DIRECTORATE FOR EDUCATION

Cancels & replaces the same document of 07 July 2009

Teacher Evaluation: Current Practices in OECD Countries and a Literature Review

OECD Education Working Paper No. 23

By Marlène Isoré

Contact: Paulo Santiago [Tel: +33(0) 1 45 24 84 19; e-mail: paulo.santiago@oecd.org]

JT03272425

OECD DIRECTORATE FOR EDUCATION

OECD EDUCATION WORKING PAPERS SERIES

This series is designed to make available to a wider readership selected studies drawing on the work of the OECD Directorate for Education. Authorship is usually collective, but principal writers are named. The papers are generally available only in their original language (English or French) with a short summary available in the other.

Comment on the series is welcome, and should be sent to either edu.contact@oecd.org or the Directorate for Education, 2, rue André Pascal, 75775 Paris CEDEX 16, France.

The opinions expressed in these papers are the sole responsibility of the author(s) and do not necessarily reflect those of the OECD or of the governments of its member countries.

Applications for permission to reproduce or translate all, or part of, this material should be sent to OECD Publishing, rights@oecd.org or by fax 33 1 45 24 99 30.

----

www.oecd.org/edu/workingpapers

ABSTRACT

This paper discusses the most relevant issues concerning teacher evaluation in primary and secondary education by reviewing the recent literature and analysing current practices within the OECD countries. First, it provides a conceptual framework highlighting key features of teacher evaluation schemes. In particular, it emphasises the importance of clarifying the purposes of teacher appraisal, whether summative when designed to assure that the practices enhancing student learning are undertaken or formative when conducted for further professional development objectives. It also encompasses the diverse criteria and instruments commonly used to assess teachers as well as the actors generally involved in the process and potential consequences for teachers' professional life. Second, it deals with a number of contentious points, including the question of the use of student outcomes to measure teaching performance, the advantages and drawbacks of different approaches given the purpose emphasised and resource restrictions, the implementation difficulties resulting from different stakeholders' interests and possible ways to overcome these obstacles. Finally, it provides an account of current empirical evidence, pointing out mixed results stemming from difficulties in assessing the effects of such evaluation schemes on teaching quality, teachers' motivation and student learning. It concludes by considering the circumstances under which teacher evaluation systems seem to be more effective, fair and reliable. Developing a comprehensive approach to evaluate teachers is critical to make demands for educational best practice compatible with teachers' appropriation of the process as well as to enhance the decisive attractiveness and recognition of the teaching profession.

RÉSUMÉ

Ce papier examine les principales questions relatives à l'évaluation des enseignants du primaire et du secondaire en passant en revue la littérature récente et en analysant des pratiques actuelles au sein des pays de l'OCDE. Premièrement, il fournit un cadre conceptuel mettant en évidence les éléments clés entrant dans les processus d'évaluation des enseignants. En particulier, il souligne l'importance de clarifier les objectifs de l'évaluation, qu'ils soient de nature sommative lorsqu'ils visent à assurer que les pratiques favorisant l'apprentissage des élèves sont à l'œuvre ou de nature formative lorsqu'ils sont conduits à des fins de formation professionnelle continue. Il comprend également les différents critères et instruments communément utilisés pour évaluer les enseignants ainsi que les acteurs généralement impliqués dans le processus et les conséquences potentielles sur la vie professionnelle des enseignants. Deuxièmement, il traite d'un certain nombre de points conflictuels, parmi lesquels la question de l'utilisation des résultats des élèves pour mesurer la performance des enseignants, les avantages et inconvénients de différentes approches compte tenu de l'objectif mis en exergue et de ressources limitées, ou encore les difficultés de mise en place résultant de divergence d'intérêts et les moyens possibles d'y remédier. Enfin, il examine l'évidence empirique sur le sujet et explique en quoi ses résultats nuancés tiennent aux difficultés d'estimer les effets de tels processus sur la qualité de l'enseignement, la motivation des personnels et l'apprentissage des élèves. Pour conclure, il considère les circonstances dans lesquelles l'évaluation des enseignants semble plus efficace, équitable et fiable. Développer une approche d'évaluation compréhensive est cruciale pour concilier les exigences d'enseignement et l'appropriation du processus par les enseignants, tout en recherchant une nécessaire amélioration de l'attractivité et de la reconnaissance du métier d'enseignant.

TABLE OF CONTENTS

TEACHER EVALUATION: CURRENT PRACTICES IN OECD COUNTRIES AND A LITERATURE REVIEW

1. INTRODUCTION

1.   This paper examines the current academic and policy literatures on teacher evaluation in primary and secondary education. It updates and expands the corresponding Section in *Teachers Matter: Attracting, Developing and Retaining Effective Teachers* published by the OECD in 2005.

2.   Evaluation of teacher practice and performance is not a recent concern. Demands for instructional quality have led many countries to set up one form or another of teaching performance assessment. Nevertheless, teacher evaluation has always been a highly controversial subject, with both mixed empirical evidence about its effects on student learning and conflicts of interest between key actors of education systems. As a consequence, evaluation has often been a meaningless exercise, stemming from required bureaucratic rituals in schools, and endured by both teachers and evaluators (Danielson, 2001; Holland, 2005; Marshall, 2005). Only recently some countries have demonstrated a growing interest in establishing evaluation systems as an integral part of broader teacher and school policies (Peterson, 2006; TDA 2007a).

3.   Existing schemes of teacher evaluation in OECD educational systems take multiple forms. Scope and methods of teacher evaluation, criteria and standards used and data gathering instruments differ largely from one country to another, according to the educational context and tradition, the actors involved in the design and implementation of the evaluation system and the purpose of evaluation emphasised. Consequences of evaluation processes for teacher careers are also diverse. Although the single promotion table and the single salary schedule remain widespread, several countries attempted to link their teacher appraisal system either to recognition and rewards, whether financial or not, or to professional development opportunities.

4.   This paper has three further sections. Section two examines the key dimensions of teacher evaluation schemes found in the literature or in OECD education systems. Section three discusses the advantages and disadvantages of different approaches, as well as the difficulties in implementing effective evaluation programmes resulting from different stakeholders' arguments. Finally, section four summarises current evidence on the effects of teacher evaluation systems.

2. KEY FEATURES OF TEACHER EVALUATION SCHEMES

5.	This section describes the key dimensions of teacher evaluation schemes found in the literature and within the education systems of the OECD area.

### 2.1	Purposes of evaluation

6.	Teacher evaluation has two major purposes. One the one hand, it is aimed at ensuring that teachers perform at their best to enhance student learning. On the other hand, it seeks to improve the teacher own practice by identifying strengths and weaknesses for further professional development. These two approaches refer to assessments of different nature, respectively summative and formative.

#### 2.1.1	*Summative assessment and quality assurance*

7.	If the ultimate goal of education systems is to provide improved learning for all students, and if teacher performance and practice is the most important factor in this, then teacher evaluation may be considered as a quality assurance mechanism (Danielson and McGreal, 2000; Kleinhenz and Ingvarson, 2004). Assuming that the quality of teachers and the quality of teaching matter, an evaluation process should ideally be directed towards both educational efficiency – ensuring that teaching meets the academic standards for students to live in knowledge societies – and educational equity – ensuring that attainment opportunities are accessible to all students regardless of their background. Thus, *summative* evaluation of teaching is a way to assess that teachers are adopting the actions and 'best' practices which improve student outcomes.

8.	Conducting a *summative* assessment is the most visible and recognisable way to evaluate someone, which consists of providing summary statements of a teacher's capabilities through examinations, in order to measure aptitude and knowledge, to ensure that required standards are met, or to promote level of performance for immediate recognition. Teacher summative evaluation gives crucial information about the current practices and performance of the teacher being evaluated relatively to what is considered as standards of 'good' teaching. Hence, summative evaluation is an indispensable source of documentation to hold teachers accountable for their professionalism. Stronge and Tucker (2003) for example emphasise the necessity of such a quality assurance mechanism: "The accountability purpose reflects a commitment to the important professional goals of competence and quality performance. This accountability function (…) relates to judging the effectiveness of educational services".

9.	The need for accountability mechanisms in teaching comes from asymmetric relationships, typical of the 'Principal-Agent Problem' well-known to the economic and political science theories. The 'principal' lacks critical information to know if his employee, the 'agent', behaves in conformity with his outcome expectations. In our particular case, parents, authorities in charge of educational quality, or even school principals, have only limited means to know the degree to which teachers act in accordance with their students' learning expectations. As explained by Mizala and Romaguera (2004), "Accountability is fundamental, because an information gap separates schools from families. It is costly for families to obtain relevant, up-to-date information on what is happening with their schools, and schools are not necessarily given incentives to provide information to parents. Moreover, depending on their cultural and socioeconomic level, families' ability to obtain information about schools varies." Given this asymmetric information, Casson (2007) argues that, in absence of incentive mechanisms, teachers have an incentive to

exert less effort (given cost associated with more work) because "the school district cannot distinguish between low student performance due to a lack of teacher effort and low student performance due to low student ability". While Casson argues in favour of measures aligning teachers' behaviour to authorities' interests *ex post* to overcome this problem, a well-designed summative evaluation system may suffice to encourage teachers to adopt the best practice because it closes the information gap *ex ante*.

10.   Besides the informational purpose *per se*, results of summative assessments allow the making of consequential decisions concerning the teacher being evaluated. According to Avalos and Assael (2006), "most forms of evaluation are justified either because diagnostic information is needed or because they provide evidence for decision making. The same is true for teacher performance evaluation". Evaluating teachers in relation to specific criteria makes comparisons possible, the latter being useful for hiring and tenure decisions, promotion opportunities or, under particular conditions, responses to ineffective teachers.

11.   Summative evaluation of teacher performance can also be used as a basis for recognition and celebration of a teacher's work. There are concerns about the image and status of teaching in a number of OECD countries, including teachers' feeling that their work is undervalued. Evaluation provides opportunities to recognise and reward teaching competence and performance, which is essential to retain effective teachers in schools as well as to make teaching an attractive career choice (OECD, 2005). For instance, the US National Board for Professional Teaching Standards (NBPTS) has developed "rigorous professional standards for what accomplished teachers should know and be able to do" as a basis for a national voluntary system certifying teachers who meet these standards. Since its creation in 1987, more than 64,000 American teachers have been celebrated for their outstanding performance.

12.   In its summative form, evaluation firstly responds to the needs of assuring that teaching is directed towards student achievement. It also provides opportunities for social recognition of teacher's skills and commitment to work. These are two major concerns in our knowledge societies.

2.1.2   *Formative assessment and professional development*

13.   Aside from accountability and recognition purposes, teacher evaluation can be conducted in order to improve the teacher's practice itself. *Formative* evaluation refers to a qualitative appraisal on the teacher current practice, aimed at identifying strengths and weaknesses and providing adequate professional development opportunities for the areas in need of improvement. As explained by Stronge and Tucker (2003) "The performance improvement purpose relates to the personal growth dimension and involves helping teachers learn about, reflect on, and improve their practice. This improvement function generally is considered *formative* in nature and suggests the need for continuous professional growth and development." As opposed to a summative assessment designed to make judgements about a performance (assessment *of* teaching), the role of a formative assessment is to underline ways to improve the current practice (assessment *for* teaching).

14.   Formative evaluation is a process by which evaluators give constructive feedback to the teacher, pointing out at what level the teacher is performing on each of the relevant criteria, and suggesting ways to enhance his practice. Conversations with evaluators or colleagues engage teachers in self-reflection about their work. As put by Danielson and McGreal (2000) "As teachers consider the wording of different components of teaching and their elements and compare their impressions and practices with one another, they trade techniques and learn new strategies from their colleagues. These conversations are rich – focused on the quality of teaching and contributing much to the professional learning of those participating." Empowering individual teachers in their own skills betterment goes far beyond the quality assurance purpose of teacher evaluation.

15.   Furthermore, the results of formative assessment allow schools to adapt their professional development programmes to the needs of their teachers in accordance to their educational objectives. Schools can learn from the strengths of effective teachers – emphasised by formative evaluations – and implement professional development programmes that respond to their weaknesses. In Finland, the school principal is the pedagogical leader, responsible for the teachers in her school and for the implementation of measures needed to enhance teaching quality. As a consequence, most of Finnish schools have a system that includes annual discussions aimed at evaluating the teacher's fulfillment of individual objectives set up during the previous year and analysing individual objectives and needs for the next year (UNESCO, 2007).

16.   In the same way, institutions in charge of teacher education can also benefit from the feedback provided by formative assessments. Pecheone and Chung (2006) describe the Performance Assessment for California Teachers (PACT) (a system for pre-service teachers competence evaluation for the purpose of teacher licensure) as "a powerful tool for teacher learning *and* programme improvement". Indeed, they add that this system has introduced many professional dialogues "about what constitutes effective teaching and about what graduates should know and be able to do", which in turn have led programmes "to reexamine the way they support and prepare candidates to teach".

17.   In its formative form, evaluation can be considered as a basis for teaching improvement and lifelong professional development opportunities. Summative and formative aspects of teacher evaluation are often conflicting – but not necessarily incompatible – purposes. In practice, countries rarely use a pure form of teacher evaluation model but rather a unique combination that integrates multiple purposes and methodologies (Stronge and Tucker, 2003).

2.2   Key elements of teacher evaluation schemes

18.   This section summarises the aspects involved in teacher evaluation systems, such as the actors engaged in designing and implementing the process, the scope of evaluation, the data gathering instruments and methods, and the criteria and standards used to assess teachers.

2.2.1   *Actors involved in the conception and implementation of evaluation systems*

19.   *Governments.* Governments play a major role in the conception of evaluation schemes, since they set the national learning outcome objectives, often by law. In the US for instance, the No Child Left Behind (NCLB) Act of 2001 aims to improve the performance of US primary and secondary schools by increasing the standards of accountability for States, school districts, and schools. Efficient evaluation systems should be directed towards the achievement of the national goals. In addition, governments sometimes play a direct role in the implementation and in the monitoring of teacher evaluation procedures. The extent of this function depends on the degree of decentralisation of the country (UNESCO, 2007). In this sense, France is a paradigmatic case: most of legal decisions are made at the central level, including *a fortiori* decisions relative to public schools; thus, the Ministry of Education is in charge of determining the different aspects of the evaluation system. More generally, national authorities are greater actors in countries where teachers are public servants.

20.   *Local authorities*. Local authorities in charge of education policies have generally one of the two following duties in relation to teacher evaluation. In some OECD countries, local authorities are accountable for the achievement of the national objectives, and therefore implement procedures considered as desirable to assure the educational quality of the schools under their responsibility. For instance, Heneman *et al.* (2006) summarise the characteristics of teacher evaluation models which are proper to four US districts (Cincinnati, Washoe, Coventry and Vaughn). In other countries, local authorities can take part in the conception of the evaluation scheme, but are above all in charge of implementing and monitoring the

teacher evaluation measures decided by the central State. For example, Chilean municipalities provided advice for the design of the evaluation process, but now all apply the same national system, which was enacted by law in 2004 (Avalos and Assael, 2006).

21.   *School leaders.* The more decentralised the country is, the more school leaders take an important part in designing and implementing the evaluation process. In Finland, whose educational system is characterised by a very high degree of school autonomy, all decisions relative to teachers (including evaluation) are made within the schools (UNESCO, 2007). The role of school leaders as proper evaluators will be discussed further.

22.   ***Educational researchers and experienced teachers.*** Researchers and teachers may be consulted as experts for designing the system. They are in a good position to know what 'good' teaching practices are, by dint of their studies or own experience as teachers, and then help to identify the relevant criteria and instruments to evaluate teachers (Ingvarson, Kleinhenz and Wilkinson, 2007). The NBPTS, mostly constituted of teachers, "examined the pros and the cons of different research methods, and then applied their own experiences to what they heard and learned – always reflecting on the intersection of large-scale empirical data, their own development as expert teachers, and the nature of the students they teach and serve. They deliberated and debated among themselves, and reached out to colleagues to generate additional perspectives and insights" (NBPTS, 2007).

23.   *Teacher unions.* Teacher unions may also be consulted to design and implement procedures in respect with teachers' day-to-day practices and difficulties. Teacher unions are supposed to represent all teachers' stakes, whatever their level of performance. Assuming that teachers may have reluctance or fears about some particular aspects of teacher evaluation schemes, teacher unions are indispensable for designing a process that will take their interest into account and lead to a wide agreement. Heneman *et al.* (2006) argue that mechanisms for lessening resistance must be incorporated into the initial design of the plan. These include communicating extensively and continually with teachers and administrators. They add that a commitment to a transformation in how teacher performance is defined, measured, and supported is needed, and that such commitment needs to address teachers' and administrators' apprehensions.

24.   *Parents.* Parents are rarely, if ever, directly involved in the designing or implementation of teacher evaluation systems, since their educational stakes are represented by the national and local authorities mentioned above. Their role as evaluators will be discussed further.

2.2.2   Scope of evaluation and teachers evaluated

25.   Teacher evaluation procedures do not necessarily apply to all teachers within a country; on the contrary, the scope of evaluation and the teachers who are the subject of evaluation significantly differ across OECD educational systems. The main differences are as follows.

26.   *Regional procedures.* The same procedures do not necessarily apply to the whole country but may vary according to the region considered. Procedures are more likely to fluctuate on a regional basis when the federal structure or the high degree of decentralisation of the country allows it. For instance, in Germany, the Ministries in charge of education in each Land determine their own orientations for all aspects of teacher evaluation: people in charge of the process, the criteria for evaluation, time devoted to evaluation, the data gathering instruments and the consequences of the evaluation results (UNESCO, 2007).

27.   *School type.* The system may be limited to public schools but may also apply to some private schools, particularly schools which are at least partly subsidised by the State although privately owned and managed. For example, the *Teacher Growth, Supervision and Evaluation Policy* of the Canadian's

province of Alberta is equally applicable to charter schools (schools with a semiautonomous organisation but completely publicly funded) and to accredited private schools (only partly subsidised by public funds) (Alberta Education, 1996, 2003).

28.   *Teacher's level of experience.* Teacher evaluations may differ according to the teacher's level of experience. In England for example, the characteristics of teachers are defined at each career stage (TDA, 2007b), as their roles and responsibilities evolve throughout their career. However, many OECD countries do not differentiate teachers according to their level of experience once tenure is obtained: "teaching, alone among the professions, makes the same demands on novices as on experienced practitioners. The moment first-year teachers enter their first classrooms, they are held to the same standard – and subjected to the same procedures – as their more experienced colleagues. (…) Although the school district must ensure that all teachers (including beginning teachers) have at least a certain level of skill, the procedures used might be somewhat different for novices than for their more experienced colleagues." (Danielson and McGreal, 2000). No agreement exists about how evaluation should be differentiated according to the level of experience. For instance, it can be argued that pre-service teachers should be subject to more frequent and more complete evaluations since they have to demonstrate the adequate practice and performance for licensure. Darling-Hammond *et al.* (2004) pointed out that the NBPTS's efforts to identify and recognise good teaching has incited the creation of a similar system to assess Californian pre-service teachers (PACT system). On the other hand, such evaluations may not be necessary, given that beginning teachers' internship and induction programme provides them with formative opportunities. Thus, the American Federation of Teachers emphasises that more than half of the American states have no requirement that a teacher completes a successful year or two of teaching in order to be fully licensed (AFT, 2001). Some argue that summative assessments should be relatively more important for experienced teachers since they are supposed to better understand what is expected from a 'good' teacher (Danielson and McGreal, 2000) whereas others emphasise their prime need for formative feedback in order to keep their motivation intact (Day and Gu, 2007).

29.   *Periodicity of evaluation.* A major issue to consider is whether formal evaluations are part of the teacher's regular work or occur in special instances (and informal evaluations remain otherwise). In the Netherlands, 38 per cent of primary schools and 62 per cent of secondary schools evaluate their teachers regularly; in most of these cases, teachers are evaluated annually (UNESCO, 2007). In other countries, there is a compulsory process of formal evaluation only when teachers are the subject of a complaint. In Italy, once tenure is granted, formal evaluations may occur when the school administrators start a procedure for 'exemption from services' because of an inadequate teaching or an insufficient performance is observed over a significant period; a teacher can also ask to be evaluated for rehabilitation if he is under the yoke of a disciplinary sanction (UNESCO, 2007).

30.   *Compulsory vs. Voluntary evaluations.* Aside from compulsory evaluation procedures, some countries offer teachers the possibility to be voluntarily evaluated, in order to apply for a salary increment or a higher position. In Spain for example, a teacher of secondary education can exceptionally be promoted after the examination of her evaluation results by a professional board, the *cuerpo de catedráticos* (UNESCO, 2007). Formal compulsory evaluations and voluntary or informal evaluations can be combined since their purpose is generally not identical. Indeed, a formal evaluation procedure is more likely destined to minimal summative purposes while an informal one provides the most formative opportunities, and a voluntary one is preferred for further consequential summative purposes. For instance, two forms of evaluation coexist in Chile: a compulsory periodical evaluation for all public school teachers and a voluntary evaluation for promotion (Avalos and Assael, 2006).

31.   *Pilot implementation.* The scope of evaluation can be limited to a pilot implementation during several years before a full implementation. The PACT system in California began with two years of pilot implementation (in 2002-2003 and 2003-2004), with restricted teaching areas tested, in order to validate

the methods used and improve the programme for the following years (Pecheone and Chung, 2006). The more teachers' professional lives will be affected by the evaluation process, the more important the pilot implementation is.

2.2.3 Criteria and standards

32.   A fair and reliable teacher evaluation scheme needs criteria and standards to evaluate teachers relatively to what is considered as 'good' teaching. Teaching competences and responsibilities should be listed in order to build a comprehensive definition of what teachers should know and be able to do in the exercise of their profession. A reference contribution in this area is the Danielson's *Framework for Teaching* (1996, 2007), which is articulated to provide at the same time "a 'road map' to guide novice teachers through their initial classroom experiences, a structure to held experienced professionals become more effective, and a means to focus improvement efforts".

33.   The *Framework* groups teachers' responsibilities into four major areas further divided into components:

- **Planning and Preparation**: demonstrating knowledge of content and pedagogy, demonstrating knowledge of students, selecting instructional goals, designing coherent instruction , assessing student learning;

- **The Classroom Environment**: creating an environment of respect and rapport, establishing a culture for learning, managing classroom procedures, managing student behavior and organising physical space;

- **Instruction**: communicating clearly and accurately, using questioning and discussion techniques, engaging students in learning, providing feedback to students, demonstrating flexibility and responsiveness;

- **Professional Responsibilities**: reflecting on teaching, maintaining accurate records, communicating with families, contributing to the school and district, growing and developing professionally, showing professionalism.

34.   Each of these components consists of several elements to evaluate. For example, the teacher's knowledge of students encompasses elements such as knowledge of characteristics of age groups, knowledge of students' varied approaches to learning, etc. Each element of a component is associated with four levels of performance: 'unsatisfactory', 'basic', 'proficient', and 'distinguished'.

35.   Danielson underlines that the levels of performance are especially useful in supervision and evaluation but can also be employed to help with self-assessment or to support mentoring or coaching relationships, to inform a professional discussion and suggest areas for further growth. Thus, the *Framework* can serve both summative and formative purposes. Danielson also cautions against potential misuses of the components, arguing that, if the components are generic and designed to apply to any teaching situation, their actual manifestations however differ in various contexts. Therefore, evaluators need to examine the applicability and weighing of each component as well as to translate the elements into specific, observable examples in particular contexts.

36.   Kleinhenz and Invargson (2004) caution against "the absence of standards that adequately explicate the work of teaching – what it is that teachers can be expected to know and be able to do in specific domains of practice", which necessarily lead to a weak "technical core of teachers' knowledge and skills". They add that "it is now widely accepted that comprehensive, congruent, domain specific standards

provide the only credible basis for making useful judgements of teacher competence". Among the sets of standards currently used to assess teachers, some were proposed previously to Danielson's *Framework*, but they are related to workers in more specific situations. For example, the NBPTS's standards impact accomplished teachers who are voluntarily evaluated to be recognised for the high level of performance they have acquired throughout their experience in teaching. Regarding beginning teachers on the contrary, the Interstate New Teacher Assessment and Support Consortium (INTASC) developed standards of teaching knowledge and skills for teacher licensing systems (CCSSO, 1992).

37. A number of teacher evaluation systems in the United States have set a list of criteria based on Danielson's *Framework for Teaching*. The four US districts of Cincinatti, Washoe, Coventry and Vaughn adopted customised versions of the *Framework*'s competency model (Milanowski, 2004; Borman and Kimball, 2005; Heneman *et al.*, 2006). So did the province of Quebec in Canada. Chile's four domains and twenty criteria of assessment were also largely inspired by the *Framework* (Avalos and Assael, 2006). UNESCO's analysis of the European and Latin American teacher evaluation systems emphasises the content knowledge, the pedagogical skills, the abilities to assess students and the professional responsibilities vis-à-vis the school, the students and their families as key domains to evaluate teachers. One should note that the analysis does not mention the engagement in professional development programmes as a common teaching standard in European systems, with a subsequent risk to undervalue the teacher's engagement and willingness to enhance his own practice. Nevertheless, England has recently implemented a framework for professional standards, close to Danielson's one, which includes professional development criteria for the five levels of teaching performance (the award of Qualified Teacher Status, teachers on the main scale, Post Threshold Teachers, Excellent Teachers, and Advanced Skills Teachers) (TDA, 2007b).

38. Among the eight standards (including 42 criteria) of the State of Iowa, the following six are very close to Danielson's *Framework*: competence in content knowledge, competence in planning and preparing for instruction, methods for instruction and assessment of student learning, competence in classroom management, engagement in professional growth and fulfilment of professional responsibilities (Iowa Department of Education, 2002). However, the remaining two criteria differ from Danielson's definition of what a good teaching is. The first one refers to the strategies used to deliver instruction that meets the multiple learning needs of students. Danielson argues that equity is implicit in the entire *Framework*, particularly for the domains related to interaction with students ("in an environment of respect and rapport, *all* students feel valuated"), yet acknowledging that "an awareness of developmental appropriateness can be extended to include a sensitivity to students with special needs". By contrast, the Iowa's standard focuses explicitly on the ability of the teacher to improve equity, for example using strategies that address the full range of cognitive levels, or connecting students' prior knowledge, life experiences, and interests in the instructional process. The second Iowa's standard that differs from the *Framework* is the teachers' demonstration of ability to enhance academic performance. For instance, the teacher must provide evidence of student learning to students, families, and staff. This point is more contentious because not only a multitude of other factors may influence student learning outcomes, but also and more importantly, it may not be part of the teacher's role to demonstrate to a third party that he is at the roots of his students' academic success. Student outcomes are sometimes used by evaluators as an instrument to evaluate teacher performance, but requiring this proof from the teacher is generally not a criterion of good practice *per se* subject to evaluation.

2.2.4 *Data gathering instruments*

39. While wide consensus is generally reached about the criteria of good teaching, much more contentious are the instruments for collecting evidence on the teacher's current practice. Since the way of gathering evidence about a particular teacher may influence the assessment results, the choice of

instruments is of chief importance in designing and implementing systems to evaluate teacher performance.

40.   ***Classroom observations.*** Classroom observations are the most common source of evidence used in OECD countries, whether American (*e.g.* Canada, Chile, United States), European (*e.g.* Denmark, France, Ireland, Spain) or Asian-Pacific (*e.g.* Australia, Japan, Korea). This process permits to observe if the teacher adopts adequate practices in his more usual workplace: the classroom (UNESCO, 2007). However, depending on the evaluator and the context, the usefulness and informativeness of the evidence collected may differ. Peterson (2000) explains that the observation of the content expertise of a teacher plays a minor role in some situation but is very important in others.

41.   ***Interviews of the teacher.*** Interviews of teachers may take multiple forms, be highly structured or not. In seldom cases, they are useful for direct judgments of a teacher's competences and skills, but they are more adequately used for professional growth, asking teachers in which ways they would like or need to improve. For example, the English schools that supply in-depth professional development to teachers rely on performance management interviews to identify the staff's individual needs (Ofsted, 2006). Nonetheless, teachers' propensity to reveal their real weaknesses and fears during interviews depends on their confidence in the interviewer and their perceptions of the possibility to receive relevant and constructive feedback from the evaluation process.

42.   ***Portfolio prepared by the teacher.*** Portfolios require teachers to gather documentation about their current work. Different elements can compose teacher-prepared portfolios: lesson plans and teaching materials, instruction videotapes, samples of student work and commentaries on student assessment examples, teacher's self-reported questionnaires and reflection sheets. Beck, Livne and Bear (2005) emphasise that an important dilemma in designing portfolios is whether the portfolio is primarily a vehicle for teacher assessment or for teacher development and whether these two objectives are compatible. This point will be detailed later. Moreover, because portfolios are a complex source of evidence, Wertzel and Strudler (2006), Strudler and Wertzel (2008), Jun et al. (2007) and Jacobs, Martin and Otieno (2008) argue that the constitution of complete portfolios is particularly useful in the evaluation of pre-service or beginning teachers. Teacher education programmes may also benefit from beginning teachers' assessment results on portfolios (Jacobs, Martin and Otieno, 2008).

43.   ***Student outcomes.*** Student outcomes are not commonly used as sources of evidence for teacher evaluation in OECD countries (OECD, 2005; UNESCO, 2007). Student achievement results may reflect teaching performance, especially when measured in value-added gains rather than in absolute terms, *i.e.* after controlling for the previous results of individual students the teacher taught (Braun, 2005). The Californian Teacher Performance Assessment for example measures student learning improvements relatively to the districts' standards in order to recommend teacher for a credential. Nonetheless, student learning is rarely used as a measurement of teacher performance in existing schemes, either because there are no regular student standardised tests allowing viable comparisons, or because it encounters strong rejections from teachers and scholars judging this instrument as flawed, ineffective or unfair (Weingarten, 2007). Advantages and disadvantages of such a direct measure of 'pure performance' will be discussed further.

44.   ***Teacher test.*** Exceptionally, teachers' curricular knowledge and pedagogical skills are assessed through written tests. It is actually the case for new teachers in Chile (Avalos and Assael, 2006) or teachers applying for promotion in Mexico (OECD, 2005).

45.   ***Questionnaires and surveys.*** Questionnaires on the teacher's practice could be completed by the school principal, parents or students, *i.e.* the ones that may testify for teaching quality through their continuous interaction with the teacher, and not only during the evaluation process (Peterson, 2000;

Peterson *et al.*, 2000, 2005; Jacob and Lefgren, 2005b). This precise category of questionnaires and surveys therefore excludes evaluator reports resulting from classroom observations or interviews of the teacher; it is restricted to questionnaires as sources of evidence *per se*. Student surveys are tools of teacher evaluation in Mexico, the Slovak Republic, Spain or Sweden, generally for teachers applying for a promotion; to our knowledge, there is no existing case in compulsory teacher evaluation schemes. While their utilisation can provide some interesting insights, cautions have to be taken because the evaluators are not teaching experts and do not necessarily value the same qualities than the ones which are supposed to enhance student learning (Peterson *et al.*, 2000, 2003; Jacob and Lefgren, 2005b). Research studies on the use or reliability of such procedures remain unfortunately very rare.

2.2.5 *Evaluators*

46.   ***Internal review.*** In most countries, teacher evaluations involve the school principal or other senior school staff (Peterson, 2000; OECD, 2005, 2008; UNESCO, 2007). However, the engagement of school leaders in the evaluation process differs between and within countries. In 2003, 100 per cent of the US students were enrolled in secondary schools where principals reported that they made classroom observations in the preceding year whereas it was the case of only 5 per cent of students in Portugal, with the OECD average at 60 per cent (OECD, 2008). Within countries, school leaders vary in the time and capacity they have to take this important responsibility (Marshall, 2005; Jacob and Lefgren, 2005a, 2008). The advantages and disadvantages to have the evaluation done by principals rather than other evaluators will be examined later.

47.   ***External review.*** Some countries have implemented evaluation schemes where teachers are evaluated by peers or by accomplished teachers, either exclusively (Ireland) or as part of a panel which includes the school principal (France). On the one hand, 'peers' are other teachers who are equivalent in assignment, training, experience, perspective, and information about the setting for the practice under review, but should neither teach at the same school as the teacher being evaluated nor be socially or politically connected with him (Peterson, 2000). On the other hand, 'accomplished teachers' are recognised as having in-depth subject knowledge and pedagogical expertise, as highly proficient and successful practitioners, able to guide and support others in the teaching process (MCEETYA, 2003). Both have relative advantages in the extent that the former evaluators take part in the process on an equal footing with the teachers assessed while the latter evaluators provide a proficient perspective.

48.   ***Self-evaluation.*** Engaging teachers in 'empowerment evaluations' is essential both to gain agreement from teachers on the evaluation process and to enhance teacher performance (Peterson, 2000; Kennedy, 2005). Portfolios are particularly adequate instruments for teacher self-reflection because the proper decision made by the teacher to include particular artifacts (lesson plan, videotape of lesson, sample of student work, narrative comments) instead of others is a judgement that requires determining how the features of one artifact are superior to others (Danielson, 1996, 2007; Darling, 2001; Mansvelder-Longayroux *et al.*, 2007). Combined with other evaluator's review, documents prepared by the teacher may be used for a summative purpose. However, the formative purpose is predominant since the reflection process enables the teacher to be aware of his own strengths and weaknesses, and to identify her needs for improvement, professional development or coaching.

49.   ***Parents.*** Parents generally play an indirect role in the evaluation process when principals' reports include their complaints about, or on the contrary their requests for, a particular teacher. They are less frequently direct evaluators, via questionnaires for example. The tiny current evidence on that subject shows that they value teacher characteristics that surprisingly depart from student achievement: 'the teacher's ability to promote student satisfaction' (Jacob and Lefgren, 2005b), 'humane treatment of students', 'support for pupil learning', and 'effective communication and collaboration with parents' (Peterson *et al.*, 2003). Even if their perspective could be taken into account, their distance from the

teaching professional standards, their ignorance about what happens in the classrooms, and their emotional implication suggest that their appraisals are far from sufficient for a comprehensive teacher evaluation scheme.

50.   ***Students.*** Students are also rarely consulted as evaluators. Mexico, Spain and Sweden use student surveys, but generally for limited education grades or in special cases of teacher evaluation (on a voluntary basis for a promotion, or in a complaint procedure for example). Studies on students' teacher appraisals for primary and secondary education levels are extremely rare. Peterson *et al.* (2000) argue that students respond with viability and reliability about teacher quality if questions are formulated in a simple and relevant way. They have proposed three sets of items that they argue to work well, for prereader, elementary and middle- and high-school students respectively. Nevertheless, they consider students as "clients" although this point is highly questionable. Indeed, students do not directly pay for educational services (while their parents may pay for but are not the consumers of educational services) and, even more importantly, students are involuntary enrolled members – they are not free to leave the organisation, to choose their school or teachers, or to influence what or how they are taught – (Greenfield, 1995).

2.3   Links to recognition and rewards

51.   The evaluation process may be linked with teacher recognition and rewards mechanisms. In most OECD countries there is a single salary schedule for teachers and few formal incentives for and recognition of good practice. This raises concerns about the attractiveness of teaching as a career choice and maintaining teacher motivation throughout the career. The Australian Department of Education argues that while people who have chosen teaching as a career are chiefly motivated by 'intrinsic' rewards, extrinsic factors such as remuneration are the most significant factors influencing people not to choose teaching as a career – especially for prospective high quality entrants –, and influencing teachers to leave the profession (DEST, 2007).

52.   Heneman *et al.* (2006) argue that standards-based teacher evaluation systems should be used as a foundation for knowledge- and skill-based pay. They support an incentive strategy that requires the design and implementation of alternative teacher compensation systems which depart from the single salary schedule. This new strategy, currently being pursued by several American States (Peterson, 2006), links pay to combinations of assessments of teacher performance, acquisition of new knowledge and skills, and student test score gains. Few European countries provide a direct pay increase from the salary base for good performance. For example, Romania has set a system in which the best teachers can compete for a temporary salary rise from fifteen per cent during a year to twenty per cent during four years (UNESCO, 2007).

53.   Other countries do not directly link teacher evaluation results with teacher pay but link them to career progression. France, Germany, Greece, Poland, Portugal and the United Kingdom are among them. The English teachers who meet the standards for 'Post Threshold, Excellent and Advanced Skills Teachers' also access the relevant pay scale (TDA, 2007b). By contrast, continuous negative evaluation results are linked to deferrals of promotion in a number of OECD countries (OECD, 2005). One should note that linking the evaluation process to salary increments through the promotion schedule may not suffice in countries where the top stages of the career schedule are reached early in the teaching career. In Australia for instance, the incentive of salary increments linked to performance review does not apply to the majority of teachers which are already at the top of the incremental salary scale after 10-12 years of teaching, and thus, does not seem to be particularly effective (Kleinhenz and Ingvarson, 2004).

54.   Finally, some countries link teacher assessments with opportunities for vertical promotions to school leadership positions. In Spain, one of the conditions to be elected as the head of the Teaching Council (*Consejo Escolar*) is to pass the teacher evaluation process (UNESCO, 2007). Likewise, in the

years after the 10th or 12th year Australian career progression stage, some highly accomplished teachers are promoted to administrative positions up to and including principal positions (Kleinhenz and Ingvarson, 2004). Nevertheless, the practice of linking outstanding teacher performance to vertical promotions can be criticised, for two main reasons. On the one hand, a good teacher is not necessarily a good manager or leader. On the other hand, this practice may have adverse effects while aiming at recognising the teaching profession: paradoxically, the best teachers are rewarded by no longer doing what they do best – teaching (UNESCO, 2007). Indeed, the outstanding teachers who choose to keep on teaching are thus considered as "shadowy creatures who occupy the netherworld of the classroom", whose knowledge and skills are seldom recognised. Taking the opposite direction, some Australian schools have thus implemented a teacher classification between the highest stage in the automatic salary scale and administrative classifications in order to reward teachers who choose to remain in the classroom rather than move into administration (Kleinhenz and Ingvarson, 2004).

2.4   Links to professional development opportunities and school broader priorities

55.   Few countries link reviewed performance with ongoing professional development. Yet, a logical chain between the performance assessment and continuing professional development opportunities is essential to improve teaching practice (Ofsted, 2006). The identification of individual teachers' strengths and weaknesses is important to choose from a wide range of possible professional development activities the ones that meet individual teachers' own needs against each of the priorities in the school improvement plan.

56.   New plans and initiatives were launched in this direction in the United Kingdom. Since 2005-2006, the Training and Development Agency for Schools (TDA) is in charge of coordinating professional development for all English school staff. In September 2007, new teaching standards were introduced in order to provide a framework for teacher evaluation in accordance with the school broader policies. The link is emphasised between what is expected from a 'good' teacher at each stage of the career on the one hand, and occasions for improvement towards the next career stage on the other hand. "The framework provides a backdrop to discussions about how a teacher's performance should be viewed in relation to their current stage and the stage they are approaching. The relevant standards should be looked at as a whole in order to help teachers identify areas of strength and areas for further professional development. A teacher who aspires to access to a higher career stage will need to reflect on and discuss how they might plan their future development so they can work towards meeting the standards, and performance management would provide evidence for the teacher's future application" (TDA, 2007b). The schools that associate the identified individual needs with the school priorities, and that also manage to develop the corresponding professional development activities, are likely to perform well (Ofsted, 2006).

57.   However, much remains to be done in this domain. Margo *et al.* (2008) emphasise that many problems are currently standing in the way of achieving a fully effective teaching workforce in England, among which are 'inconsistent quality of training' and 'inadequate professional development'. To overcome these problems, they recommend reinforcing again the link between continuing professional development (CPD) and the appraisal process, through more frequent evaluations, integrated CPD requirements, and obligations for teachers reviewed as poor performers to access appropriate training before they re-enter teaching.

58.   Seeing the evaluation procedures as a basis for future practice improvement is critical to implement a system in which every single teacher will feel concerned by the evaluation and the relevant professional growth opportunities, whatever the current level of performance. Evaluation procedures are certainly necessary for responding to ineffective teachers and ensuring that teachers adopt appropriate practices. Nevertheless, without a link to professional development opportunities, the evaluation process is not sufficient to improve teacher performance, and as a result, often become a meaningless exercise that

encounters mistrust – or at best apathy – on the part of teachers being evaluated (Danielson, 2001; Milanowski and Kimball, 2003; Margo *et al.* 2008; Pochard, 2008). As regards the French system, Pochard (2008) deplores that the professional development programmes are not shaped to constitute a response to the training needs clearly identified by both the teacher and the institution. It is argued that evaluation alone is not sufficient to implement the necessary changes to favour improvements in the efficacy and equity of the educational system. Also, it is argued that any evaluation highlighting dysfunctions in a school should result in the designing of a new educational plan supported by an external team.

3. ADVANTAGES AND DISADVANTAGES OF DIFFERENT APPROACHES TO EVALUATE TEACHERS

59.   This section discusses a number of contentious issues in designing and implementing teacher evaluation systems. The first controversial aspect in teacher evaluation schemes is whether or not student outcomes should be used as a measure of teacher performance. The second debate illustrates that the advantages and disadvantages of different methods are generally related to the purposes of teacher evaluation emphasised, and that, given resource restrictions, trade-offs between the arguments in favour of summative approaches and the ones for a formative system are inevitable. Finally, the arguments in favour or against the different approaches reflect different stakeholders' views, resulting in implementation difficulties.

3.1   Advantages and drawbacks of using student outcomes as a measure of teacher performance

60.   Student learning outcomes is an appealing measure to assess teaching performance, since the ultimate goal of teaching is to improve student learning. Not surprisingly, much research has focused on the use of student achievement as measured by standardised tests to evaluate teachers. For instance, Leigh (2007) recently examined the test scores in literacy and numeracy of three cohorts of students, and concluded that the changes in the relative positions of classes of students provided a basis for the identification of effective and ineffective teachers. Braun (2005) argues that considering student scores is a promising approach for two reasons: first, it moves the discussion about teacher quality towards student learning as the primary goal of teaching, and second, it introduces a quantitative – and thus, objective and fair – measurement of teacher performance. In this respect, the development of "value-added" models represents significant progress relative to methods based on the absolute proportion of students meeting a given achievement level. "Value-added" models are designed to control for the individual students' previous test scores, and therefore have the potential to identify the contribution an individual teacher made to students' achievement.

61.   In Florida, the "Special Teachers are Rewarded" (STAR) scheme links salary or bonus awards for individual teachers to value-added measures of student learning (Ingvarson, Kleinhenz and Wilkinson, 2007). Nevertheless, this type of link between a direct measure of performance and pay remains extremely rare, given the numerous statistical and theoretical challenges associated with the use of these methods. Indeed, Braun (2005) emphasises the marked contrast between the enthusiasm of those who would like to use such measurements, mainly policymakers, and the reservations expressed by the researchers who have studied their technical characteristics.

62.   Using student achievement on standardised tests to evaluate teacher performance presents numerous statistical challenges. Most authors (Lockwood, Louis and McCaffrey, 2002; Kupermintz, 2003; Braun, 2005; Aaronson, Barrow and Sander, 2007; Goe, 2007) are not convinced that the current generation of value-added models is sufficiently valid and reliable to be used for fairly evaluating individual teachers' effectiveness. Statistical limitations first refer to the noticeable lack of reliable data, mainly due to the fact that individual students rarely take annual standardised tests. Rowley and Ingvarson (2007) criticise Leigh (2007)'s methodology, which consists of creating a hypothetic test score in the missing data year at the midpoint of two available test results, arguing that it does not allow to fairly attribute the students' success to the different teachers involved. Second, when data are available, sampling variations can cause imprecision in test score measures; this problem is particularly striking in elementary

schools, where the limited number of students per classroom creates large idiosyncrasies of the particular sample of students being tested (Kane and Staiger, 2002).

63.   Broader methodological criticisms stress that value-added models, whatever their degree of sophistication, can neither fully integrate all factors influencing student achievement scores – qualitative by nature – nor reflect all student learning outcomes. Family background and support, school attendance, peer and classroom climate, school policies, availability of adequate materials, and children effects influence student learning (CAESL, 2004; Ingvarson, Kleinhenz and Wilkinson, 2007; Goe, 2007, Weingarten, 2007). Specific factors at the time of the test – "a dog barking in the playground, a severe flu season, a disruptive student in a class" – can also affect one student's results independently from his teacher's contribution (Kane and Staiger, 2002). Moreover, good teachers are likely to have an impact on children's achievement during several years after having taught to them; and conversely, after several years of ineffective teachers, students may never be able to catch up academically. These teacher 'cumulative effects' cannot be accurately measured at discrete points in time (Hanushek, 1986; Sanders and Rivers, 1996; CAESL, 2004). Finally, teaching impact on students is not restricted to areas assessed through student standardised tests, – generally limited to reading and numeracy –, but also include transfer of psychological, civic and lifelong learning skills (Margo *et al.*, 2008). While Xin, Xu and Tatsuoka (2004) tried to decompose single test scores into several categories of cognitive abilities in four countries (Japan, Korea, the Netherlands and the United States), they found that teachers' attributes used in pay decisions have no consistent positive impact on any type of cognitive skills, despite their attention to controlling for individual and family background. These are sources of skepticism for using such statistical methods.

64.   Theoretical limitations also need to be considered. First, a statistical correlation is not a causal relationship: the fact that teachers matter for student learning does not necessarily indicate that student learning is the result of good teaching. Second, the standardised tests used to differentiate students are not specifically designed for the purpose of assessing teachers. Following Popham (1997), Goe (2007) argues that they were not engineered to be particularly sensitive to small variations in instruction or to sort out teacher contributions to student learning. Thus they do not provide a solid basis on which to hold teachers accountable for their performance. Third, using student tests scores to evaluate teachers may induce unexpected distortions and constrictions in teacher behaviour towards the sole achievement on standardised tests. High-stakes incentive schemes based on standardised tests can incite teachers to concentrate exclusively on teaching areas assessed in the tests – therefore reducing the curriculum to the basic skills generally tested – (Jacob and Lefgren, 2005, Weingarten, 2007), incite teachers to concentrate on the specific students who are close to passing mark at the expense of children who are behind or ahead (Weingarten, 2007), and even provoke serious cases of teacher cheating on standardised tests (Jacob and Levitt, 2003; Jacob, 2005). Furthermore, test results may identify teachers who are ineffective or should professionally develop but do neither permit to fairly discriminate between the wide range of effective teachers nor identify which professional development activities should be established in order to improve their performance (Braun, 2005). Finally, it may lead to holding teachers responsible for the whole student performance whereas one should instead recognise that successful teaching is a shared responsibility between governments, schools and the teaching profession (Ingvarson et al., 2007).

65.   As a consequence, despite the attractiveness of the idea, there are numerous caveats against the use of student scores to evaluate teachers. In particular, there is a wide consensus in the literature around two specific directions: student outcomes should not be used as the *sole* measurement of teacher performance, and student outcomes should not be naively used for career decisions concerning the teacher, including the link to pay, because this incorporates a substantial risk to punish or reward teachers for results beyond their control (Kane and Staiger, 2002; Kupermintz, 2002; McCaffrey *et al.*, 2003; CAESL, 2004; Raudenbush, 2004; Braun, 2005; Ingvarson, Kleinhenz and Wilkinson, 2007; Rowley and Ingvarson, 2007). These rejections from teachers and scholars have materialized, for instance, in the New York State's legislature decision to ban the use of test scores in evaluating teachers in April 2008.

3.2   Designing a coherent set of methods and instruments aligned with the purpose of teacher evaluation emphasised

66.   Fenstermacher and Richardson (2005) distinguish two approaches to determine teacher quality: "successful teaching" is a measure of pure performance whereas "good teaching" focuses on the quality of opportunities provided for student learning in classrooms relative to teaching standards. Given that there is no direct, reliable and certain relationship between teacher quality and her students' achievement on standardised tests, there is a wide agreement to preferably evaluate teachers for their practice ("good teaching") rather than their performance ("successful teaching") (Ingvarson *et al.*, 2007). As a result, gathering *multiple* sources of evidence about teacher practice meets the needs for accuracy and fairness of the evaluation process, taking into account the complexity of what a 'good' teacher should know and be able to do (Danielson, 1996, 2007; Peterson, 2000).

67.   However, while the multiplication of instruments and evaluators is more likely to provide a solid basis to evaluate teachers, limited resources make trade-offs inevitable. Comprehensive teacher evaluation schemes imply greater direct and indirect costs at every stage of the process: reaching agreements on the design of the system requires time for discussions and consultations with all stakeholders (Avalos and Assael, 2006); training evaluators is expensive (Danielson, 1996, 2007); conducting evaluation processes induces additional workload for both teachers and evaluators, unless offsetting is made by reducing workload with other responsibilities (Heneman *et al.*, 2006); aligning broader school reforms such as professional development opportunities requires more educational resources (Heneman *et al.*, 2007; Margo *et al.*, 2008). For these reasons, countries have often decided to implement only some of the aspects of evaluation schemes. But making trade-offs between different methods is not simple since the advantages of particular instruments or methods for summative purposes are generally disadvantages for formative purposes, and vice-versa.

68.   With regard to the sources of evidence, instruments such as student outcomes, teacher tests, questionnaires and surveys completed by parents and students, and classroom observations are more summative in nature, whereas interviews with the teacher and documentation prepared by the teacher are generally more useful for formative purposes. Accountability mechanisms require to quantitatively and objectively rate teachers according to a unique framework composed of few professional items (Nabors Oláh, Lawrence and Riggan, 2008). This allows to compare teachers in conformity with identical well-defined standards, and to easily aggregate the scores obtained on the different criteria. A global index of teacher performance obtained this way may provide a fair scale to reward and celebrate teacher work. By contrast, when the purpose is to help teachers improve their practices and provide them with professional growth opportunities, qualitative and customised instruments and criteria are preferred. For a formative purpose, adapted collection of evidence is more adequate than one set of standards to fit all possible situations. It must allow both to identify domains of strength and weakness in teaching and to give the teacher a constructive feedback including possible ways of improvement, according to the teacher's level of experience and the school context.

69.   Portfolios are particularly interesting to the extent that they contain artifacts of teacher work which can be differently combined according to the purpose emphasised. On the one hand, Klecker (2000), Campbell *et al.* (2000) and Tucker *et al.* (2002) argue that portfolios provide assessment information to hold teachers accountable for meeting educational standards. On the other hand, Darling (2001) argues that teacher development should take precedence in designing portfolios and that 'narrative reflection' is the best way to foster such development. Beck, Livne and Bear (2005) observe that formative portfolios that focus on teacher development better support professional outcomes, and consequently argue that portfolios should not be used for the summative assessment of teachers. Another question is whether electronic portfolios are most effective and fair than paper portfolios. Electronic portfolios facilitate a constant access to work samples and prompt feedback from evaluators, permit to capture on-going reflection of teachers'

professional development, present an easy display of multiple data points (Wetzel and Strudler, 2006; Jun et al., 2007), while a drawback compared to the paper portfolio is that a teacher with greater technology skills is at an advantage, even if her teaching capacities are precisely the same (Wetzel and Strudler, 2006). Strudler and Wetzel (2008) also emphasise the significant amount of time and effort expended on the creation and the revision of portfolios and the potential lack of compatibility with institutions' beliefs, values and concerns.

70.   With regard to evaluators, it is widely agreed that school principals, peers and experts, parents and students do not value the same teaching capacities and knowledge, do not refer to the same collection of evidence, and have different perceptions and degrees of objectivity. Consequently, the participation of *multiple* evaluators is often seen as a key to successful practices; at least more than one person should be involved in judging teacher quality and performance (Peterson, 2000; Stronge and Tucker, 2003). Danielson and McGreal (2000) explain that the '360-degree evaluation systems', which incorporate the participation of many kinds of evaluators, support the idea that a teacher's competence may be seen from several different perspectives and that it should be exemplary (or at least adequate) from all those different angles.

71.   The literature distinguishes the relative advantages and drawbacks of including different types of evaluators according to the nature of the evaluation process. Principals seem to be particularly effective at identifying the very best and the very worst teachers (those at the top and bottom 10-20 percent) (Jacob and Lefgren, 2005a, 2008), and their supervision and leadership role *a priori* put them in a position to make assessments of teachers. This suggests that they can be relevant evaluators for summative assessments. However, evidence also suggests that principals show very little ability to distinguish between teachers in the vast majority of the distribution, systematically discriminate on the basis of some teacher characteristics (*e.g.* gender, age, education), and are often influenced by a number of affective or non-performance factors such as the likability of the subordinate or the first appraisal they made on the particular teacher being evaluated (Lefkowitz, 2000; Bolino and Turnley, 2003; Jacob and Lefgren, 2005a, 2008). Levin (2003) and MacLeod (2003) demonstrate that principal's compression and leniency in performance evaluations, often found in practice, are features of the optimal contract between a risk-neutral principal and a risk-averse agent when rewards are based on subjective performance evaluation. Finally, it is difficult to consider principals as impartial judges since they are in day-to-day contact with the teachers being evaluated. These arguments show that a principal-based evaluation system may be useful for dismissal decisions but may not be accurate enough to be linked to pay increments. By contrast, external reviewers assess teachers relative to frameworks of professional standards, know the specificities of content and skills for each teaching area, but are relatively less able to adapt the process to the school context, problems and values (Anderson and Pellicer, 2001).

72.   In relation to the formative purpose, there are also debates about who is in the position to accurately define teachers' needs for improvement and provide the most constructive feedback. Peers and colleagues who have the same characteristics, teach the same subject to the same student grade are more likely to obtain the confidence of the teacher being evaluated. The teacher may therefore more easily engage in self-reflection about her practices, and express her feelings and concerns during interviews, without fearing potential sanctions. Peers can also provide qualitative feedback based on their own experience in the relevant teaching area. But principals are essential to link the teacher revealed needs for improvement to the further professional development opportunities according to school goals, and other personnel needs. They are also more likely to provide informal continuing feedback to the teacher throughout the year and not only during the formal evaluation process. More generally, they are essential to make performance improvement a strategic imperative, and help considering teacher evaluation indispensable to teacher and school broader policies (Heneman *et al.*, 2007; Robinson, 2007; OECD, 2008).

3.3   Difficulties in implementing teacher evaluation schemes

3.3.1   Conflicts of interest between different stakeholders

73.   The choice of evaluation methods is made even more complex by significant divergences of views and interests between different stakeholders. The relative importance of the summative and formative purposes is particularly contentious. On the one hand, policymakers and parents tend to value quality assurance and accountability. "They make the point that public schools are, after all, public institutions, supported by tax payer money, and that the public has a legitimate interest in the quality of the teaching that occurs there. It is through the system of teacher evaluation that members of the public, through their legislators, local boards of education, and administrators, ensure the quality of teaching. A parent, in other words, in entrusting the education of a child to the public schools, has a right to expect a certain minimum level of performance." (Danielson and McGreal, 2000). On the other hand, teachers and their unions expect opportunities of social recognition of their work and opportunities for professional growth through the development of a formative system of teacher evaluation (Avalos and Assael, 2006).

74.   Teachers also often reject reforms shaped by policymakers because they consider them as incompatible with their own concerns and day-to-day practices. Kennedy (2005) argues that "highly dedicated" teachers' reform rejections do not come from their unwillingness to change or improve, but from "the sad fact that most reforms don't acknowledge the realities of classroom teaching". Reformers have high expectations regarding rigorous and important school content, intellectual engagement, and universal access to knowledge. But reforms fail, mainly because "the circumstances of teaching prevent teachers from altering their practices". For instance, teachers encounter daily constraints and unexpected events, face academic calendar pressures, and serve a heterogeneous and 'compulsory clientele'.

75.   Table 1 summarises the arguments in support of and against teacher evaluations, according to some key dimensions.

Table 1. A summary of the arguments for and against teacher evaluation as portrayed in the literature

	Arguments supporting teacher evaluation	Arguments against teacher evaluation
Accountability	The current system does not hold teachers accountable for their practice and performance. It is a bureaucratic meaningless exercise that needs change.	Designing a fair and accurate evaluation system for accountability purposes is vain because performance cannot be determined objectively and 'good teaching' can take several forms.
	Local authorities and parents have the 'right' to institute quality assurance mechanisms. Performance review increases political and public support for education systems.	Teaching needs a safe environment, far from political, social and business pressures.
	Evaluations allow the identification of good performers in a way similar to other markets.	Market mechanisms have no place in education.
Incentive mechanisms and links to recognition and rewards	Evaluations provide a basis for pay increments that depart from the single salary schedule solely based on experience. Essential to make the profession more attractive.	Teachers are not motivated by financial rewards but by 'intrinsic' aspects (*e.g.* desire to teach, work with children) and favourable working conditions (*e.g.* flexible schedules). Some teachers can be discouraged as a result of evaluation procedures.
	There is a need to respond to ineffective teachers.	It is a patronising policy which stigmatises teachers.
Professional development	Evaluations allow teachers to identify strengths and weaknesses in relation to school goals and to assess professional development needs. Essential to keep teachers motivated by their work.	- The choice of professional development activities should not stem from evaluation results but be made by the teacher unilaterally. - Schools do not provide professional development activities in areas in which improvement is needed.
Cost	The current system wastes time, energy and money.	A comprehensive teacher evaluation scheme is expensive and time consuming.
Effects	Evaluations enhance teacher practice and improve student learning.	Evaluations produce a range of negative effects such as the narrowing of the curriculum and the neglect of some students.
	Teacher evaluations enhance co-operation between teachers – through professional discussions and a sharing of their practice – and between teachers and school leaders – from whom they expect feedback and coaching.	Teacher evaluations reduce teacher co-operation between teachers – because of competition effects – and between teachers and school leaders – because of hierarchical or 'appraiser-appraisee' relationships.

76.   While the arguments in favour of evaluations generally come from policymakers and the arguments against from teachers' unions, the reality is less clear-cut. For instance, the American Federation of Teachers and the National Education Association, the two largest US teacher unions (Hess and West, 2006), promote the NBPTS' Certification as "a proven way to strengthen the skills, knowledge, professionalism and recognition of teachers" (AFT and NEA, 2008). However, as mentioned earlier, the NBPTS evaluation is voluntarily undertaken by teachers who want to examine their practice against the profession's highest standards, and does not have any negative consequence. Unions' response could have been different for a compulsory generalised evaluation scheme in which all teachers would be tested, whatever their level or quality, or in which potential dismissals or deferrals of promotion would be at stake.

77.   In the same way, there is a large heterogeneity in teachers' responses to evaluation schemes. For instance, younger teachers may be more likely to accept summative evaluations and links between evaluation results and pay than more experienced ones who are at a higher level in the salary scale. Studying new teachers' potential acceptance of performance-based pay or knowledge- and skill-based pay, Milanowski (2007) found that beginning teachers may view new pay strategies more favourably that their more experienced colleagues. Similarly, teachers who have invested a considerable amount of time and money into obtaining the educational credits required by the single salary schedule are more likely to reject the reform (Odden and Kelley, 2002). Among older teachers, heterogeneity is also striking. Day and Gu (2007) found that teachers teaching for more than twenty-five years were in extreme professional scenarios: one sub-group "showed a continuing interest in updating and improving their classroom knowledge" while the other sub-group "increased feelings of fatigue and disillusionment". This split could imply that the former are much more likely to accept – or even promote – review of, and feedback on, their practice than the latter.

3.3.2   Ways to overcome obstacles

78.   In spite of the challenges, time and resources should be dedicated to the (re)design and implementation of a well-accepted teacher evaluation system. Otherwise, concerns associated with current wasteful and demotivating bureaucratic procedures will remain (Danielson, 2001; Milanowski and Kimball, 2003; Holland, 2005; Marshall, 2005). The literature points out the crucial importance of the following elements to overcome obstacles to implementation.

79.   ***Engaging in dialogue and consultations.*** The initial conception of the system should include the wide participation of all key actors, especially teachers and their unions, from the start of discussions (Avalos and Assael, 2006). Teachers will accept more easily to be evaluated if they are consulted in the design of the process. In addition to taking their fears and claims into account, the participation of teachers recognises their professionalism, the scarcity of their skills, and the extent of their responsibilities (Hess and West, 2006), as well as their indispensable position to estimate the feasibility and the relevance of the teacher evaluation system. Teachers are the 'technical core' of education systems and their engagement is essential in the development of evaluation systems given the depth of their professional knowledge and practice. If teacher evaluation procedures are unilaterally designed at the level of the administrative superstructure, without addressing and including the core of teaching practice, then there will be a 'loose coupling' between administrators and teachers, that will both fail to provide public guarantees of quality, and will discourage reflection and review among teachers themselves (Elmore, 2000; Kleinhenz and Ingvarson, 2004). Thus, administrators and teacher unions need to work hand in hand to create the level of trust and co-operation required to enable reform to move forward in a productive way (Odden and Kelley, 2002).

80.   ***Supporting teachers in understanding and appropriating the evaluation.*** Guaranteeing that teachers are provided with support to understand the evaluation procedures is also vitally important. Teachers must know what is expected from them to be recognised as 'good' teachers before the process starts. This requires not only complete transparency in the evaluation criteria and procedures but also ensuring that teachers appropriate the process through support and coaching. For instance, the *Guide to understanding National Board Certification* responds to teachers' concerns in relation to the characteristics of the NBPTS evaluation. It is important to explain the system (who is concerned, what the process consists of, how the scores are established, etc.) and to give advice to help teachers succeed (what to include in a portfolio, which exercises to get prepared, examples and ideas from past candidates and trainers) (AFT and NEA, 2008).

81.   ***Conducting a pilot implementation before the full implementation.*** Conducting a pilot implementation is a cost effective way to ensure the viability and reliability of the system before full

implementation. Associated with key actors' gathered perceptions, it allows the review of the process and adjustments in light of potential flaws. Heneman *et al.* (2006) argue that at least one pilot year is needed to work the glitches out of the evaluation systems. However, going to scale after the pilot sometimes reveals other implementation problems which in turn lower the credibility of the system to teachers and reduce acceptance. Therefore, caution in choosing representative schools or teachers for the pilot implementation is required.

4. EMPIRICAL EVIDENCE ON THE EFFECTS OF TEACHER EVALUATION SCHEMES

82.   This section summarises the current empirical evidence of the effects of teacher evaluation on teacher practice and student learning. First, it highlights the difficulties in assessing quantitative effects associated with teacher evaluation schemes, and the resulting mixed findings. Then, it considers the circumstances under which teacher evaluation systems seem to be more effective.

4.1   Quantitative evidence

4.1.1   Difficulties in measuring teacher quality and the impact of teacher evaluation

83.   The literature largely establishes that teachers matter in student outcomes, in the sense that they are powerful contributors to students' academic achievement (Vandervoort *et al.*, 2004; OECD, 2005). However, the literature is more hesitant in demonstrating which teacher aspects are relevant to teacher quality and what is the relative importance of teacher quality *vis-à-vis* other factors that theoretically influence student learning, including family, student and school factors. Aaronson, Barrow and Sander (2007) stress that the literature on this subject "remains somewhat in the dark" in spite of data improvements. Not surprisingly, measuring the impact of teacher evaluation in terms of student learning through 'education production functions' is even more difficult.

84.   Measuring the effect of teacher evaluation faces a number of challenges. First, it needs to control for the broad set of qualitative variables which are likely to influence student learning. These variables encompass teacher characteristics (*e.g.* age, gender), teacher education and experience, students' family factors (*e.g.* parents' background, parents' support), school factors (*e.g.* school policies, school incentives, peer and classroom effects) and student factors (*e.g.* motivation, cognitive abilities, cumulative experience). The complex realities of education prevent researchers from accurately assimilating these factors as traditional inputs into production functions (Hanushek, 1986). Second, because of its qualitative and heterogeneous nature, the output itself – student learning – is not a traditionally measurable 'end product', and this makes the decomposition between different factor contributions even more difficult (Hanushek, 1986; Ingvarson *et al.*, 2007). This does not mean that doing any quantitative study in education is vain but rather than it requires particular attention to analytical issues or potential misinterpretations of the results. A particular focus should be placed on the fact that each factor omission or measurement problem – including lack of data – creates a potential quantitative bias in the estimated relationship between teacher quality and student achievement (Xin, Xu and Tatsuoka, 2004).

85.   As a consequence, the empirical literature that primarily indicated that teacher evaluation may have an important role in student learning came from a process of elimination. By contradicting or restricting the respective roles of individual teachers' apparent features (whether characteristics, education, experience, or financial incentives), numerous studies concluded that it was teacher practices – and, by extrapolation, evaluation of these practices – that indeed matter. The first influential contribution was Hanushek's distinction between observable aspects of teachers, such as teacher background, gender, or race, and teachers' unquantifiable "skills" (Hanushek, 1986, 1992). According to Hanushek, if the previous literature has found no significant impact of teacher quality on student achievement, it was because it concentrated on observable attributes of teachers – teacher's holding of a master degree for example – while teacher quality was instead related to their "skills" or "idiosyncratic choices of teaching and methods" (such as

classroom management, methods of presenting abstract ideas, communication skills, and so forth), i.e. their practice.

86.   Relying on Hanushek's model, a body of new empirical studies provides evidence of little relationship between teacher attributes and student outcomes in order to support the evaluation of teacher practice. For instance, Munoz and Chang (2008) stress the vague relationship between teacher education and years of experience and student academic growth, with the intention of raising the profile of the review of teachers' instructional practices among educational policies. Aaronson, Barrow and Sander (2007) found that traditional measures relative to teachers, including the ones that determine current teacher compensation, explain little of the variation in estimated quality. These findings question the current single salary schedule based on teacher education and experience, at the benefit of pay increments that consider broader determinants, including the link to appraisal systems of teacher practice and performance.

4.1.2   *Mixed empirical evidence on teacher evaluation systems*

87.   Unlike the studies mentioned in the previous subsection, the following refer to the quantitative evidence surrounding teacher evaluation systems *per se*. However, it should be noted that most of the empirical research does not focus on estimating the direct effects on student outcomes (neither considering one particular teacher evaluation system nor longitudinally). Rather, the quantitative literature primarily encompasses the two following categories of empirical studies.

88.   The first subgroup examines the variation in the statistical relationship between teachers and student outcomes when teachers pass one particular evaluation process and when they do not. This body of evidence does not assess the effects of teacher evaluation on student outcomes, since it compares two distinct groups of teachers (one subject to evaluation, the other not) instead of comparing the impact of one particular group of teachers on student outcomes before and after the considered evaluation process. Rather, it provides an indication of the capacity of the implemented evaluation process to effectively differentiate proficient teachers from other teachers. These studies are essential since they establish the viability and the reliability of an evaluation scheme, indispensable for fairness in summative procedures and in the potential link to rewards.

89.   For instance, numerous studies examined the viability of the NBPTS' evaluation system, because it both represents one of the most complex and comprehensive approaches to teacher evaluation, and leads to a formal recognition – the National Board Certification (NBC). A number of authors (Cavalluzzo, 2004; Goldhaber and Anthony, 2007; Vandervoort *et al.*, 2004; Smith *et al.*, 2005) found that students of teachers who have obtained the NBC do better on standardised tests than students of non certified teachers. This indicates, first, that teacher practices are important for student achievement, and second, that the NBC correctly identifies the teachers who have adopted the best practices. Moreover, Goldhaber and Anthony (2007) and Cavalluzzo (2004) also conclude that student scores particularly improved for minority students and special needs students, thus suggesting that the NBC properly identifies teachers who adopt the practices which enhance educational equity in addition to overall efficacy. However, other authors (McColskey and Stronge, 2005; Sanders *et al.*, 2005; Harris and Sass, 2007) showed, by contrast, that students of teachers who obtained the NBC did not perform significantly better than other students, in spite of improvements in some grades or areas.

90.   The empirical evidence is also mixed for systems of compulsory teacher evaluation. Milanowski (2004) estimated the relationship between teacher evaluation ratings and a measure of value-added student achievement for the US district of Cincinnati, which has implemented a comprehensive standards-based teacher evaluation scheme as a basis for a knowledge- and skills-based pay system. He found significant positive correlations, and concluded that if scores from a rigorous teacher evaluation system are substantially related to student achievement, then this provides validity evidence for the use of the teacher

scores as a basis for a financial reward system. Borman and Kimball (2005) studied the teacher evaluation system of the district of Washoe County, with a two-level model. After controlling for student background and teachers' experience, they assessed the relation between teacher quality as measured by the evaluation system and both overall classroom mean achievement and within-classroom effects on social equality. They found that teachers with high evaluation scores are related to better student learning outcomes across grades and teaching areas (reading and math). But these teachers do not appear to be reducing gaps in achievement between low- and high-achieving students and students from low-income or minority background. This is a source for skepticism when looking at the validity of the evaluation system to distinguish between teachers who adopt practices directed towards equity and those who do not.

91.   The second subgroup of quantitative studies on teacher evaluation systems focuses on the effects on the enhancement of teacher practice and motivation, as perceived by the teacher who is evaluated. If teachers report enhanced practices owing to the evaluation process – and assuming that the corresponding practices are relevant to student learning –, then the evaluation system is supposed to be effective at indirectly improving student outcomes. For instance, by requiring from teachers the creation of portfolios and reflection about their practices, the NBPTS provides a cost-effective opportunity for professional development through the evaluation process (Cohen and Rice, 2005), by leading the teacher to focus on what makes a strong curriculum and what makes an accurate assessment of student learning. Several authors (Bond *et al.*, 2000; Lustick and Sykes, 2006) stress that teachers applied in the classroom what they learnt from the evaluation process. Teachers seem to have also gained new enthusiasm for the profession – regarding how long they plan to stay in teaching – as a result of going through the evaluation process (Vandervoort *et al.*, 2004; Lustick and Sykes, 2006; Sykes *et al.*, 2006; NBPTS, 2007). Finally, the accomplished teachers who go through the evaluation process are likely to contribute to school leadership by adopting new roles including mentoring and coaching of other teachers who recognise certified teachers as helpful (Petty, 2002; Freund et al., 2005; Sykes et al., 2006). These studies bring considerable insights in the formative aspect of teacher evaluation system.

92.   As mentioned above, very little evidence exists about the direct correlation between teacher evaluation and student achievement. Figlio and Kenny (2007) attempted to introduce teacher evaluation as a teacher incentive mechanism towards student achievement – alongside measures of financial incentives – in a longitudinal regression. They used the US longitudinal database on schools to distinguish between schools which evaluate their experienced teachers annually from the ones that do so less frequently, with the expectation that more frequent performance review improves teacher performance. They also controlled for a broad set of student and school variables, in addition to the teacher financial incentives and threat of dismissals. Unfortunately, they still found that teacher evaluation was not statistically significant whereas financial incentives were positively and significantly correlated with student achievement.

93.   Figlio and Kenny's results underline the difficulty to estimate the direct impact of teacher evaluation on student achievement. But the fact that financial rewards are identified as positive incentive mechanisms towards student success suggests that a teacher evaluation system is indispensable, since it allows the identification of the effective teachers that will be rewarded. Moreover, it may not be the frequency of evaluation but the quality and the sophistication of the evaluation process that matter. That probably explains why empirical studies often focus on the design of successful evaluation schemes rather than on quantitative measurements of their impacts.

4.2   Qualitative evidence

94.   Evidence also indicates that teacher evaluation seems to be more effective under particular circumstances.

95.   ***Involving teachers at every stage of the process.*** Teachers should be consulted on the strengths and the flaws of the system, from its design to its full implementation and review. First, teachers must agree with the framework which defines the standards of the profession. "Active teacher participation in the construction and refinement of the model is essential" (Heneman *et al.*, 2006). Danielson's *Framework for Teaching* (1996, 2007), or the standards developed by the NBPTS could be used as starting points, with further adjustments to meet the local educational goals. The creation of Teaching Councils, as in Ireland in 2005-2006, provides great opportunities to involve teachers in the setting of high-level profession-led standards, and more generally, to fully integrate teachers in (re)defining the profession for further policy development (OECD, 2005).

96.   Second, *all* teachers must be supported in understanding what the evaluation expects from them to be recognised as good teachers and in preparing adequately for the evaluation process. This requires both transparency on the methods used and coaching towards empowerment evaluation. As put by Heneman *et al.* (2006) "For teachers, early training should focus on the nature of the performance competencies on which the system is based, the purposes and mechanics of the evaluation system, and knowledge and skills needed to function effectively within the new system". Clear expectations related to the evaluation process and the corresponding teacher training may be integrated in the school principal's leadership actions (Ovando and Ramirez, 2007). The need for removing the 'loose coupling' between the administrative superstructure and the technical core of teaching (Ingvarson *et al.*, 2007), i.e. the transition from the current bureaucratic procedures where evaluation is *done to* teachers to an evaluation made *with* teachers, asks for the full appropriation of the reform by the teachers.

97.   Finally, teachers should also be provided with opportunities to express their perceptions and concerns on the evaluation process after the system is installed. Interviews and surveys are common methods used by analysts to collect teacher reactions on the evaluation system. The items generally include the understanding of the process, the acceptance of the standards, the fairness of the process and of the results, the capability and objectivity of the evaluator, the quality of the feedback received, the perceived impact of the evaluation process on teaching and the overall impression of the the system (Milanowki and Heneman, 2001, 2004; Kimball, 2002). Milanowski and Heneman (2001) found that teachers' overall favourableness toward a system newly implemented in a medium-sized US school district was correlated with acceptance of the teaching standards, the perceived fairness of the process, the qualities of the evaluator, and the perception that the evaluation system has a positive impact on their teaching. Teachers received 1.5 days of training to understand the domains and standards evaluated, to get acquainted with the aspects evaluators look at, and were provided with information about the development of a portfolio.

98.   ***Training evaluators.*** The literature largely agrees on the need for an in-depth training for the evaluators. First, evaluators should be trained to rate teachers according to the limited evidence they gather, the criteria of good teaching and the corresponding levels of teacher quality. This is particularly important when the evaluators are school principals, which may have limited knowledge on the content and pedagogical skills needed for the subject taught by the teacher being evaluated. Second, evaluators should be trained to provide constructive feedback and coaching to the teacher for further practice improvement.

99.   ***Releasing both evaluators and teachers from other tasks.*** Comprehensive teacher evaluation systems require time and other resources. This may be costly but is indispensable for designing a consistent and fair system, approved and appropriated by the teachers. A consequence is that both teachers and evaluators should be partly released from other duties. Milanowski and Heneman (2001) found that even if teachers accept the standards and the need for an evaluation system, they may also manifest reluctance when the system adds too much to their workloads. Hence, teachers should have time to reflect on their own practice, especially when the process requires the constitution of a portfolio. As emphasized by Heneman *et al.* (2006), "System designers need to carefully review what is required of teachers to minimise burden. (…) Perhaps some small reduction in other responsibilities while teachers are

undergoing evaluation would decrease the perception of burden and sense of stress." Policymakers should also aim at reducing the administrative workload for evaluators, especially school principals, in order to provide them with more time for teacher evaluation, feedback and coaching (Marshall, 2005).

100.	***Conducting a pilot implementation and a continuous review of the process***. A pilot implementation is a cost effective way to ensure that the system is efficient, fair and consistent with local needs before a full implementation. Interviewing teachers during the pilot implementation is essential to correct the potential flaws and concerns related to the system. Researchers or practitioners should also concentrate on validity and reliability studies. Pecheone and Chung (2006) show that score consistency should be cautiously examined. For example, the standards-based scores are relevant if they correspond to the holistic ratings of the teachers' performance and if supervisors familiar with the teachers heavily agree on the level of performance. Milanowski and Heneman (2001) and Milanowski (2004) emphasise the necessity to control the reliability of the evaluators via a 'calibration process' that consists of comparing experimental ratings of some evaluators with expert judges and discussing differences. Only accredited evaluators should be allowed to evaluate teachers after the full implementation. Reviews of the process should also be conducted after the full implementation. Teachers are more likely to accept the process today if they know that they will be able to express their concerns and provide advice on the necessary adjustments as the process evolves.

101.	***Blending teacher evaluation into broader teacher quality and support policies.*** Evaluation should not replace other means of guaranteeing teacher quality, such as teacher education and licensing programmes, induction programmes, professional development and continuous informal feedback and advice, and broader recognition of teacher expertise and commitment to work (AFT, 2001; Corcoran, 2007).

5. CONCLUSION

102.   The first section of the paper emphasised the wide variety of teacher evaluation schemes in the OECD area. Countries largely differ in each of the relevant features, including the respective roles of key stakeholders, the scope and purpose of the evaluation, the methods and instruments used to assess teachers, as well as the criteria of 'good' teaching, and the links to rewards or professional development. In addition to the debates on particular features and on the consequences for teachers' careers, teacher evaluation appears to be particularly contentious in countries where teacher evaluation stems from required bureaucratic procedures instead of being an integral part of broader teacher and school policies.

103.   More and more countries are showing a growing interest in implementing comprehensive teacher evaluation systems, as a response to the demands for high educational quality. Although little empirical evidence is currently available, the literature primarily agrees on the need for clarifying the purpose emphasised and the importance of including a diverse set of evaluators and criteria to better reflect the complexity of defining what good teaching is. There is also a broad consensus about the involvement of teachers throughout the development of the evaluation process. An effective, fair and reliable evaluation scheme requires teachers' overall acceptance and appropriation of the system. Developing a comprehensive approach may be costly but is critical to conciliate the demands for educational quality, the enhancement of teaching practices through professional development, and the recognition of teacher knowledge, skills and competencies.

ANNEX 1: CONCEPTUAL FRAMEWORK FOR TEACHER EVALUATION

Key agencies or organisations involved / Stakeholders:

- National governments (Ministries / Departments of Education)
- Decentralised authorities in charge of educational policies (districts, municipalities)
- School leaders
- Teachers and Teacher Unions
- Parents / Students

Scope of evaluation / Teachers evaluated:

- Whole country *vs.* procedures on a regional basis
- School type: public schools, private schools
- Periodicity of evaluation: part of the regular work *vs.* evaluation in special cases (promotion, complaint)
- Compulsory *vs.* voluntary
- All teachers are the subject of the same evaluation *vs.* customised evaluation according to the teacher's experience
- Pilot implementation *vs.* full implementation

Evaluators:

- Internal reviews (by principals or senior school staff)
- External reviews (by peers or accomplished teachers within the same teaching content area)
- Self-evaluation
- Parents
- Students

Criteria and standards:

- Content knowledge on the subject taught
- Pedagogical skills
- Knowledge of students
- Ability to enhance student performance
- Competence in instruction planning
- Knowledge on assessing student learning
- Ability to create a favourable classroom environment
- Ability in managing classroom procedures
- Capacity to engage students in learning and to interact with them
- Communication and monitoring skills
- Ability to meet the needs of diversified student populations; demonstration of flexibility and responsiveness
- Professionalism: communication with families, school staff and leaders
- Engagement in professional growth and development: reflection on teaching, in-service training

EVALUATION OF TEACHER PRACTICE AND PERFORMANCE

Methods and instruments:

- Classroom observations
- Interviews with the teacher
- Teacher-prepared portfolios (video clips, lesson plans, reflection sheets, self-reported questionnaires, samples of student work)
- Student achievement results (absolute performance or value-added gains)
- Teacher tests
- Data from questionnaires and surveys completed by parents and students

---

Summative assessment:

- **Accountability and quality assurance for policymakers and parents**
  - Improving student learning and performance through better teaching practices
  - Reducing inequity in student achievement
- **Recognition and/or rewards for teachers:**
  - Recognition of skills and commitment
  - Promotion
  - Salary increments
  - Non financial rewards (working conditions)
  - Responses to ineffective teachers (deferrals of promotion, dismissals)
  - → **Making teaching an attractive career choice**
  - → **Retaining effective teachers in schools**

Formative assessment:

- **Professional development to enhance teaching:**
  - Identifying the teacher's strengths and weaknesses
  - Providing constructive feedback on the teacher's practices
  - Guiding teachers towards adequate professional development programmes and opportunities to develop their capacities
  - → **Keeping teachers motivated throughout their careers**
- **Improving school leadership:**
  - Adapting schools' professional development programmes to identified needs
  - Improving teacher monitoring and coaching from principals
  - → **Engaging teachers in policy development and implementation**

ANNEX 2: EXAMPLES OF TEACHER EVALUATION SYSTEMS IN OECD COUNTRIES

1. Teacher evaluation for summative purposes with links to pay: The US District of Cincinnati
[Milanowski, 2004]

Context: Cincinnati is a large urban district with 48,000 students and 3,000 teachers in more than 70 schools and programmes. Its average level of student achievement is low compared to the surrounding suburban districts. Cincinnati has also had a history of school reform activity, including the introduction of new whole-school designs, school-based budgeting, and teams to run schools and deliver instruction. The union-management relationship has generally been positive. Like many other urban districts, state accountability programmes and public expectations have put pressure on the district to raise student outcomes.

Implementation: In response to the obsolescence of the existing teacher performance evaluation system, and ambitious goals for improving student achievement, the District designed a knowledge- and skill-based pay system and a new teacher evaluation system during the 1998-1999 school year. The assessment system was piloted in the 1999-2000 school year and is used for teacher evaluation district wide since the 2000-01 school year.

Criteria: The assessment system is based on a set of teaching standards derived from the Framework for Teaching (Danielson, 1996). Seventeen performance standards are grouped into four domains: (i) planning and preparation; (ii) creating an environment for learning; (iii) teaching for learning; and (iv) professionalism. For each standard, a set of behaviourally anchored rating scales called rubrics describe four levels of performance: unsatisfactory, basic, proficient, and distinguished.

Instruments: Teachers are evaluated using the rubrics based on two major sources of evidence: six classroom observations and a portfolio prepared by the teacher. The portfolio includes artifacts such as lesson and unit plans, attendance records, student work, family contact logs, and documentation of professional development activities.

Evaluators: Four classroom observations are made by a teacher evaluator hired from the ranks of the teaching force and released from classroom teaching for three years. Principals and assistant principals do the other two observations.

Aggregation of scores: Based on summaries of the six observations, teacher evaluators make a final summative rating on each of the standards in domains (ii) and (iii), whereas principals and assistant principals rate teachers on the standards in domains (i) and (v), primarily based on the teacher portfolio. Standards-level ratings are then aggregated to a domain-level score for each of the four domains.

Scope and frequency of the evaluation: The full assessment system is used for a comprehensive evaluation of teachers in their first and third years and every five years thereafter. A less intensive assessment is done in all other years, conducted only by principals and assistant principals and based on more limited evidence. The annual assessment is intended to be both an opportunity for teacher professional development and an evaluation for accountability purposes.

Training on the evaluation process: Both teachers and evaluators receive considerable training on the new system. Evaluators are trained using a calibration process that involves rating taped lessons using the rubrics and then comparing ratings with expert judges and discussing differences. To ensure consistency among evaluators, the district eventually requires that all evaluators, including principals, meet a standard of agreement with a set of references or expert evaluators in rating videotaped lessons. Since the 2001-02 school year, only those who meet the standards are allowed to evaluate.

Direct consequences: For beginning teachers (those evaluated in their first and third years), the consequence of a poor comprehensive evaluation could be the termination of the contract. For tenured teachers, consequences of a positive evaluation could include eligibility to become a lead teacher. A poor evaluation could lead to placement in the peer assistance programme and to the eventual termination of the contract.

Link to pay: The performance evaluation system was designed in part to provide the foundation for the knowledge- and skill-based pay system. This system defines career levels for teachers with pay differentiated by level. The new pay system was originally scheduled to come into effect in the 2002-03 school year, resulting in relatively high stakes evaluations for the district's teachers. However, the link between the evaluation system and pay was rejected by teachers in a special election held in May 2002.

2. Teacher evaluation for formative purposes and as part of broader school policies

2a. Finland [UNESCO, 2007]

Context: In Finland, school teachers have positions comparable to national or municipal public servants. However, school leaders are in charge of teacher selection – once the required license is obtained – and in charge of all the policies that are considered as necessary to the enhancement of teaching quality, among which teacher evaluation. Finland is a paradigmatic case where the former system of 'teachers and schools inspection and supervision' was removed in 1990 but not replaced by another similar external system. As a consequence, teacher evaluation currently goes hand in hand with other policies within each particular school.

Methods / Evaluators: The Finnish scheme of teacher evaluation is characterised by the very high level of confidence placed in school and teacher competencies and professionalism as a basis to improve teaching quality. Thus, teacher self-evaluation is considered as a prime means of professional optimisation. School leaders also have a crucial role in engaging teachers in self-reflection about their own practice, and in developing a culture of evaluation alongside ambitious goals, according to the school context and challenges. The majority of schools have implemented annual discussions between school leaders and teachers to evaluate the fulfillment of the personal objectives set up during the previous year and to establish further personal objectives that correspond both to the analysis of the teacher and the needs of the school.

2b. England [Ofsted, 2006; TDA, 2007]

Context: The English system was originally designed with summative purposes, aiming at evaluating teachers' performance, and providing them with opportunities to access a higher career stage and the corresponding pay scale. However, numerous concerns about the fairness of the process and the potential perverse impacts of the procedure on teacher performance itself were addressed (Kleinhenz and Ingvarson, 2004). Hence, the recent developments of the system – including new professional standards from September 2007 – indicate an increased formative approach, embodied by a willingness to reinforce the link between the teacher appraisal system and teacher professional development needs relative to the

school goals. More generally, the system, completed within a wider framework for the whole school workforce, aims to improve school leadership and to be an integral part of the school's broader policies.

Scope/Methods: The evaluation is differentiated according to the career stage of the teacher being evaluated. Five professional stages are identified: (i) the award of the Qualified Teacher Status (Q); (ii) teachers on the main scale (Core) (C); (iii) teachers on the upper pay scale (Post Threshold Teachers) (P); (iv) Excellent Teachers (E); and (v) Advanced skills Teachers (A).

Criteria: At each stage, teaching professional standards encompass three domains. The first one refers to the teacher's *professional attributes*, including relationships with children and young people; attitude *vis-à-vis* the framework and the implementation of new school policies; communicating and working with others; and professional development activities. The second domain is composed of the teacher's *professional knowledge and understanding*, including knowledge on teaching and learning; understanding of assessing and monitoring; subjects and curriculum knowledge; literacy, numeracy and ICT skills; understanding the factors affecting the achievement of diversified student groups; and knowledge on student health and well-being. The last domain refers to the teacher's *professional skills*, including planning, teaching, assessing, monitoring, giving feedback competencies; ability to review and adapt teaching and learning; ability to create a learning environment; capacities to develop team working and collaboration. All of these standards are statements of good teaching which do not replace the professional duties and responsibilities of teachers.

Consequences on teacher professional growth and links to school expectations and policies: The standards support teachers in identifying their professional development needs. Where teachers wish to progress to the next career stage, the next level of the framework provides a reference point for all teachers when considering future development. Whilst not all teachers necessarily want to move to the next career stage, the standards also support teachers in identifying ways to broaden and deepen their expertise within their current career stages. These frameworks are a basis for professional responsibility and contractual engagement to engage all teachers in effective, sustained and relevant professional development throughout their careers. They provide a continuum of expectations about the level of engagement in professional development that provides clarity and appropriate differentiation for each career stage. They also set expectations about the contribution teachers make to others, taking account of their levels of skills, expertise and experience, their role within the school, and reflecting on their use of up-to-date subject knowledge and pedagogy. In all these cases, performance management is the key process that provides the context for regular discussions about teachers' career aspirations and their future development, within or beyond their current career stage.

For further information:

- Training and Development Agency for Schools (TDA):
  http://www.tda.gov.uk/teachers/professionalstandards.aspx and
  http://www.tda.gov.uk/teachers/continuingprofessionaldevelopment.aspx

- Office for Standards in Education (Ofsted): http://www.ofsted.gov.uk

3. Conciliating the summative and formative purposes in a comprehensive approach: Chile [Avalos and Assael, 2006]

Context: The historical context of the Chilean educational system has doubtlessly played a critical role in understanding the necessity for a comprehensive and conciliating teacher evaluation scheme. In 1980, the military government [1973-1990] transferred the management of schools to the municipal authorities, which also implied a change of status of teachers from public servants to salaries employees of

municipalities. At the end of the dictatorial regime, a major concern was that teachers' conditions did not evolve in line with those for public servants, which had an enormous impact on how teachers perceived and valued themselves, as well as on public opinion. In the 1990s the teaching profession suffered from a dramatic deterioration of the quality of applicants to teaching and from worsened working conditions. At the same time, evidence of unsatisfactory student learning results put a strong pressure on the government to include a clause in the new Teacher Statute (1991) that required a yearly evaluation of teachers. But while teachers continued to make their case for improved salaries and working conditions, they rejected the implementation of the evaluation system. This was followed by a long period of discussions and negotiations on the teacher evaluation model to be implemented.

Design and implementation of the system: The system was enacted by law in August 2004, that is, some seven years after the initial discussions. The system is directed toward the improvement of teaching and learning outcomes. It is designed to stimulate teachers to further their own improvement through the learning of their strengths and weaknesses. It is based on explicit criteria of what is evaluated, but without prescribing a model of teaching. It rests on the articulation of its different elements: criteria sanctioned by the teaching workforce, an independent management structure, especially prepared evaluators, and a coordinated set of procedures to gather the evidence required by the criteria.

Key actors in the system: The Centre for In-service Training located in the Ministry of Education (*Centro de Perfeccionamiento, Experimentación e Investigación Pedagógica*) manages the system. A consultative committee composed of academics and representatives from the Teachers' Union, the Chilean Association of Municipalities and the Ministry of Education, monitors and provides advice on the process. A university centre is contracted to implement the process: production and revision of instruments, selection and preparation of evaluators and scorers, and analysis of evidence gathered from each evaluation process. The application process itself is decentralised so that in every district there is a committee that is directly responsible for organising the evaluation procedures. The evidence gathered is processed at the district level and sent to the central processing unit at the university, together with contextual information that can help interpret results. This central form of processing the evidence follows a request by teachers with the purpose of greater objectiveness.

Criteria: The Ministry of Education took the lead in defining the assessment criteria, producing a set of standards based on the work done earlier for the initial teacher education standards and on Danielson's *Framework for Teaching*. The result is a framework for competent teaching formulated in four teaching domains (planning, learning environment, professionalism and teaching strategies for the learning of all students) and twenty criteria/standards. The framework was the subject of wide consultations among teachers until an agreement was reached. The criteria are linked to four levels of quality/performance: 'unsatisfactory', 'basic', 'competent' and 'excellent'.

Instruments: The evidence used to evaluate the teachers, structured around the Framework, includes four sources: (i) a portfolio with samples of teachers' work and a video of one of their lessons; (ii) a structured self-evaluation form; (iii) a structured interview with a peer evaluator; and (iv) a report from the school management and pedagogic authorities. The evaluation takes place every four years.

Training of evaluators: The peer evaluators are specifically prepared for their task and must pass a test to be accredited. Although they should be familiar with the context in which the evaluated teacher is based (*e.g.* socio-economic and working conditions) they may not be teachers in the same school.

Consequences of the evaluation: One of the main challenges that needed to be addressed during the negotiation process referred to the potential implications for the individual teacher evaluated. It was agreed that teachers rated as being at a 'basic' level are provided with specific professional development opportunities in order to improve. Teachers rated as performing 'unsatisfactorily' are also provided with

professional development opportunities, but are evaluated again one year later; if the teacher fails to perform satisfactorily in two consecutive evaluations, he or she is dismissed. By contrast, teachers assessed as 'competent' or 'exceptionally competent' are given priority in promotion opportunities and in professional development activities of their interest. They may also apply for a salary bonus provided that they take a test on curricular and pedagogical knowledge. The system has both summative and formative elements instead of being primarily dedicated to one of the purposes, which is the result of the negotiation process which had taken the multiple stakeholders' interests into account. For instance, the summative elements neither include a link between teacher's performance and student results (something the union strongly opposed) nor a link to the career ladder. The link to professional development is emphasised and differentiated on the basis of the teacher's level of performance.

For further information: Chile's laws on the Teaching Statute: Ley N°3.500; Ley N°19.070; Ley N° 19.933; Ley N° 19.961.

4. Teacher evaluation stemming from bureaucratic procedures: France [Haut Conseil de l'évaluation de l'école, 2003; Pochard, 2008]

Context: French teachers are classified in three distinct categories according to their education and initial certification: primary education teachers (*professeurs des écoles*), secondary education teachers with a regular certification (*enseignants certifiés*), and secondary education teachers with a higher level of certification (*enseignants agrégés*). All teachers are public servants but are placed in one of these three career tracks. These differ in terms of conditions and hours of work, administrative pay scale, and teaching practice (multitask primary education teachers *vs.* subject-specialised secondary education teachers). France does not generally suffer from teacher shortages and examinations to enter the profession continue to be selective. However, France has concerns regarding the societal status of teaching, and the skills necessary to respond to school needs. The current teacher evaluation system is often described as 'not very fair', 'not very efficient', and 'generating malaise and sometimes suffering' for both evaluated teachers and evaluators, because it is based on administrative procedures rather than a comprehensive scheme with a clear improvement purpose.

Periodicity of evaluation/evaluators: Teacher evaluation is supposed to be undertaken on a regular basis, as an integral part of the work and duties of the teacher. Primary education teachers are evaluated by a teaching inspector (*inspecteur*), while secondary level teachers are evaluated by a panel composed of an inspector – who defines 60 % of the final score – and the school principal – responsible for the other 40 %. However, the intended frequent evaluations often fall short of expectations. First, the frequency of evaluations is not legally fixed, and is arbitrarily determined by the inspectors' availability. This is a cause for concern regarding the fairness of the system – because teachers working under the same rules receive feedback at diverse intervals – as well as regarding its efficacy – the average interval between two evaluations being 3-4 years in primary education and 6-7 years in secondary education, deemed much too long. Moreover, the workload is such that concerns might be raised regarding the value of the feedback. An inspector takes responsibility for between 350 and 400 teachers, which is excessive for the feedback to be effective in improving teachers' practices. As a consequence, the inspectors themselves report malaise and frustration associated with the evaluation process, mainly because they feel that they have little impact on teaching practices and cannot develop their competences and skills for teaching enhancement. Their role is sometimes *de facto* restricted to control the abuses within the profession.

Instruments: Evidence on the teacher's practice is gathered through the observation of a teaching session, followed by an interview with the teacher. Criticisms of this approach include: (i) the fact that a single classroom observation might not be enough to forge a fair and accurate view of the teacher's abilities and knowledge; and (ii) in the interview teachers focus on reacting to the inspectors' criticisms instead of

discussing their particular needs for improvement. The whole procedure does not seem to give much room for self-evaluation and teachers' reflection on their own practice and performance.

Criteria: Both 'pedagogical' and 'administrative' aspects are observed and rated but with no reference to a framework which defines what 'good' teaching is. Concerns are numerous. The nature of the different 'pedagogical' skills assessed, as well as their weight in the overall appreciation of the teacher, remains largely at the discretion of each inspector. This reinforces subjective appraisals, unpredictable and random results, at the expense of fairness and accuracy in the process. Teachers report not knowing how and on what criteria they are evaluated. The most objective and understood criteria used to evaluate teachers are the 'administrative' ones such as punctuality and attendance. As a result, the rating obtained by a teacher often remains primarily determined by their certification rating (*i.e.* result of entrance examination).

Consequences: The consequences of the teacher's evaluation on the career are limited, except in cases of serious misconduct. Teachers' salaries are determined by a single salary scale in which progression depends on years of service and the initial qualifications and entrance examination. Commitment to work is rarely recognised and valued, as well as merit, outstanding performance, or initiatives seeking to improve student learning. In addition, there is no link to professional development activities, the latter being very limited and disconnected from teachers' identified weaknesses. The evaluation process does not provide opportunities for self-reflection on teaching practices or for peer mutual learning, and entails little advice and coaching.

For further information:

- *Haut Conseil de l'évaluation de l'école* (2003):
  http://cisad.adc.education.fr/hcee/ documents/rapport_annuel_2003.pdf

- Rapport des inspections générales:
  http://lesrapports.ladocumentationfrancaise.fr/BRP/054004446/0000.pdf

- Ministry of Education: http://www.education.gouv.fr/cid263/l-evaluation-des-personnels.html

About other teacher evaluation systems:

- The Canadian Province of Alberta:
  http://www.education.alberta.ca/department/policy/k12manual/section2/teacher.aspx

- The US State of Iowa: http://www.iowa.gov/educate/content/view/1450/1617

REFERENCES

Aaronson, D.; Barrow, L. and Sander, W. (2007) "Teachers and Student Achievement in the Chicago Public High Schools", *Journal of Labor Economics*, Vol. 25, No. 1, pp 95-135.

American Federation of Teachers (2001) "Beginning Teacher Induction: The Essential Bridge", Educational Issues Policy Brief No. 13, AFT, 2001.

American Federation of Teachers and National Education Association (2008) *A Guide to Understanding National Board Certification*, AFT and NEA, Washington, DC.

Anderson, L. and Pellicer, L. (2001) *Teacher Peer Assistance and Review*, Corwin Press.

Avalos, B. and Assael, J. (2006) "Moving from resistance to agreement: The case of the Chilean teacher performance evaluation", *International Journal of Educational Research*, Vol. 45, No. 4-5, pp 254-266.

Beck, R.; Livne, N. and Bear, S. (2005) "Teachers' self-assessment of the effects of formative and summative electronic portfolios on professional development", *European Journal of Teacher Education*, Vol. 28, No. 3, pp 221-244.

Bolino, M. and Turnley, W. (2003) "Counternormative impression management, likeability, and performance ratings: the use of intimidation in an organizational setting", *Journal of Organizational Behavior*, Vol. 24, No. 2, pp 237-250.

Bond, L.; Smith, T.; Baker, W. and Hattie, J. (2000) "The Certification System of the National Board for Professional Teaching Standards: A Construct and Consequential Validity Study", NBPTS, 2000.

Borman, G. and Kimball, S. (2005) "Teacher Quality and Educational Equality: Do Teachers with Higher Standards-Based Evaluation Ratings Close Student Achievement Gaps?", *The Elementary School Journal*, Vol. 106, No. 1, pp 3-20.

Braun, H. (2005) "Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models", Educational Testing Service (ETS), 2005.

Casson, H. Jr (2007) "Reducing Teacher Moral Hazard in the U.S. Elementary and Secondary Educational System through Merit-pay: An Application of the Principal – Agency Theory", *Forum for Social Economics*, Vol. 36, No. 2, pp 87-95.

Campbell, D.; Melenzyer, B.; Nettles, D. and Wyman, R. (2000) *Portfolio and performance assessment in teacher education*, Needham Heights, MA, Allyn & Bacon.

Cavalluzzo, L. (2004) "Is National Board Certification An Effective Signal of Teacher Quality?", The CNA Corporation, Alexandria, Virginia, 2004.

Center for Assessment and Evaluation of Student Learning (2004) "Using Student Tests to Measure Teacher Quality", CAESL Assessment Brief No. 9.

Center for Teaching Quality (2008) "Measuring What Matters: The Effects of National Board Certification on Advancing 21st Century Teaching and Learning", CTQ, 2008.

Cohen, C. and King Rice, J. (2005) "National Board Certification as Professional Development: Design and Cost", NBPTS, 2005.

Corcoran, T. (2007) "Teaching Matters: How State and Local Policymakers Can Improve the Quality of Teachers and Teaching", Consortium for Policy Research in Education (CPRE) Policy Briefs RB-48.

Danielson, C. (2001) "New Trends in Teacher Evaluation", *Educational Leadership*, Vol. 58, No. 5, pp 12-15.

Danielson, C. (1996, 2007) *Enhancing Professional Practice: A Framework for Teaching*, 1st and 2nd editions, Association for Supervision and Curriculum Development (ASCD), Alexandria, Virginia.

Danielson, C. and McGreal, T. (2000) *Teacher Evaluation to Enhance Professional Practice*, Association for Supervision and Curriculum Development (ASCD), Alexandria, Virginia.

Darling, L. (2001) "Portfolio as practice: the narratives of emerging teachers", *Teaching and Teacher Education*, Vol. 17, pp 107-121.

Darling-Hammond, L.; Pecheone, R. and Stansbury, K. (2004) "Beginning Teacher Quality: What Matters for Student Learning?", Research proposal from Standford University to the Carnegie Corporation of New York, available at www.pacttpa.org/_files/Publications_and_Presentations/ Carnegie_grant_proposal.doc.

Day, C. and Gu, Q. (2007) "Variations in the conditions for teachers' professional learning and development: sustaining commitment and effectiveness over a career", *Oxford Review of Education*, Vol. 33, No. 4, pp 423-443.

Department of Education, Science and Training (2007) "Performance-based rewards for teachers", DEST Research Papers, 2007.

Elmore, R. (2000) *Building a New Structure for School Leadership*, Albert Shanker Institute, Winter 2000.

Figlio, D. and Kenny, L. (2007), "Individual teacher incentives and student performance", *Journal of Public Economics*, Vol. 91, No. 5-6, pp 901-914.

Freund, M.; Kane Russell, V. and Kavulic, C. (2005) "A Study of the Role of Mentoring in Achieving Certification by the National Board for Professional Teaching Standards", NBPTS, 2005.

Goe, L. (2007) "The Link Between Teacher Quality and Student Outcomes: A Research Synthesis", National Comprehensive Center for Teacher Quality, 2007.

Goldhaber, D. and Anthony, E. (2007) "Can Teacher Quality Be Effectively Assessed? National Board Certification As a Signal of Effective Teaching", *The Review of Economics and Statistics*, Vol. 89, No. 1, pp 134-150.

Greenfield, W. (1995) "Toward a Theory of School Administration: the Centrality of Leadership", *Educational Administration Quraterly*, Vol. 31, No. 1, pp 61-85.

Halverson, R.; Kelley, C. and Kimball, S (2004) "Implementing Teacher Evaluation Systems: How Principals Make Sense of Complex Artifacts to Shape Local Instructional Practice" in *Educational Administration, Policy and Reform: Research and Measurement Research and Theory in Educational Administration*, Vol. 3, W.K. Hoy and C.G. Miskel (Eds.) Greenwish, CT: Information Age Press.

Hanushek, E. (2004) "Does School Accountability Lead to Improved Student Performance?", NBER Working Papers n°10591.

Hanushek, E. (1992) "The Trade-Off between Child Quantity and Quality", *Journal of Political Economy*, Vol. 100, No. 1, pp 84-117.

Hanushek, E. (1986) "The Economics of Schooling: Production and Efficiency in Public Schools", *Journal of Economic Literature*, Vol. 24, No. 3, pp 1141-1177.

Hanushek, E.; Kain, J.; O'Brien, D. and Rivkin, S. (2005) "The Market for Teacher Quality", NBER Working Papers n°11154.

Harris, D. and Sass, T. (2007) "The Effects of NBPTS-Certified Teachers on Student Achievement", Center for Analysis of Longitudinal Data in Education Research (CALDER), Working Paper No. 4.

Haut Conseil de l'évaluation de l'école (2003) *Rapport annuel*, HCéé, 2003.

Heneman, H. and Milanowski, A. (2003) "Continuing Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System", *Journal of Personnel evaluation in Education*, Vol. 17, No. 2, pp 173-195.

Heneman, H.; Milanowski, A. and Kimball, S. (2007) "Teacher Performance Pay: Synthesis of Plans, Research, and Guidelines for Practice", Consortium for Policy Research in Education (CPRE) Policy Briefs RB-46.

Heneman, H.; Milanowski, A.; Kimball, S. and Odden, A. (2006) "Standards-Based Teacher Evaluation as a Foundation for Knowledge- and Skill-Based Pay", Consortium for Policy Research in Education (CPRE) Policy Briefs RB-45.

Hess, F. and West, M. (2006) "A Better Bargain: Overhauling Teacher Collective Bargaining for the 21st Century", Cambridge, MA: Program on Education Policy and Governance, Harvard University.

Holland, P. (2005) "The Case for Expanding Standards for Teacher Evaluation to Include an Instructional Supervision Perspective", *Journal of Personnel Evaluation in Education*, Vol. 18, No. 1, pp 67-77.

Ingvarson, L.; Kleinhenz, E. and Wilkinson, J. (2007) *Research on Performance Pay for Teachers*, Australian Council for Educational Research (ACER), 2007.

Interstate New Teacher Assessment and Support Consortium (1992) "Model Standards for Beginning Teacher Licensing, Assessment and Development: A Resource for State Dialogue", INTASC, Council of Chief State School Officers (CCSSO), 1992.

Jacob, B. (2004) "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools", *Journal of Public Economics*, Vol. 89, No. 5-6, pp 761-796.

Jacob, B. and Lefgren, L. (2008) "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education", *Journal of Labor Economics*, Vol. 26, No. 1, pp 101-136.

Jacob, B. and Lefgren, L. (2005b) "What Do Parents Value in Education: an Empirical Investigation of Parents' Revealed Preferences for Teachers", NBER Working Paper n°11494.

Jacob, B. and Lefgren, L. (2005a) "Principals as Agents: Subjective Performance Measurement in Education", NBER Working Papers n°11463.

Jacob, B. and Levitt, S. (2003) "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating", *The Quaterly Journal of Economics*, Vol. 118, No. 3, pp 843-878.

Jacobs, C.; Martin, S. and Otieno, T. (2008) "A Science Lesson Plan Analysis Instrument for Formative and Summative Program Evaluation of a Teacher Education Program", *Science Teacher Education* (Articles online in advance of print).

Jun, M.-K.; Anthony, R.; Achrazoglu, J. and Coghill-Behrends, W. (2007) "Using ePortfolio for the Assessment and Professional Development of Newly Hired Teachers", *TechTrends*, Vol. 51, No. 4, pp 45-50.

Kane, T. and Staiger, D. (2002) "Volatility in School Test Scores: Implications for Test-Based Accountability Systems", *Brooking Papers on Education Policy*, Wahington, DC.

Kennedy, M. (2005) *Inside Teaching*, Harvard University Press, London, England, 2005.

Kimball, S. (2002) "Analysis of Feedback, Enabling Conditions and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems", *Journal of Personnel Evaluation in Education*, Vol. 16, No. 4, pp 241-268.

Kimball, S.; Milanowski, T. and McKinney, S. (2007) "Implementation of Standards-Based Principal Evaluation in One School District: First Year Results From Randomized Trial", Paper presented at the annual conference of the American Educational Research Association, available at http://cpre.wceruw.org/publications/KimballMilanowskiMcKinney.pdf.

Klecker, B. (2000) "Content validity of preservice teacher portfolios in a standards-based program", *Journal of Instructional Psychology*, Vol. 27, No. 1, pp 35-38.

Kleinhenz, E. and Ingvarson, L. (2004) "Teacher Evaluation Uncoupled: A Discussion of Teacher Evaluation Policies and Practices in Australian States and Their Relation to Quality Teaching and Learning", *Research Papers in Education*, Vol. 19, No. 1, pp 31-49.

Leigh, A. (2007) "Estimating Teacher Effectiveness From Two-Year Changes in Students' Test Scores", Research School of Social Sciences, Australian National University, available at http://econrsss.anu.edu.au/~aleigh/pdf/TQPanel.pdf.

Levin, J. (2003) "Relational incentive contracts", *American Economic Review*, Vol. 93, No. 3, pp 835–57.

Lustick, D. and Sykes, G. (2006) "National Board Certification as Professional Development: What are Teachers Learning?", *Education Policy Analysis Archives*, Vol. 14, No.5.

MacLeod, B. (2003) "Optimal contracting with subjective evaluation", *American Economic Review* 93, No. 1, pp 216-240.

Mansvelder-Longayroux, D.; Beijaard, D. and Verloop, N. (2007) "The portfolio as a tool for stimulating reflection by student teachers", *Teaching and Teacher Education*, Vol. 23, No. 1, pp 47-62.

Margo, J.; Benton, M.; Withers, K. and Sodha, S. (2008) *Those Who Can?*, Institute for Public Policy Research (IPPR) Publications, 2008.

Marshall, K. (2005) "It's Time to Rethink Teacher Supervision and Evaluation", *Phi Delta Kappan*, Vol. 86, No. 10, pp 727-735.

McColskey, W. and Stronge, J. (2005) "A Comparison of National Board Certified Teachers and non-National Board Certified Teachers: Is there a difference in teacher effectiveness and student achievement", NBPTS, 2005.

Milanowski, A. (2007) "Performance Pay System Preferences of Students Preparing to Be Teachers", American Education Finance Association, 2007.

Milanowski, A. (2004) "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence From Cincinatti", *Peabody Journal of Education*, Vol. 79, No. 4, pp 33-53.

Milanowski, A. and Heneman, H. (2001) "Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System: A Pilot Study", *Journal of Personnel Evaluation in Education*, Vol. 15, No. 3, pp 193-212.

Milanowski, A. and Kimball, S. (2003) "The Framework-Based Teacher Performance Assessment Systems in Cincinatti and Washoe", CPRE Working Paper Series TC-03-07.

Ministerial Council on Education, Employment Training and Youth Affairs (2003) "A National Framework for Professional Standards for Teaching", MCEETYA, Carlton South, Australia, 2003.

Mizala, A. and Romaguera, P. (2004) "School and teacher performance incentives: The Latin American experience", *International Journal of Educational Development*, Vol. 24, No. 6, pp 739-754.

Muñoz, M. and Chang, F. (2007) "The Elusive Relationship Between Teacher Characteristics and Student Achievement Growth: A Longitudinal Multilevel Model for Change", *Journal of Personnel Evaluation in Education*, Vol. 20, No. 3-4, pp 147-164.

Nabors Oláh, L.; Lawrence, N. and Riggan, M. (2008) "Learning to learn from benchmark assessment data: How teachers analyze results", Paper presented at the Annual Meeting of the American Educational Research Association, New York, 2008, available at http://www.cpre.org/images/stories/cpre_pdfs/ aera2008_olah_lawrence_riggan.pdf .

National Board for Professional Teaching Standards (2007) "A Research Guide on National Board Certification of Teachers", NBPTS, Arlington, VA, 2007.

Odden, A. and Kelley, C. (2002) *Paying Teachers for What They Know and Do: New and Smarter Compensation Strategies to Improve Schools*, Corwin Press, Thousand Oaks, California, 2002.

Office for Standards in Education (2006) "The logical chain: continuing professional development in effective schools", OFSTED Publications n°2639, United Kingdom, 2006.

Organisation for Economic Co-Operation and Development (2008) *Improving School Leadership*, OECD, Paris, 2008.

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

Ovando, M. and Ramirez, A Jr (2007) "Principals' Instructional Leadership Within a Teacher Performance Appraisal System: Enhancing Students' Academic Success", *Journal of Personnel Evaluation in Education*, Vol. 20, No. 1-2, pp 85-110.

Pecheone, R. and Chung, R. (2006) "Evidence in Teacher Education: The Performance Assessment for California Teachers (PACT)", *Journal of Teacher Education*, Vol. 57, No. 1, pp 22-36.

Peterson, K. (2000) *Teacher Evaluation: A Comprehensive Guide to New Directions and Practices*, 2nd edition, Thousand Oaks, CA: Corwin Press.

Peterson, K.; Wahlquist, C. and Bone, K. (2000) "Student Surveys for Teacher Evaluation", *Journal of Personnel Evaluation in Education*, Vol. 14, No. 2, pp 135-153.

Peterson, K.; Wahlquist, C.; Esparza Brown, J. and Mukhopadhyay, S. (2003) "Parents Surveys for Teacher Evaluation", *Journal of Personnel Evaluation in Education*, Vol. 17, No. 4, pp 317-330.

Peterson, K. (2006) "Teacher Pay Reform Challenges States", Stateline.org: where policy and politics news click, available at http://www.stateline.org/live/ViewPage.action?siteNodeId=136&languageId=1&contentId=93346.

Petty, T. (2002) "Identifying the Wants and Needs of North Carolina High School Mathematics Teachers for Job Success and Satisfaction", NBPTS, 2002.

Ping Yan Chow, A.; King Por Wong, E.; Seeshing Yeung, A. and Wan Mo, K (2002) "Teachers' Perceptions of Appraiser-Appraisee Relationships", *Journal of Personnel Evaluation in Education*, Vol. 16, No. 2, pp 85-101.

Pochard, M. (2008) *Livre vert sur l'évolution du métier d'enseignant*, Rapport au ministre de l'Education nationale, La Documentation française, Collection des rapports officiels, 2008.

Popham, J. (1997) "Consequential validity: Right Concern – Wrong Concept", *Educational Measurement: Issues and Practice*, Vol. 16, No. 2, pp 9-13.

Robinson, V. (2007) "School Leadership and Student Outcomes : Identifying What works and Why", Australian Council for Educational Leaders, ACEL Monograph Series No. 41.

Sanders, W.; Ashton, J. and Wright, P. (2005) "Comparison of the Effects of NBPTS Certified Teachers with Other Teachers on the Rate of Student Academic Progress", NBPTS, 2005.

Smith, T.; Gordon, B.; Colby, S. and Wang, J. (2005) "An Examination of the Relationship Between Depth of Student Learning and National Board Certification Status", Office for Research on Teaching, Appalachian State University.

Stronge, J. and Tucker, P. (2003) *Handbook on Teacher Evaluation: Assessing and Improving Performance*, Eye On Education Publications, 2003.

Stronge, J.; Ward, T.; Tucker, P. and Hindman, J. (2007) "What is the Relationship Between Teacher Quality and Student Achievement? An Exploratory Study", *Journal of Personnel Evaluation in Education*, Vol. 20, No. 3-4, pp 165-184.

Strudler, N. and Wertzel, K. (2008) "Costs and Benefits of Electronic Portfiolos in Teacher Education: Faculty Perspectives", *Journal of Computing in Teacher Education*, Vol. 24, No. 4, pp 135-142.

Training and Development Agency for Schools (2007a) "Models Performance Management Policy for Schools", TDA, United Kingdom, 2007.

Training and Development Agency for Schools (2007b) "Professional Standards for Teachers: Why Sit Still in Your Career?", TDA, United Kingdom, 2007.

Tucker, P.; Stronge, J. and Gareis, C. (2002) *Handbook on teacher portfolios for evaluation and professional development*, Larchmont, NY, Eye on Education.

UNESCO (2007) *Evaluación del Desempeño y Carrera Profesional Docente: Una panorámica de América y Europa*, Oficina Regional de Educación para américa Latina y el Caribe, UNESCO Santiago, 2007.

Vandervoort, L.; Amrein-Beardsley, A. and Berliner, D. (2004) "National Board Certified Teachers and Their Students' Achievement", *Education Policy Analysis Archives*, Vol. 12, No. 46.

Weingarten, R. (2007) "Using Student Test Scores to Evaluate Teachers: Common Sense or Nonsense?", United Federation of Teachers, available at http://www.uft.org/news/randi/ny_times/UFT_WMM_Mar07_v32.pdf.

Wertzel, K. and Strudler, N. (2006) "Costs and Benefits of Electronic Portfiolos in Teacher Education: Student Voices", *Journal of Computing in Teacher Education*, Vol. 22, No. 3, pp 69-78.

Xin, T.; Xu, Z. and Tatsuoka, K. (2004) "Linkage Between Teacher Quality, Student Achievement, and Cognitive Skills: A Rule-Space Model", *Studies in Educational Evaluation*, Vol. 30, pp 205-223.

EXISTING OECD EDUCATION WORKING PAPERS

No. 1   *Teacher Demand and Supply: Improving Teaching Quality and Addressing Teacher Shortages* (2002), Paulo Santiago.

No. 2   *Teacher Education and the Teaching Career in an Era of Lifelong Learning* (2002), John Coolahan.

No. 3   *Towards an Understanding of the Mechanisms That Link Qualifications and Lifelong Learning* (2003), Friederike Behringer and Mike Coles.

No. 4   *Measuring Educational Productivity in Standards-Based Accountability Systems: Introducing the SES Return on Spending Index* (2005), Martin Hampel.

No. 5   *PISA 2000: Sample Weight Problems in Austria* (2006), Erich Neuwirth.

No. 6   *Funding Systems and their Effects on Higher Education Systems – International Report* (2007), Franz Strehl, Sabine Reisinger and Michael Kalatschan.

No. 7   *On the Edge: Securing a Sustainable Future for Higher Education* (2007), OECD/IMHE-HEFCE.

No. 8   *Globalisation and Higher Education* (2007), Simon Margison and Marijk van der Wende.

No. 9   *Understanding the Regional Contriubution of Higher Education Institutions: A Literature Review* (2007), Peter Arbo and Paul Benneworth.

No. 10   *Effects of Tertiary Expansion – Crowding-out Effects and Labour Market Matches for the Higher Educated* (2007), Bo Hansson.

No. 11   *Skilled Voices? Reflections on Political Participation and Education in Austria* (2007), Florian Walter and Sieglinde K. Rosenberger.

No. 12   *Education and Civic Engagement: Review of Research and a Study on Norwegian Youths* (2007), Jon Lauglo and Tormod Oia.

No. 13   *School Accountability, Autonomy, Choice, and the Level of Student Achievement: International Evidence from PISA 2003* (2007), Ludger Wössmann, Elke Lüdemann, Gabriela Schütz and Martin R. West.

No. 14   *School Accountability, Autonomy, Choice, and the Equity of Student Achievement: International Evidence from PISA 2003* (2007), Gabriela Schütz, Martin R. West, Ludger Wössmann.

No. 15   *Assessment of learning outcomes in higher education: a comparative review of selected practices* (2008), Deborah Nusche.

No. 16   *Approaches and Challenges to Capital Funding for Educational Facilities* (2008), Ann Gorey.

No. 17   *Recent Developments in Intellectual Capital Reporting and their Policy Implications* (2008), W. Richard Frederick.

No. 18   *Employers' Perspectives on the Roles of Human Capital Development and Management in Creating Value* (2008), L. Bassi and D. McMurrer.

No. 19   *Job-related Training and Benefits for Individuals: A Review of evidence and explanations* (2008), Bo Hansson.

No. 20   *A Framework for Monitoring Transition Systems* (2008), Rolf van der Velden.

No. 21   *Final Report of the Development of an International Adult Learning Module (OECD AL Module)* (2008), Bo Hansson and Helmut Kuwan.

No. 22   *What Works in Migrant Education? A Review of Evidence and Policy Options* (2009), Deborah Nusche.

THE OECD EDUCATION WORKING PAPERS SERIES ON LINE

The OECD Education Working Papers Series may be found at:

The OECD Directorate for Education website: www.oecd.org/edu/workingpapers

The OECD's online library, SourceOECD: www.sourceoecd.org

The Research Papers in Economics (RePEc) website: www.repec.org

If you wish to be informed about the release of new OECD Education working papers, please:

Go to www.oecd.org

Click on "My OECD"

Sign up and create an account with "My OECD"

Select "Education" as one of your favourite themes

Choose "OECD Education Working Papers" as one of the newsletters you would like to receive

For further information on the OECD Education Working Papers Series, please write to:
edu.contact@oecd.org.