://www.hoover.org/publications/ednext/3347781.html

Hanushek E., Raymond M. (2006), 'School Accountability and Student Performance', *Federal reserve bank of St Louis Regional economic development*, vol 2-1, p. 51-61

Hargreaves D. H. (2002), *Assessing assessment*, Communication présentée au RSA Lecture programme, 13 février 2002, (www.rsa.org.uk)

Hargrove T. W., Bradford L. H., Richard A., Corrigan S. Z.; Moore C. (2004), 'No Teacher Left Behind: Supporting Teachers as They Implement Standards-Based Reform in a Test-Based Education Environment', *Education*, vol 124-3, p 567-581

Harris D. N., Herrington C. D. (2006), 'Accountability, Standards and the Growing Achievement Gap: Lessons for the Past Half-century', *American Journal of Education*, 112, p. 209-239

Helgoy I., Homme A. (2007) 'Towards a New Professionalism in School? A Comparative Study of Teacher Autonomy in Norway and Sweden', *European Educational Research Journal*, vol 6-3, p. 232-249

Hoffman J., Assaf L., Paris, S. (2001), 'High-stakes testing in reading: Today in Texas, tomorrow?' *The Reading Teacher*, vol 54, p. 482-492

Hong W-P., Youngs P. (2008) 'Does high-stakes testing increase cultural capital among low-income and racial minority students?', *Education Policy Analysis Archive*, vol 16-6, p. 2-18

House of Commons (2007), *Testing and assessment: Third report of Session 2007-08*, Children, Schools and Families Committee, London

Hoxby C. M. (2002), 'The Cost of Accountability,' *NBER Working Papers* 8855, National Bureau of Economic Research, Inc

IGEN-IGAENR (2005), *Les acquis des élèves, pierre de touche de la valeur de l'école?,* Paris: IGEN-IGAENR

Jacob B. A. (2001), 'Getting tough? The impact of high school graduation exams', *Educational Evaluation and Policy Analysis*, vol 23-2, p. 99-121

Jacob B. A. (2002), *Accountability, incentives, and behavior: The impact of high-stakes testing in the Chicago Public Schools*, National Bureau of Economic Research Working Paper 8968

Jones B. (2007), 'The unintended outcomes of High-Stakes Testing', *Journal of Applied School psychology*, vol 23-2, p. 65-86

Jones B. D., Egley R. J. (2004), 'Voices from the frontlines: Teachers' perceptions of high-stakes testing'. *Education Policy Analysis Archives*, vol 12-39

Jones M. G., Jones B. D., Hargrove T. Y. (2003), *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Johnson J., Duffett A. (2003), *Where We Are Now. 12 Things you need to Know about Public Opinion and Public Schools*. New York, NY: Public Agenda

Klieme E. *et al.* (2004), *Le développement de standards nationaux de formation: Une expertise*. Bonn: Ministère fédéral de l'Education et de la Recherche (BMBF)

Lee J. (2008), 'Is Test Driven External Accountability Effective? Synthesizing the Evidence From Cross-State Causal-Comparative and Correlational Studies', *Review of Educational Research*, 78-3, p. 608-644

Lee J., Wong K. K. (2004), 'The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes', *American Educational Research Journal*, vol 41-4, p. 797-832

Levacic R. (2001), *An Analysis of Competition and its Impact on Secondary School Examination Performance in England*, National Center for the Study of Privatization in Education, Teachers College, Columbia University Occasional Paper n°34

Linn R. L. (2000) 'Assessment and accountability', *Educational Researcher*, vol 29-2, p. 4-16

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems.* Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Lipman P. (2004), *High stakes education: Inequality, globalization, and urban school reform,* New York: Routledge Palmer.

McDonnell L. M. (2005), Assessment and accountability from the policy-marker's perspective, In E. H.Haertel, J. L. Herman (Eds) *Use and misuses of data for educational accountability and improvement.* The 104th Yearbook of the National Society for the Study of Education, Part II, p. 1-34, Malden MA: Blackwell

Maroy C. (2008), 'Vers une régulation post-bureaucratique des systèmes d'enseignement en Europe?', *Sociologie et Société*, vol 40-1, p. 31-55

Maroy C. (2005). – «Vers une régulation post-bureaucratique des systèmes d'enseignement en Europe?», *Cahier de recherche en Éducation et Formation n° 49*, Louvain-la-Neuve: Université de Louvain

Maroy C., Cattonar B. (2002). – «Professionnalisation ou déprofessionnalisation des enseignants? Le cas de la Communauté française de Belgique», *Cahier de recherche en 'ducation et Formation n° 18*, Louvain-la-Neuve: Université de Louvain

Manin B. (1996), *Principes du gouvernement représentatif*, Paris: Flammarion

McNess E., Broadfoot P., Osborn M. (2003), 'Is the effective compromising the affective?', *Bristish Educational Research Journal*, vol 29-2, p. 243-257

Goodwin B. (2003), 'Digging deeper: where does the public stand on standards-based education?', *Issues Brief*, Aurora, Colorado: McRel

Mons N. (2007), *Les nouvelles politiques éducatives*, Paris: PUF

Mons, N., Pons, X. (2006). *Les standards en éducation dans le monde francophone. Une analyse comparative,* Neuchâtel: IRDP

Mons, N., Pons, X. (2009), Pourquoi le pilotage par les résultats? Une mise en perspective théorique et historique de ce nouveau mode de gouvernance, In Mons N, Emin J.-C., P. Santana, *Piloter par les résultats,* Paris: CNDP

Nichols S. L. (2007), 'High-Stakes Testing: Does It Increase Achievement?', *Journal of Applied School Psychology*, vol 23-2, p. 47-64

Nichols S. L., Glass G.V., Berliner D.C. (2006), 'High-Stakes Testing and Student Achievement: Does Accountability Pressure Increase Student Learning?', *Education Policy Analysis Archives*, vol 14-1

McNeil L. M. (2005), Faking equity: High-stakes testing and the education of Latino youth. In A. Valenzuela (Ed), *Leaving children behind: How 'Texas-style' accountability fails Latino youth* (p.57-111). Albany, NY: State University of New York Press

OCDE (2007), *PISA 2006 Science Competencies for Tomorrow's World, Volume 1*, Paris: OCDE

Osborn M. (2006), 'Changing the context of teachers' work and professional development: a European perspective', *International Journal of Educational Research*, vol 45, p. 242-253

Public Agenda (2006), *Is support for standards and testing fading?* Reality Check 2006, Issue n°3, N.Y.: Public Agenda

Phelps R. P. (2005), *Defending standardized testing*, Mahwah, NJ: Erlbaum

Pons X. (2008*), L'évaluation des politiques éducatives et ses professionnels. Les discours et les méthodes (1958-2008),* thèse de doctorat de science politique, IEP Paris

Treisman P. U., Fuller E. J. (2001), *Comment by Philip Uri Treisman and Edward J. Fuller*. In D. Ravitch (Ed.), Brookings papers on education policy, p. 208-218, Washington DC: The Brookings Institution

Roderick M., Jacob B. A., Bryk A. S. (2002), 'The impact of high stakes-testing in Chicago on student achievement in promotional gate grades', *Educational Evaluation and Policy Analysis*, vol 24-4, p.333-357

Roderick M., Nagaoka J. (2005), 'Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless?' *Educational Evaluation and Policy Analysis*, vol 27-4, p. 309-340

Rosenshine B. (2003), High-stakes testing: Another analysis. *Education Policy Analysis Archives*, vol 11-24

Woessmann L. (2007), 'International Evidence on School, Competition, Autonomy and Accountability: a Review', *Peabody Journal of Education*, vol 82-2-3, p. 473-497

Zanten Van A. (1999), Les chefs d'établissements et la justice des systèmes d'enseignement en Angleterre et en France: les pratiques et les éthiques professionnelles à l'épreuve de la concurrence entre établissements. In D. MEURET (Ed). *La justice du système éducatif.* Bruxelles: Département De Broeck Université