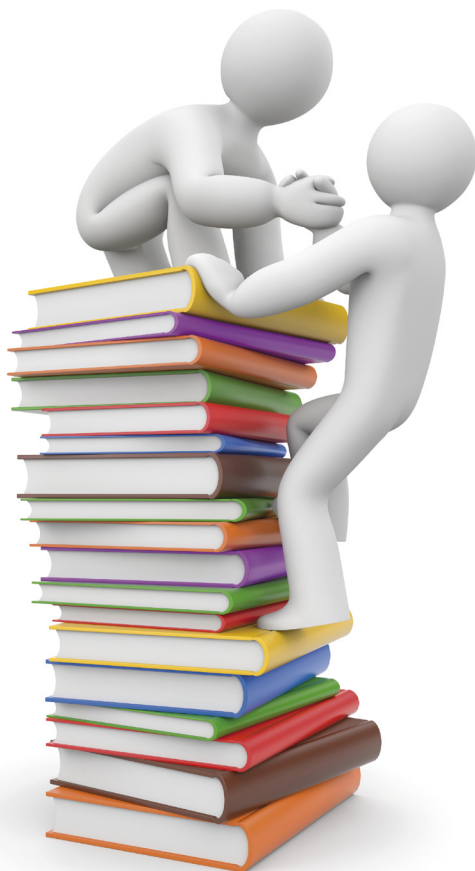


# VALORIZZA

A HOLISTIC AND CONTEXTUAL APPROACH  
FOR TEACHER APPRAISAL



«The Italian Valorizza experiment is an extremely interesting new approach and a major contribution to the international debate on teachers evaluation...

It is to be hoped that its results will be amply spread throughout the international educational community.»

*(Dirk Van Damme, head of CERI – OECD)*

# VALORIZZA

A HOLISTIC AND CONTEXTUAL APPROACH  
FOR TEACHER APPRAISAL

"How to identify and rewards teachers who are highly  
and widely respected for their professional ability  
within the context of each school"

**The content of this booklet can be downloaded on**

<http://www.treelle.org/valorizza-experiment>

[http://www.fondazionescuola.it/magnoliaPublic/iniziative/ricerca\\_valorizza/Teacher-Summit.html](http://www.fondazionescuola.it/magnoliaPublic/iniziative/ricerca_valorizza/Teacher-Summit.html)

**Contact**

For further information please contact:

[info@treelle.org](mailto:info@treelle.org)

FONDAZIONE PER LA SCUOLA  
DELLA COMPAGNIA DI SAN PAOLO  
Piazza Bernini, 5  
10138 Torino - Italy  
Tel. +39 011 4306511  
Fax +39 011 4333135  
[info@fondazionescuola.it](mailto:info@fondazionescuola.it)

ASSOCIAZIONE TREELLE  
Palazzo Pallavicino  
Via Interiano, 1  
16124 Genova - Italy  
Tel. + 39 010 582 221  
Fax + 39 010 540 167  
[info@treelle.org](mailto:info@treelle.org)

# INDEX

<b>Summary</b> .....	5
----------------------	---

<b>Part One</b> .....	7
-----------------------	---

## **Is it possible to evaluate teachers?**

1. Teacher evaluation .....	7
2. The evaluation of the “ethical professions” .....	8
3. The Italian situation .....	9

<b>Part Two</b> .....	13
-----------------------	----

## **Valorizza: an experiment by the Ministry of Education**

4. Key concepts and hypothesis .....	13
5. The Valorizza experiment .....	14

<b>Part Three</b> .....	19
-------------------------	----

## **Analysis of the results**

6. Analysis of the results .....	19
7. Valorizza: criticism and response .....	32
8. Valorizza - Strengths and weaknesses .....	38
<i>(by Dirk van Damme – Head of CERI-OECD)</i>	

<b>Part Four</b> .....	41
------------------------	----

## **Can the model be systematically deployed throughout the educational system?**

9. A gradual approach .....	41
10. Is the proposal coherent with the strategic objectives connected of teacher evaluation? .....	43



## Summary

### ***How to identify and reward teachers who are highly and widely respected for their professional ability within the context of each school***

The Italian Ministry of Education launched an experiment in 2011 to base the appraisal of merit-worthy teachers on the experience and views of the key stakeholders in the school: principals, teachers, parents and students. The basic assumption was that “within each school everyone knows who the best teachers are, by reputation and in that specific context”.

Thirty three schools were voluntarily admitted to join this experiment and, within each school, teachers were invited to volunteer for evaluation. Each candidate teacher completed a self-evaluation questionnaire, which focused on their professional behaviour, attitudes and practices and provided a CV containing information about their professional background. All the parents of the school and the students of the last two years of high school were also given an evaluation questionnaire to fill in, in which they were asked to indicate up to three teachers, whom they considered as the most highly and widely respected for their professional behaviour. The response rate was quite high at around 60%.

The teachers assembly elected two of its members who joined the school head to form an evaluation committee (nucleus). Each of the three evaluators carefully reviewed the questionnaires and the professional background of the candidates and subsequently the evaluation forms completed by parents and students. The evaluators had to bear in mind, as a guideline for their appraisal, the professional profile of the teachers, broadly defined in their national labour contract. No single element of this evaluation process was given priority over others, nor a specific weight. The analysis was carried out individually, without prior agreement and without the exchange of views among the three members during the evaluation process. In conclusion each of the evaluators drew up a personal list of those teachers considered most highly and widely respected for their professional behaviour (up to 30% of the candidates in each school). Finally, the three evaluators gathered for the first time and compared their lists: the teachers who had been chosen by all the three evaluators were selected. Subsequently, candidates who were included in two lists out of three were examined and the “comparatively better” were added to the general list, with the aim that the total would not exceed 30% of the candidates in the school. More than 900 teachers were evaluated in the 33 participant schools and 276 of them were selected. The Ministry rewarded them with a bonus equivalent to a month’s salary.

It is interesting to note that two thirds of the selected candidates were unanimously, yet independently, chosen by the three evaluators and that nearly all, in the remaining third, were chosen by two of the evaluators. It is also worth noting that those teachers who were most appreciated by the parents and the students were also, by a wide majority, among those selected by the evaluating committee.

These outcomes suggest that the key stakeholders in a school share very similar views as to whom the best teachers are, even if those stakeholders may not be able to define or agree upon the exact weight of the criteria used to evaluate professional behaviour.

The Ministry asked two respected and independent Foundations (Associazione Treille and Fondazione per la Scuola della Compagnia di San Paolo) to produce a scientific report about the extent to which the outcomes from the evaluating committee were considered appropriate by the other stakeholders in each school. A new questionnaire was distributed to all the teachers (both candidates and non) and to the parents and the students of each school in the sample. Once again, the response rate was around 50% and two thirds of the respondents fully agreed with the results.

It is noteworthy that all the schools accepted the experiment favourably. What was mostly appreciated was the recognition provided to the best teachers for their professional behaviour and the display of trust in the capacity of school communities for appraising their own members, without having to rely on external experts. The high level of the consensus on the selected candidates also helped to avoid conflicts and disputes about the evaluation process and its outcomes.

The high level of convergence in the perspective of the different stakeholders, each one giving his point of view separately from others, also suggests a high level of validity for the process and that a good reputation is a shared opinion even though it may be based on a different appreciation of the criteria or motivations. It seems that the Valorizza model cannot be qualified as too “subjective” or “arbitrary”: it is rather an “inter-subjective” holistic and contextual approach, especially well suited to small communities, such as schools. It is reliable and effective, since it identifies and rewards the teachers who are widely recognized as the best in their school. In addition it is cheap, quick and non-bureaucratic.

Finally, the two independent Foundations provided three suggestions, should the Valorizza method become in the future the standard means for identifying and rewarding merit worthy teachers:

- that the selected teachers receive a yearly allowance equal to two months of their ordinary salary for three consecutive years;
- that the appraisal process be repeated in each school every three years, always on a voluntary basis, in order to extend the acknowledgment to a wider number of teachers;
- that the process be diffused progressively, following a bottom-up approach: i.e. that the schools may decide to take part or not in the process, within an annual budget provided by the Ministry specifically for the development of the Valorizza model in the school system. (See chapter 9).

### **NOTE for the reader**

*The parts to be read more attentively for a thorough comprehension of the “idea” underlying the Valorizza process are chapters 4, 5, and 8.*

*Chapter 6 contains a selection of the data from the validation process and the related analysis. The key tables are 9, 10 and 11 on page 23 and page 24 (“Level of convergence in the decision of the three evaluators school by school).*

# Part One

## Is it possible to evaluate teachers?

### 1. Teacher evaluation

The first hurdle in the process is well known: how can substantially self regulated performance be evaluated? Excluding that it can be measured in terms of mere fulfilment, almost all methods imagined so far (and also put into practice in different countries) rely upon two methodologies both of which are based upon indirect measurement.

#### *a. On indirect “objective” criteria*

This approach is based on the quest for an as complete and analytical as possible definition of professional practice, a form of a break down into its composite parts. For example how lesson time is organised, what criteria are used to measure learning, what qualifications have been attained. The next step is to attribute a value to each element so identified and develop appropriate identification and measurement techniques.

Another way to express the same concept refers to the so-called “objectivity” of the evaluation. That is the attempt to identify indicators of quality in the individual’s behavior or professional activities (time dedicated to continuing professional development, specialization courses attended and so forth) and to infer an overall evaluation, from their presence or absence, of the professionalism demonstrated.

These approaches have not yet produced satisfactory results mainly because it is not easy to “isolate” individual elements that are themselves significant for the evaluation, regardless of relationship and context. For example, no one doubts that a deep knowledge of the subject is an important requisite for a teacher, but in practice this attribute can have different relevancies depending on whether the student is socially and culturally predisposed to study or is demotivated and at risk of misconduct.

In short, a good teacher is not an abstract theoretical concept, but is a good teacher in practice. The worth of a teacher is not easily broken down into elementary characteristics that are good for every situation, the sum of which can provide a uniform and “objective” evaluation that is independent of the evaluator and the context in which the evaluation takes place.

#### *b. On the results obtained by students.*

A different approach to the subject of the evaluation of teachers - developed since external systematic surveys of learning by students became available – consists in starting from the result obtained in learning outcomes to identify the value of the teacher professional behaviour.

This inference appears natural and also seductive: if the student has learnt, the teacher has been good. And, apparently, to measure what students have learned in school is fairly “easy”, although the relationship between theoretical knowledge and the skills which will later be demonstrated in practice still needs to be understood. But, when one tries to use these data to assess the professional quality of teachers, one is immediately confronted with further questions: to what extent are the overall results obtained



by the student attributable to the individual teacher and to what extent to the team to which the teacher belongs or to the context in which the school operates? And, last but not least, to what extent do the results depend on the individual student and their commitment to study, over and above their intellectual capability? Important debates are ongoing regarding this, but almost everyone agrees that the results of students learning can be only one of many aspects to consider, certainly not the only one.

In any case, such an approach - in addition to providing results that are as yet uncertain and debatable - is time consuming and absorbs important resources. "Longitudinal", i.e. repeated over time, surveys are required; measurement of the socio-cultural parameters that surround the school; complex operations to eliminate various "distorting" variables so as to isolate and identify the so-called "added value." Which is then (with some approximation) attributable to the activity of the school, but can almost never be convincingly attributed to the contribution of the individual teacher.

In substance, there are no inherently objective and purely quantitative methods for assessing the performance of teachers. The reason is quite simple: evaluation is inherently subjective because it is the expression of an opinion of the evaluator on the worth of the subject of evaluation, in this case the performance of a teacher.

There is a further difficulty, which is not restricted to the teaching profession and which also needs to be considered when dealing with these issues, and will be addressed in the next section.

## 2. The evaluation of the "ethical professions"

There are professions, such as that of the judge or the medical doctor, upon whose decisions the freedom or even the lives of individuals may depend. Such is the profession of the teacher, in a less dramatic though not entirely different way, to whom it is entrusted - along with the transmission of knowledge - the duty of care to accompany, guide and correct the personal and civic development of the student.

For these professions the parameters of evaluation applied to other disciplines are not valid: the importance of the subjects under their care and the impact that they have upon them is of such significance that they must have on one hand absolute decision-making freedom, and on the other operate within a particularly rigid ethical framework.

These professions have always been granted the right to complete self determination in their professional decisions: in the interest of those who rely upon or are entrusted to them. A judge who is not free to form an opinion, or a doctor who cannot select in their opinion the most appropriate therapy would constitute a grave risk to society. At the same time the limits to that freedom need to be defined, the red line that separates that freedom from the arbitrary and the irresponsible.

The traditional formula dictates: "according to my ability and judgment". I.e. on the basis of accepted standards and recognised best practice, but answering only to one's personal conviction.

The case of a teacher is similar even though not identical. Learning is a process that takes place in the mind of the student but cannot occur without the trigger that is con-

stituted by the actions of the teacher. That action must, by definition, be flexible as each student is different from the next. An absolute and complete didactic method cannot be prescribed: and therefore it is not possible to evaluate the quality of a teacher with respect to a code of professional conduct or the scrupulous application of the best pedagogic theory.

This much is true that the same teacher, in the same class, with the same process obtains very different results from one student to the next, which makes it impossible to evaluate based only upon the outcome of the learning process.

### 3. The Italian situation

In Italy the only aspect of the educational system that is regularly evaluated is that of the results obtained by students, based on national tests. Of course, Italy takes part in some of the most known international surveys on students' learning, such as these promoted by OECD. There is a national institute for the evaluation of the educational system (INVALSI - Istituto nazionale per la valutazione del sistema di istruzione) that has since 2009 published quantitative and (to a lesser extent), qualitative data for all Italian students at each key stage of their academic progression: the second, fourth, sixth, eighth and tenth years of study. The results cover only two subjects, Italian and Mathematics. In reality – and whilst waiting for the creation of a broader and more historic data base – this is a simple measurement rather than a true and proper evaluation.

No evaluation of the system in its entirety exists, nor of the individual schools, which have somewhat limited autonomy and are not able to rely upon the support of a specialised inspecting body. As for the school heads, as a result of union pressure, the system tends to ascribe to them a mainly bureaucratic function.

That said, a governmental provision is on the point of being implemented that will subject all schools and their heads to evaluation, without there being at the moment the intention to include the evaluation of the individual teachers, due to strong resistance from the unions. This situation produces increasing difficulties, given that the teachers constitute by far the single most important variable in the quality of the learning process.

In short, the principal problems given a lack of teacher evaluation are:

- It is not possible to structure an economic progression path for teachers that is not based, as now, upon years of service.
- It is not possible to clearly and systematically identify those teachers with particular personal capabilities in addition to teaching skills, to whom coordination, organisational, training, coaching and similar functions could be ascribed, the so called *middle management*.
- It is not possible to motivate the best to continuously improve themselves, nor elicit emulation among those that could do better, as in any event there will be no recognition, not even symbolic or in status.

It is not possible to effectively deal with critical cases of teachers who are not suited to the profession, which, though few in number, exist in any system and are able to cause significant damage.

What is really essential is to overcome the taboo of a fake professional uniformity of so many people. It is true that all teachers do the same job: it is not true (and it is neither fair nor intellectually honest to pretend) that all do their job in the same way and with the same results. Many do well, many more could do better.

### *The opinion of the teachers*

In this situation, the most attentive observers are in agreement that an acceptable means to overcome these weaknesses needs to be found. What is new is that for the first time a significant number of teachers themselves are becoming convinced of this, as has been demonstrated by different research studies.

Among these it is worth highlighting the Third Survey of Teachers (2010), performed by the IARD Institute and the Ricerca ANP-NOMISMA (2009) carried out on a vast and representative sample. Both surveys revealed that a significant percentage of teachers considered the need for an evaluation of professional performance to be well overdue, even if they were divided in their opinion about the means and content of the form such an evaluation should take.

In all cases there was a distinct preference for an internal evaluation carried out by individuals with direct experience of what is going on within the school. It is interesting to note that an increasing number of teachers are ready to accept differentiations in remuneration and career paths based on evaluation, and almost two out of three teachers are in favour of a system of merit recognition.

We are dealing with an attitude that is rapidly evolving: just ten years before, the preceding IARD survey showed a very different reality whereas today an increasing number of Italian teachers, at least half of them, are willing to play the game.

The understanding that which is never valued ends up losing value has developed: in the eyes of the public but also in the eyes of the protagonists. An entire body of professionals, denied the possibility of measuring themselves against some kind of reference, has lost social visibility and has come up against a worrying decline in collective self esteem.

The time has come to move from the recognition of the need of an evaluation to the study of the most useful and practical means of obtaining one. Along this path, beyond union resistance, there are other obstacles:

- The lack of a clearly defined professional profile given the high level of professional self determination traditionally accorded to teachers.
- The large number of teachers (over seven hundred thousand)
- The limited availability of financial resources that demand simple and economic solutions.

However perhaps the greatest difficulty is conceptual in nature: whether it is and how

is it possible to define a “good teacher”. This is a traditionally controversial subject about which different theories have clashed without as yet finding common ground. The Valorizza experiment was proposed to test an original approach to the question.

*The turning point: a new law on the appraisal of all the civil servants*

In 2009 a law was passed in parliament that set out new rules for the evaluation of all public servants. There is however a waiver: because of the difficulties that have been described regarding schools, and in particular for teachers, their evaluation was postponed until a future decree that would adapt the general principles to the characteristics of instructional services.

In this context, the Ministry decided to form a Technical Scientific Committee (TSC), included in which were a few international experts, with the mandate of formulating a proposal for a National system for the evaluation of the schools service aimed at “enhancing the value of merit.”

The individual positions within the committee were polarised around two approaches: those that wanted to define a model of evaluation for the individual teacher and those who wanted to direct attention at the effectiveness of the educational teams and of the schools. It was decided to pursue both approaches: the project directed at the evaluation of individual teachers took the name **“Valorizza”** [*Enhancing Value*] and constitutes the subject of this publication.

This was an “experimental” project in the sense that the TSC advised the Ministry to test the model on a small scale in order to highlight its potential and limits. This suggestion coincided with the intention of the Ministry which wanted – before acting on a wide scale – to fine tune and test the instrument on a reduced dimension which would be more easily monitored.

In the following section we will illustrate in further detail the assumptions and the working model that the project was based upon.



## Part Two

### Valorizza: an experiment by the Ministry of Education

#### 4. Key concepts and hypothesis

##### *A “holistic” and contextual evaluation*

The internal debate of the Committee produced an original methodological hypothesis, different from those previously mentioned: to adopt a “holistic” approach to teacher performance rather than an analytical procedure.

The key concept was taken from an original idea by the TreelLe Association which had been developed in its publication “Which teachers for the school of autonomy?” - 2004. From this the TSC took the fundamental concept rather than the operational aspects.

The hypothesis formulated was based on certain fundamental considerations:

- that professional teaching is based on the possession and use of certain basic competences: *subject matter; pedagogical and psychological competencies; teaching methodology; organisational – relational; research – documentation and evaluation*. That said, the relative proportion of these competencies varies and it is not possible to assign a “weight” to each in the evaluation if one isolates one from the other;
- teaching expertise is to a certain extent “situational” i.e. it has value and can be the object of appreciation in a given context;
- in schools everyone knows by reputation who the most highly esteemed teachers are. It is a matter of common experience that favorable judgments are based on different considerations depending on point of view, but tend to centre on certain individuals. In short the reasons why a teacher is appreciated by their principal, their colleagues, by auxiliary staff, by parents or students change, but those appreciated remain the same. It is a matter of finding a transparent procedure for bringing out this consensus.

On these premises the conviction is derived that it is to no account to identify and separately evaluate specific professional competencies: no ideal recipe exists, valid for all situations. The “quality” of a teacher – taken also as their capacity to act in a given context – is appreciated in its entirety and above all within that given context. It is neither necessary nor possible to compare qualities observed in different contexts. Different users and working environments characterised by particular problems call for different approaches. At the extreme, one could say that one “unique” style, an “ideal-type” of teacher, designed without having reference to any concrete situation, could not even be appraised, because you do not know his suitability with respect to any particular school.

It follows that the evaluation of teachers makes sense within a single school: it would otherwise be unthinkable to try and construct a ranking of over 700,000 teachers spread across 40,000 educational locations.

These considerations inspired the creation of an evaluation model that, in each school, brought together many different “subjective” contributions to build up, as far as possible, a consistent and shared evaluation. What makes it a reliable judgment is not the point of view of each single evaluator, but the spontaneous convergence of multiple positive appreciations. In short, the fact that many independent judgments agree ran-

domly on the same teachers just elide the original randomness and also eliminates what could be arbitrary in each of them taken alone. It is all about finding a way to express these multiple points of view independently of each other and without giving them a chance of reciprocal influence and of passing errors and prejudices among them.

Naturally, a completely new model needed to be tested to see if it worked and thus the decision for a trial on a small scale. An experiment is not valid unless its results are verified, and for this reason the Committee recommended a subsequent process of verification (see chapter 6)

## 5. The Valorizza experiment.

### *a. The strategic objectives connected of teacher evaluation.*

The ultimate aim of the evaluation is to improve the learning outcomes of students. In order to obtain this goal, the quality of teaching must be improved, that is to say teacher competency. In addition, the new law of 2009 (although not fully in effect yet) requires the evaluation of individual teachers and the school.

However, there are also other reasons for doing so: reasons that regard the improvement of the school system and are aimed at the essential aspects of this. These are:

1. to link economic rewards to a mechanism of recognition of merit and not only to years of service;
2. to set in motion a dynamics of positive emulation among teachers, that broadens the range of professional excellence;
3. to identify the most esteemed personalities in each school for the purpose of the assignment of additional tasks, both didactic and organizational. In this way the development of the so called middle management could be promoted, which is essential for shared leadership. In future, head teachers and inspectors could possibly be recruited among middle managers;
4. to attract over time to the teaching profession high level graduates through the provision of higher remuneration and career prospects;
5. to encourage all teachers to develop the practice of self evaluation, a prerequisite for the general improvement of their performance;

### *b. The specific objectives of Valorizza*

It is important to keep in mind that Valorizza – because of how it is constructed and its theoretical assumptions – is not an instrument conceived for the evaluation of *all teachers*. Its objective is more limited, but no less important for this: to identify and reward the most *merit worthy* teachers.

The concept of *merit worthy* is not absolute, referring to an abstract best teacher, but relative to a given specific context (a single school), and identifies those that are considered most capable by their immediate community (principals, colleagues, parents and students).

This brings us back to the previous consideration of the difficulty of identifying valid and objective universal parameters for the evaluation of a self-regulated profession,

self regulated because of its “ethical” nature (see chapter 2).

If many agree in appreciating the way in which a specific teacher carries out their professional role, and thus the way in which they live the ethics of the profession, it is probable that that teacher is effectively worthy of merit with respect to that community. It is in any event the community recognition of worth, and the Ministerial reward that sends a positive reinforcing signal.

### **c. Expected benefits of the Valorizza model**

In its work, the TSC (in accordance with the Ministry) wanted to ensure certain aspects that were especially important for the fulfilment of the mandate. The proposed model should:

- be able to be developed in a short time;
- be easily managed without heavy bureaucratic complications;
- be economic (i.e not require great resources to work);
- have definable costs (with pre-established rewards to be attributed);
- be reliable (i.e produce results that have general consensus within the scholastic community);
- assure the active participation of the candidates in the process regarding them;
- not be open to legal contention and attack;
- be open to development and modification overtime.

### **d. Valorizza: description of the model**

These are the principal elements of the model that was tested:

- *the “reputation” model*: i.e. based upon the general evaluation by the different components of the scholastic community, which were not asked to provide an analytical choice but only to indicate the most esteemed teachers. This method arose from the adoption of a “holistic” approach, i.e. an overall view of the individual teacher and their professional behaviour without – for reasons explained – analysing and giving weight to the single components of their professional performance;
- *participants involved in the process*: In each school the choice of “merit-worthy” teachers (in the broader sense of “widely professionally esteemed”) was entrusted to a nucleus of evaluation (see below). The evaluation committee (*from now on*: the nucleus) based their decision on:
  - a) personal knowledge of the candidates by the evaluators;
  - b) documentation produced by the candidates (self-evaluation questionnaire and curriculum vitae);
  - c) the opinion of the parents of students (by way of a questionnaire);
  - d) the opinion of the students of the last two years of high school (by way of a questionnaire);
- *the evaluation nucleus*: in each school involved in the experiment, the assembly of teachers elected two teachers who, together with the principal, constituted the evaluation nucleus. The ratio of two teachers to one principal was decided to ensure on the one hand a prevailing opinion of peers and on the other the significant input of the principal. The president of the School Board,



a parent, was permitted to take part in the work of the nucleus without decisional power simply as a representative of the users and as a guarantee towards the community.

- *documentation provided by the participants*: each candidate teacher presented:
  - a) a curriculum vitae;
  - b) a self-evaluation questionnaire;
- *free will*: given the novelty and the foreseeable resistance, it was decided to adopt a double degree of free will in participation. Schools volunteered to take part only after deliberation of the Teachers Assembly and in each school only teachers who volunteered were evaluated. Only teachers who had at least three consecutive years of service in the same school could participate: this condition served to ensure that the "reputation", on which the evaluation was based, was founded upon a sound knowledge of the candidate by colleagues and users;
- *the number of "merit-worthy"*: Up to 30% of candidates could be selected in each school. This relatively high percentage took into account the fact that not all eligible teachers would have volunteered and had the scope of not overly restricting the available choice. Those selected would be rewarded with a one off payment of one month's salary;
- *"supporting the assessment process carried out by the schools"*: as a design choice, it was desired that the process took place within the school and without external interference, in order to endorse the autonomy of scholastic institutions. At the same time, to provide support to the members of the nuclei and to ensure that their behavior proved comparable (also with regard to the subsequent validation of the experiment), each school was assigned a ministerial expert (ME) and appropriately trained in and informed of the objectives and assumptions of the model. The expert participated in each phase of the nucleus' work only in so far as method was concerned, without interfering in the selection. Each expert followed three schools;
- *the sequence of operations*: the three components of the evaluation nucleus at first examined – independently of each other and without consultation – all the assembled documentation (curriculum vitae, self-evaluation questionnaires and the questionnaire filled by users); they then compared these elements with the "reputation" that each candidate carried within the community according to their opinion; finally they drafted – still without consultation – an individual list of the "merit-worthy" equal in number to the number of awards to be distributed (30% of candidates). The last stage was collaborative: the three members met and compared – for the first time – their individual lists. Those candidates who had been selected by all three evaluators were automatically chosen. Where not all awards were allocated in this phase they proceeded to examine the remaining selections;
- *from the "subjective" to the "intersubjective"*: the sequence illustrated obeyed one of the principles chosen for the model, according to which each subjective point of view can be questioned, but the convergence of several opinions (formed independently) could not be by pure chance. All the elements on which the decision was based were in themselves subjective, but came from different subjects and were based, presumably, on different criteria of evalua-

tion: what made them reliable and substantiated the choice were the circumstances that “indicated” the same people via independent channels. In other words this was the way the individual’s “reputation”, as shared appreciation of evidence of professional merit observed in practice, could be intercepted. The method was developed so as to allow the preservation of the two cornerstones of the conceptual model: the “holistic” evaluation and the shared “intersubjectivity” of the community, i.e in the operating environment, a parameter considered essential to evaluate the real effectiveness of a teacher.

#### **e. *Foreseen limitations***

From the beginning, certain limits were foreseeable and present to both the experts of the TSC and the Ministry. *Valorizza* is only a component of a more complex National evaluation system, which to a large degree remains to be defined and even more so realised. It is true that it does not solve all problems but this was never the intention.

For instance, it does not take into account the relevant problem represented by those teachers who are notoriously “inadequate” and should be put in condition not to damage their students. This issue was discussed in the TSC, but in the end the political decision was to postpone it.

Moreover, it’s likely that the reputation method by shared appreciation and the holistic and contextual evaluation that are the foundations of *Valorizza* maintain their conceptual validity in countries other than that in which they were tested. Instead, the operating procedures followed in this particular case (and which will be discussed below) are difficult to export, since they depend heavily on the characteristics of the Italian school system.

Other relevant criticism, formulated before the experiment, are presented and discussed in Chapter 7.

#### **f. *The implementation of the experiment.***

All testing of the project was carried out between January and June of 2011, in 33 schools identified by the Ministry among those who had volunteered for the experiment in three regions: Campania (12), Lombardy (10) and Piedmont (11).

The month of February was dedicated to the training of ministerial experts (ME) that had to assist schools in the evaluation process, without interfering in the decisions. Each of them followed, as a rule, three schools.

In each participating school, during the same period, the evaluation nuclei were identified: composed of the principal and two teachers elected by the Teachers Assembly. The criterion adopted was to choose between teachers with prestige and experience, esteemed by their colleagues, who explicitly renounced to apply for evaluation themselves. Almost all appointments were made by very wide majority and without particular internal tension. The components of the nuclei also participated in a training session, held in the capital of each region.

During the month of March, the ME visited schools participating in the project, explaining in detail its contents and operating method. At the same time, they also responded to requests for clarification and to the criticisms brought up by those who dis-

agreed. A selection of these criticisms are shown later in this publication, together with considerations on their merit (see chapter 7). Only after this process was the collection of applications opened.

These were to be submitted on a confidential basis to the school principal, who was committed to ensuring the anonymity of the competitors. No school has never made public the list of those who volunteered as candidates for evaluation.

The evaluation nuclei used the whole month of April and beginning of May for the examination of the documents submitted by the candidates and the questionnaires filled out by the users (parents of all students and students of the final two years of high school). This examination was carried out by each of them separately from the others and without communication, using the method described in the previous section.

Once this examination of the material was completed, the nuclei were brought together in the presence of the Ministerial expert to identify that 30% of the candidates who were considered as particularly merit worthy, based on the criteria already described. This phase took up most of the month of May.

On May 31 - the date indicated by the Ministry - all schools published the names of the selected candidates, in alphabetical order. The names of those not selected were not published. This precaution, along with the anonymity of the candidates had a specific purpose. No one, outside of the three members of the nucleus (which were sworn to secrecy), could establish a ranking among the teachers within the school based on the list of the winners, because it was not known against whom they were compared or even if those not included in the list had participated. In practice, this measure severely dampened controversy and internal tension.

There were 905 candidates from the 33 participating schools. Of these, 30% were selected as being particularly merit-worthy.

The evaluation carried a financial reward: one month extra salary, which was paid by the Ministry. The two teachers who were part of the nucleus (not the principal) received a fee equal to half of the sum paid to the others, in recognition of the work done and the waiver to apply themselves.

A small amount (of between one thousand and two thousand euro, depending on the number of candidates) went to schools to recognize the additional workload for secretarial services and office costs.

## Part Three

### Analysis of the results

#### 6. The validation of the experiment

An experiment is nothing without the study of its results and their correspondence to the objectives defined in the planning of the project. In order to ensure maximum independence in the process, the Ministry decided not to evaluate the results of the experiment only on its own, but to engage two independent foundations (**Treille Association** and the **Foundation for the School of the Compagnia di San Paolo**) which were required - through a special agreement - to conduct the necessary surveys and produce a research report (released December 7, 2011 at a conference in Rome, where the leaders of the Education Division of the OECD, the Minister and the head of INVALSI were present).

The validation process obviously based itself on the underlying assumption: the *Valorizza* project was intended to identify not the “best” teachers in absolute, but those who - in their school - had the highest professional reputation. For this reason, the evaluation intended to ascertain, with instruments designed ad hoc, the effective degree of consensus that the actual selection made by the evaluation nuclei met within each institution, also by reconstructing the procedure adopted in the school. It was necessary therefore to test whether the awarded teachers were actually the most appreciated in their own communities. This thus gave rise to a qualitative and a quantitative analysis of the whole process.

The **qualitative** analysis was carried out in 11 schools out of 33, chosen according to appropriate criteria (geographic location, users, school type, characteristics of the principal, etc..). In these the experiment was monitored by researchers from the Foundations, acting as non participating observers.

The activities they observed – but never intervened in - were: the presentation of the project by the Ministerial expert, the election of teachers to the nuclei, the training meeting of the nuclei in the regional capital, the concluding work sessions of the nuclei.

In June, the researchers went back to the eleven schools in the “sample” to carry out a series of in-depth interviews with the principal, a component of the nuclei and a non candidate teacher. The aim was that of obtaining - in light of the results by this time published, - any reactions significant for the purposes of the qualitative analysis and the validation.

The **quantitative** analysis was carried out in all 33 participating schools, although some of them - for various reasons (mainly because of overlap with exams) - did not provide all the requested responses. It was based on the analysis of two questionnaires - similar to each other – one of which was provided to all the teachers of the school (whether or not they were candidate) and the other to the users (parents and students).

There were only two questions, very simple and straightforward: 1) Is there anyone among the awarded that you feel did not deserve this recognition? And, if so, how many? 2) Are there people not on the list that you regard as more deserving? And, if so, who? The answer to this last question was optional.

The main findings of the quantitative and the qualitative valuation are given in the following sections.

### **a. The quantitative analysis: some results**

In this section we present a subset of the in-depth quantitative analysis contained in the research report available on the websites of the two foundations. It is useful to focus on the following points:

- basic quantitative data
- degree of internal convergence within the evaluation nuclei in identifying the merit-worthy
- degree of convergence between the judgments of the nuclei and those of teachers, families and students
- stakeholder perception of the ability of the method to identify the most deserving

It should be noted at the outset that the quantitative analysis relating to the first two aspects was conducted in all the 33 schools that participated in the trial. That relating to the other two, which was based on questionnaires distributed in June 2011, after the publication of the results, however suffered some limitations, since some schools were unable to distribute or collect questionnaires and others only partially provided the requested data. In any case, 23 schools responded fully to all requests, while 10 responded only partially. This partial mismatch between the analysis conducted up to publication of results and that subsequently carried out should be taken into account in the comparison of the data. Where relevant, this data is indicated in the analysis that follow.

#### **a1. Quantitative data**

**tab. 1 – participating schools by type**

scuola primaria	6	primary school
istituto comprensivo	15	comprehensive school
scuola secondaria I grado	3	lower secondary school / junior school
istituto professionale	3	further education institutes
istituto tecnico	3	vocational institutes
liceo	3	grammar school / lyceum
total	33	

**tab. 2 – participating schools by region**

Campania	12
Lombardy	10
Piedmont	11
Total	33

The thirty-three schools that participated in the trial did not constitute a true statistical sample of all Italian schools, since participation was voluntary. However, they can be considered a good representation of those oriented to the evaluation of their teachers. This provides useful information in the event that appraisal of teachers should be adopted by progressive adhesion, rather than an authoritative act of legislation (see Chapter 9).

Within these limits, Tables 1 and 2 show that the groups are representatively diverse, at least in terms of type of school and geographical distribution. Therefore, the results – all other conditions being equal - can be considered sufficiently indicative with respect to these two variables.

**tab. 3 – dimensional aspects of participating schools**

	minimum	average	maximum
enrolled students	237	738.45	1479
classes	11	35.45	68
annexed locations	1	4.22	14
total teachers	23	85.15	200
potential candidates*	16	53.93	122
candidates	3	27.42	86

\* potential candidates were required to have at least three years of service in the same school

Table 3 provides several useful insights. The 33 schools in the sample have very different sizes (from a minimum of 237 to a maximum of 1479 students) and are located in different contexts: from the center of large cities such as Naples (school 3 in Table 9) to provincial towns (school 4 in tab. 10) and even small islands (school 9 in tab. 9). The number of teachers qualifying for candidacy (three years continuous service in the same school) compared to the total number of teachers is also significant: 53.93%, just over half. This confirms the data already shown by many studies: there is an excess of staff mobility between one school and another, which does not favor the creation of long term project teams and stability of employment. If schools had a say in the choice of their teachers - as in many other countries - this fact would be different.

**tab. 4 – candidates to non-candidates by school type\***

	candidates	non candidates	total	% candidates
primary schools	208	166	374	55.61
comprehensive schools	362	487	849	42.63
lower secondary/junior school	107	64	171	62.57
further education institutes	107	176	283	37.80
vocational institutes	74	108	182	40.65
grammar school/lyceum	47	93	140	33.57
total	905	1094	1999	45.27

\* regarding only candidates, schools that provided complete data

Table 4 provides another interesting insight: the willingness to be evaluated is highest in primary schools and lower secondary schools (nearly 60%), but falls to around 40% in other types of school. High school teachers seem less cooperative: barely one in three joined the trial. A significant overall fact remains: more than 45% of those who were eli-

gible to apply did do so, almost one in two. This finding is consistent with those of the research IARD and ANP-NOMISMA mentioned in Chapter 3.

**tab. 5 – candidate teachers and non candidate teachers by region \***

	candidate	non candidate	total	% candidates
Campania	424	398	822	51.58
Lombardy	187	276	463	40.38
Piedmont	294	420	714	41.17
<b>totale</b>	<b>905</b>	<b>1094</b>	<b>1999</b>	<b>45.27</b>

\* regarding only candidates, schools that provided complete data

Table 5 adds another piece of information: the propensity to get involved in the project was highest in Campania (over 50%) compared to 40% in the other two regions. It is worth mentioning that the number of schools that had originally adhered to Valorizza was also significantly higher in this region. There is insufficient evidence to correlate this fact.

**tab. 6 – other characteristics of candidate teachers**

% female	82.51
average age	49.98
years of service	13.02

Table 6 provides further information, or rather confirmation. More than 82% of the candidates were women, proof of the prevalence of women in education: a prevalence that is set to become virtually a monopoly. The average age of fifty years is lower than the overall average of the profession, which lies between fifty-five and almost sixty. This seems to indicate - again no surprise - that the relatively younger teachers were more willing to take part. The data on length of service may instead be a surprise at first sight: thirteen years for people of an average age of fifty suggests entry into the profession at 37 years of age. But the figure should probably be read in a different way: it almost certainly indicates the period of "stable employment" and not the start of the teaching career. If it is so - as is likely - it confirms what is an already known and problematic fact in our country: the abnormal duration of a state of job insecurity before the consolidation of a long-term working relationship.

**tab. 7 – level of user participation in the process - parents**

	minimum	average	maximum
questionnaires distributed	175	646	1441
questionnaires returned	13	388	1200
rate of returns (percent)	6	<b>63</b>	100

\* data per school over a total of 33 schools

**tab. 8 – level of user participation in the process - students**

	minimum	average	maximum
questionnaires distributed	65	226	434
questionnaires returned	58	149	293
rate of returns (percent)	33	<b>68</b>	89

\* data per school over a total of 9 secondary high schools (the only type in which the students of the last two years were involved)

Tables 7 and 8 return interesting data about the degree of participation of users in the evaluation process. The data shows the minimum, maximum (absolute) and average number of questionnaires distributed in a single school. It should be noted that the parent's questionnaire was only one per family (not one for each parent). These figures, cross referenced with those of Table 3, show that almost 90% of the parents received the questionnaire and the response rate was approximately 60%. The first data measures the ability of schools to involve their users in the project, and the second indicates the level of interest with which the users accepted the proposal. A much higher than expected result: this encourages a different perspective to be taken when considering the role that parents could exercise in school life, when they perceive that their opinion matters. Regarding students it is more difficult to infer from the data, as the available data permit the calculation of the overall student population of 33 schools, but not the sub-group recipients of the questionnaire (pupils of the last two years of high school of the only 9 such schools in the sample).

## a2. Degree of internal convergence of the nuclei in their choice of the merit-worthy by region

**tab. 9 – level of convergence in the decision of the three evaluators school by school - Campania**

	total teachers in the school	eligible teachers	candidates	merit-worthy 30% candidates	selected by 3 unanimously	selected by 2 (initially)	selected by 1 (initially)
school 1	52	38	35	11	7	4	0
school 2	76	57	28	8	4	3	1
school 3	200	122	86	26	13	12	1
school 4	74	44	20	6	4	2	0
school 5	67	47	33	11	8	3	0
school 6	67	45	19	6	4	2	0
school 7	104	75	32	10	7	3	0
school 8	74	47	37	11	7	4	0
school 9	114	75	70	21	15	6	0
average (numerical)	92.00	61.11	40.00	12.22	7.67	4.33	0.22
average (percent)			64.9		<b>63.9</b>	<b>34.3</b>	1.8

\* Data refer to the nine schools in the region who at the time of drafting the report - had returned complete data. The merit-worthy were equal to 30% of the candidates. The candidates initially selected by 2 out of 3 evaluators were compared with each other to choose the most deserving. Only in two schools was it necessary to push the analysis to a comparison between those who had been indicated by a single evaluator (due to a relative fragmentation of individual judgments).



**tab. 10** – level of convergence in the decision of the three evaluators school by school - **Piedmont**

	total teachers in the school	eligible teachers	candidates	merit-worthy 30% candidates	selected by 3 unanimously	selected by 2 (initially)	selected by 1 (initially)
school 1	118	95	58	17	8	9	0
school 2	63	27	15	5	3	2	0
school 3	69	30	15	5	3	0	2
school 4	65	43	23	7	6	1	0
school 5	105	75	34	9	7	2	0
school 6	79	19	19	6	5	1	0
school 7	76	52	46	14	14	0	0
school 8	72	47	28	8	5	3	0
school 9	115	90	39	12	5	6	1
average (numerical)	82.11	53.11	30.78	9.22	6.22	2.67	0.33
average (percent)			61.9		<b>68.7</b>	<b>26.0</b>	5.3

\* Data refer to the nine schools in the region who- at the time of drafting the report - had returned complete data. The merit-worthy were equal to 30% of the candidates. The candidates initially selected by 2 out of 3 evaluators were compared with each other to choose the most deserving. Only in two schools was it necessary to push the analysis to a comparison between those who had been indicated by a single evaluator (due to a relative fragmentation of individual judgments).

**tab. 11** – level of convergence in the decision of the three evaluators school by school - **Lombardy**

	total teachers in the school	eligible teachers	candidates	merit-worthy 30% candidates	selected by 3 unanimously	selected by 2 (initially)	selected by 1 (initially)
school 1	79	29	17	6	5	1	0
school 2	91	58	12	4	4	0	0
school 3	91	69	27	8	3	4	1
average (numerical)	87.00	52.00	18.67	6.00	4.00	1.67	0.33
average (percent)			39.5		<b>73.5</b>	<b>22.3</b>	4.2

\* Data refer to the nine schools in the region who- at the time of drafting the report - had returned complete data. The merit-worthy were equal to 30% of the candidates. The candidates initially selected by 2 out of 3 evaluators were compared with each other to choose the most deserving. Only in two schools was it necessary to push the analysis to a comparison between those who had been indicated by a single evaluator (due to a relative fragmentation of individual judgments).

Tables 9, 10 and 11 are perhaps the most interesting with regard to the ability of Valorizza to bring out the wide spread opinion existing within the school community regarding the meritability of its own teachers (at least among those who were candidates). The data are shown analytically school by school and region by region, to highlight - beyond the simple average - the homogeneity of conduct of different evaluation teams. It is very significant that people from different backgrounds, who never met and who analyzed very different materials from each other arrived at conclusions so strikingly similar.

To summarise, and overlooking the details, more than two-thirds of the teachers were

deemed worthy by unanimous opinion of the three assessors of their school: judgments that each formed independently of each other and without starting from shared categories of value. Indeed, in Lombardy, the percentage exceeded 73%, almost three quarters. And even those who did not initially achieve unanimity, however, were chosen by two out of three evaluators and represent almost the entire remaining third of those selected. The remainder (9 individuals out of 276 - only 3%) were selected initially by only one of the evaluators. This happened in the few schools where there was some fragmentation of judgment, in which each evaluator indicated different names from those of the others.

**a3. Degree of convergence in the selection of the nucleus and the teachers, families and students**

**tab. 12 – stakeholder participation in the validation of the results – teachers**

	minimum	average	maximum
questionnaires distributed	23	79	200
questionnaires returned	7	36	79
rate of returns (%)	12	<b>45</b>	77

*\* figures per school against a total of 26 schools that returned results*

**tab. 13 – stakeholder participation in the validation of the results – parents**

	minimum	average	maximum
questionnaires distributed	245	651	1485
questionnaires returned	20	217	610
rate of returns (%)	2	<b>33</b>	81

*\* figures per school against a total of 26 schools that returned results*

**tab. 14 – stakeholder participation in the validation of the results – students**

	minimum	average	maximum
questionnaires distributed	72	227	435
questionnaires returned	46	126	223
rate of returns (%)	21	<b>57</b>	89

*\* figures per school against a total of 26 schools that returned results*

Some weeks after the publication of the results (at the end of June 2011), an immediate validation of the results took place, carried out by the two Foundations, the purpose of which was to measure the degree of consensus of the various components of the school (teachers, parents and students) with respect to the decisions of the nuclei. As previously mentioned, this stage experienced considerable difficulties due to the period of the academic year in which it took place (at the end of lessons and during the course of

reports and final exams). This accounts for the relatively low rate of return of questionnaires distributed on this occasion compared to those that were used in the process of the holistic assessment.

The level of stakeholder participation in this phase can be deduced from tables 12, 13 and 14. As for the teachers, the rate can be said to be significant although not very high - given the timing difficulties. In fact, cross referencing again the data of Table 3 with those of Table 12, it can be concluded that over 40% of potential recipients completed the questionnaire (an average of 36 per school against an average of 85 teachers). If we consider only the questionnaires actually distributed (79), the response rate goes up even more.

As for the parents, whose overall figures can be estimated by induction from the numbers of the students in Table 3, the response rate (compared to questionnaires distributed) is around a third. In this case it is perhaps more significant to compare the 217 average responses recorded on this occasion with the 388 of the first questionnaire (the evaluation questionnaire distributed in April): the ratio between the two data exceeds 55% and is therefore quite indicative, no less so than the judgment of those who were directly involved in the trial (and thus demonstrated greater interest in the matter).

As for the students, Table 3 does not provide the information required to estimate the subset of “eligible voters” (students in the last two years of high school). The direct response rate (questionnaires returned/distributed) is already high (57%). But even more significant is the comparison of absolute levels of response to the two questionnaires (the “evaluation” of April and the “validation” of June). This indicates that 85% of students who had expressed an opinion in April expressed their opinion again in June (129 against 146 on average for each school). This, at the end of the day, was an excellent result.

**tab. 15 – how many non merit-worthy teachers were rewarded according to the stakeholders\***

	teachers %	parents %	students %
minimum	16	10	11
average	34	32	35
maximum	66	100	78
schools considered	26	23	8

*\* the percentages indicate what percentage of winners was held to be not worthy by each of the components. It does not automatically indicate the percentage of disagreement with the decisions taken.*

**tab. 16 – how many merit-worthy teachers were rewarded according to the stakeholders\***

	teachers	parents	students
minimum	3.2	1.9	1.9
average	5.1	2.4	2.3
maximum	11.2	2.8	2.7
schools considered	26	23	8

*\* The numbers indicate the absolute number of teachers that the various components considered merit-worthy but did not feature on the list of those rewarded. Please note that given the anonymity of candidates these figures include a number (not determinable for a lack of data) of teachers that were not candidates. The gap between maximum and minimum, relating to absolute numbers, also reflect the size of the schools.*

The considerations made about tables 12, 13 and 14 are useful in appreciating the substantial relevance of those that follow, i.e. the degree of convergence between the judgment of the nuclei and those of the various stakeholders: teachers, parents, and students.

This information emerges from Tables 15 and 16. The first reflects the response of those answering the question “among the teachers selected as merit-worthy are there any who you think did not deserve to be?” (They were not asked for names, but – if possible – the number). Fluctuations, beyond isolated cases, are not relevant: taken on average, which is the only significant finding in this context, the lack of consensus remains around a third. The parents are most in agreement (only 32% of lack of consensus), followed by teachers (34%) and students (35%). The percentages refer to the proportion of teachers “awarded without merit” in the judgment of the various components. And so stakeholders do not agree, on average, on only about one third of the names chosen by the three evaluators.

Interestingly, two-thirds - as we have seen - was the degree of consensus inside the nucleus on the names of the merit-worthy: two out of three were chosen unanimously right from the start. It almost seems that the lack of consensus among the stakeholders is directed at that third for which unanimity was initially not there. But this is an inference that has no supporting elements and which therefore should be left suspended.

The second table (n. 16) reports the results of a question that mirrors the first: “Are there teachers who merit recognition and have not received it?”. The results should be read with particular caution, as respondents could not know if the teachers they regarded as merit-worthy had not been “rewarded” as a result of the decision of the three assessors or were in fact not candidates. So an answer in the affirmative is not automatically to be read as a complaint regarding the decision of the nucleus.

The results tend to diverge more than in the previous case. Parents and students are in relative agreement that, on average, just over two deserving teachers (per school) were not taken into consideration. For teachers this number, always on average, is higher and is around five (with peaks up to eleven). These are absolute numbers (taken on average per school) and not percentages, because there is no basis on which to calculate such a value (percentage of what?). Therefore, the values are also influenced by the size of the school. In a school with 200 teachers and 86 candidates, it is more likely that a larger number of those deserving but not awarded will arise.

In any case, in the worst case scenario, the five merit-worthy teachers not rewarded by the nuclei are to be compared with the average of 85 teachers of each of the schools involved, or the 58, again on average, that were not candidates (table 3). It is a comparison that has no real value in the discussion, given that not all were eligible to apply, but allows us to say that the “absolute” level of dissent seems to remain marginal.

**a4.** *Degree of convergence in the selection of the nucleus and the teachers, families and students*

**tab. 17** – profound divergence between the judgment of the nucleus and the opinion of stakeholders\*

	teachers %	parents %	students %
minimum	0	7	19
average	38	22	30
maximum	71	100	43
schools considered	26	23	8

\* by “profound divergence” it is intended the contemporary belief that the nucleus indicated merit-worthy teachers who were not so, and that they failed to reward teachers who were so. The values are weighted to returned questionnaires (not the total sent out, or the total of the population in question). In theory - but only in theory - one could assume that those who did not return the questionnaire had nothing to report, either positively or negatively (and thus were not unhappy with the outcome)

Table 17 examines the worst case scenario, the greatest divergence between the opinion of the nucleus and that of stakeholders: in which there co-exists a negative judgment on the names of the merit-worthy (some of them would not have been deemed so) and a further negative judgment on those not rewarded (some of whom deserved to be so). How many people share both of these negative evaluations?

It is hard to say exactly, since the questionnaires were anonymous, but the two questions were listed on a single sheet of paper and can therefore, with a certain approximation, be attributed to the same respondent (although not all responded to both questions). In any case, according to the available data, the widest dissent concerns only 22% of parents, 30% of students and 38% of the teachers. Even considering the latter figure, which may seem high (but is still a minority), these figures confirm the trend already seen: on average, two-thirds of the interested parties agree with the opinion expressed by the nucleus.

This could be taken further, noting that those who did not express an opinion are probably indifferent or uninterested, but they can hardly be categorized among those who disagree. So, in reality, the area of “non dissent” would be even wider. However this is once again an inference that is not supported by hard data.

A reflection: the nuclei had the task of selecting the merit-worthy to the extent of 30% of the candidates. In two cases out of three they were in agreement from the outset without ever having discussed their choices. And in two cases out of three, teachers, parents and students approved their choice.

This finding could provide the starting point for future repeats of the Valorizza experiment: it seems to indicate there is universal consensus about the first two thirds of the 30% selected, that is 20% of total candidates. It would therefore be appropriate to restrict the number of awards to this level, to reach agreement easier. The same conclusion is arrived at by the research conducted among teachers and already mentioned in chapter 3: here too respondents indicated a 20% range deemed excellent among teachers. Could this just be coincidence?

In conclusion, the quantitative analysis speaks quite clearly: not only did Valorizza reach the goal of identifying the “worthy” teachers by general professional appreciation within

their own school, but agreement without substantial difficulties among a significant majority of those who responded was found. This holds true for the components of the nucleus, for school teachers, parents and students.

Whether it can be said that the result is such that the method can appropriately be extended to all schools, is not beyond any possible objection. In any case, consensus at these levels is not easy to achieve with other methods. The judgment was expressed by – and only by – people with at least a three years direct relationship of study and work with the candidates. It is still to be seen whether other evaluators, different from these, could do the same with as much knowledge of the facts and with as much consensus within the school community.

### ***b. The qualitative analysis: conclusions***

This section summarizes the main findings of the research and qualitative analysis carried out in a representative sample of 11 schools from among those participating in Valorizza. The aim was to reconstruct the experimental process, identify strengths and weaknesses and record the input of those who took part, collecting comments, suggestions and proposals, including through in-depth interviews carried out in schools. This conclusion briefly presents the most salient aspects that emerged from the analysis.

The method, it is to be remembered, was based upon non-participatory observation during the course of the trial, supplemented by in-depth interviews conducted with some of the participants from the schools after its conclusion (principals, a teacher evaluator, a teacher who was not a candidate). Of course, what follows is in the form of considerations relating to the main significant points that emerged during the process.

#### *The appreciation of the experimental nature of Valorizza by schools*

In a troublesome context, bottom-up participation and the possibility (that the Ministry wanted to assure schools) to make proposals for improvements in the construction of an evaluation system was highly appreciated by principals and teachers, and this is the aspect that led some institutions to take part in the project.

#### *Valorizza's ability to meet the widespread desire among teachers for their professionalism to be recognized*

Valorizza met a widespread desire among teachers for recognition of their professionalism. However, it must be stressed that the study sample of the Foundations is small and clearly not representative of the universe of Italian schools, but it is representative of those schools who volunteered and may be deemed as particularly favourable to the evaluation of their teachers.

#### *The central role of the head teacher and, more generally, the leadership of the institution in the implementation of Valorizza*

The role of mediators, particularly the heads and ministerial experts was crucial especially in the first stage of the process. The leadership and management skills of principals, along with the ability of ministerial experts to clearly explain the features of the evaluation, to read the climate of the school and, subsequently, to calibrate their communication, were fundamental in ensuring the transparency of the process and in compensating for the ini-

tial information gap, thereby facilitating the creation of a climate conducive to experimentation.

*The effectiveness of the choice of an internal evaluation committee and the process of election of the nucleus*

The decision-making process, as described by the evaluation model, was effective and able to guarantee agility and independence of judgment without internal conflict in the nucleus. Going more in depth into the merits of the evaluation process, the evidence reported shows that the method of election of members of the nuclei found favor among the protagonists of the trial. It should be noted that the evaluators were elected among those who volunteered for the role; this regarded a limited number of teachers (2 to 5 per school) and the election was done by large majority through secret ballots.

*The need for more in-depth training of the nuclei members on the logic of the reputational method*

On the contrary, what emerged as a criticism of the process was the training phase of the members of the nucleus: a need was felt, mainly by teachers, for greater preparation that explained in depth, for example, the logic of the reputational method and its connection with the instruments adopted in order for the evaluators to operate more securely and professionally. One of the recurring concerns of teachers regarded the evaluation criteria that the nuclei used to define their judgment; it was difficult to understand the role of the ancillary tools with respect to the forming of an opinion on the basis of reputation. The request by some to indicate the “weights” attributed to the different documents on which the process was based is revealing.

*The tension between the reputational method and the use of standardized instruments (aimed at the production of mainly numerical output)*

Overall, the reputational method was understood and applied, but at the same time each nucleus employed elements and strategies of modulation and compensation to reach a compromise between the request of a holistic judgment and the presence of formalized assessment tools (CV, self-assessment questionnaire).

*The significance of teacher participation in the evaluation phase.*

In the eleven schools examined, half of the teachers who were eligible for evaluation actually applied, including some of those who were initially against the experiment. This fact helps in gaining a better understanding of the initial resistance of those same teachers towards Valorizza: where the model and its aims are presented in the required detail, teachers seem inclined to take up the challenge of evaluation.

*The effectiveness of the decision-making process: the proper functioning of both the individual phase and collegial work of the nuclei*

Despite the different methods of evaluation of the individual members, it is striking how well the collaborative phase of the nuclei worked: the drafting of the list of merit-worthy took place in a relaxed atmosphere, thanks to the spontaneous convergence between the lists of individual assessors. 66% of those awarded were present in all three lists produced by the evaluators in the individual phase. The data also show that 33% of those award-

ed were present in at least two lists. The data confirms, therefore, the strong convergence, effectiveness and flexibility of the decision-making process provided by the model. Criticism is mainly due to cases of fragmentation at the local level across multiple locations or different curricular sub streams. This resulted in the teachers in the nuclei having some problems in the evaluation process because of a lack of direct knowledge of all those being evaluated.

### *Student and parent participation*

One aspect that deserves a certain degree of caution concerns the participation of parents and students in the assessment process. The participation of the family was not homogeneous across the sample: Piedmont and Lombardy resulted better than expected and the evaluation unit found the preferences expressed by parents to be an important element in the definition of the list of teachers who enjoyed general appreciation; Campania, however, the participation rate of households was quite low, especially in schools situated in disadvantaged areas and with a lower middle class demography. Teachers and principals recommend a more precise identification of the aspects upon which the advice of parents should be sought. The data on the participation of students (only the last two years of high school) show that the latter participated in the evaluation seriously, delivering an informed and reliable opinion. The protagonists of the trial, therefore, hope in the future for the more extensive involvement of the pupils, as early as the first years of high school.

### *The intersection of the points of view of the various components of the school as a strength of the reputational model*

Despite the critical nature of each assessment tool made available to the nucleus, the possibility that the model offers to cross reference different points of view - the actors (teachers and principals) and users (parents and students) - in the evaluation is one of its major strengths. The holistic and plural character of the model is appreciated, which makes the process of evaluation original and effective, and provides the possibility of adopting an internal evaluation logic that makes the school community the protagonists and allows one to contextualize the model itself, basing it on the characteristics of individual school contexts.

### *Weaknesses of the model according to the protagonists of the experiment: one-off rewards, careers, professional categories and feedback*

Among the criticisms addressed to the model by the participants of the trial we can highlight the following aspects: on the one hand, a one-off reward (evidently connected with the limits of a trial) and the prudent silence of the Ministry of a possible (though foreseeable) connection between the award and forms of career advancement and/or promotions; and second, the lack of feedback to those evaluated that, highlighting the strengths and weaknesses of candidates, may support, sustain and enhance the formative value of the evaluation process.

### *The first effects of Valorizza in the schools*

In relation to the first effects of Valorizza, the in-depth interviews carried out a few weeks after the publication of the results (at the end of June 2011) show a reassuring picture: the



principals interviewed stressed that there were no apparent negative implications of the experiment on the academic climate, nor was there observed a negative impact on the level of internal cohesion. The publication of the list of the awards, despite the discontent that was created immediately in some school settings, does not seem to have had an impact on the school climate. Further tests would be desirable in the future. Some principals reported a positive impact on teachers relating to the intention to find paths towards self improvement and assuming responsibility for their professional development.

#### *Trust and full responsibility accorded to the school community*

Valorizza is an evaluation system that attributes a central role to the professional community of the school; in a widespread scope of accountability, it also makes parents and students part of the evaluation process. This structure is reinforced by the choice not to entrust such an important function to external experts (inspectors or others). In this sense, the choice of this model to enhance the professionalism of teachers is a display of faith by the Ministry in the full autonomy and responsibility of the school community as a whole.

#### *A model managed by the school community, contextual, flexible, quick and economic*

Valorizza made the internal staff of the school the protagonist of the evaluation process and centered it in the inner context of the school itself. Despite the tight schedule, all schools completed their work on time, the instruments were generally used more or less correctly and the work of the nucleus was effective.

#### *The desire for participation and continuation*

What was evident from the respondents, within a few months of the end of the evaluation process, was the fear that this experiment, like others in the past, would be a one off. All executives reported a strong desire among colleagues to be part of a continuation of Valorizza in order to optimize their experience and to ensure that the wealth of accumulated knowledge is not lost.

## **7. Valorizza: criticism and response**

### **a. Premise**

In Italy, as has been said, there is no evaluation of the profession of teachers and their unions have even always opposed - so far successfully - any attempt to introduce it. It is understandable that the decision of the Ministry to initiate an effort in this direction (although experimental) generated strong resistance, which were expressed in the form of criticism of the model before it was put to the test and then during and after the experimentation.

In the first section some of the criticism made before the trial are presented. All, or almost all, are about the principles on which it was based, or the feared consequences. This does not necessarily mean that they are the result of prejudice: several are perfectly understandable, at least in a context in which the evaluation of teachers has always been absent. These points have been the subject of deep discussion and debate in the TSC: the guidelines that ensued were taken in agreement with the Ministry. This account, then, has principally an informative function.

Below (see c) are a selection of the most frequent criticisms that were recorded during the trial period from the participants involved, collected by ministerial experts or through observations and interviews of the researchers appointed by the Foundations.

Many of them are reasonable and justifiable, mainly because they are based on the experience of the process and not on ideological considerations. In the text that follows, they are discussed taking into account the empirical evidence gathered and the results of the experimentation available at this time.

### ***b. Criticism before the experiment***

All major changes intimidate and require a leap into the dark with a certain amount of risk: that is, the abandonment of an established condition, understood, perhaps not satisfactory, but which appears at that moment preferable to another undefined condition, which generates anxiety precisely because the full implications and consequences are unclear. The main criticism are:

1. *An experimental model that introduces individual awards to deserving teachers is likely to generate competition among teachers, rather than collaboration.*

Competition and collaboration are not mutually exclusive. In all working environments where many people contribute to the same end, both dynamics are present. One of the explicit objectives of the project is to generate positive not competition but positive emulation among the teachers.

Obviously, this is not an attempt to start a war of all against all: after all, what is at stake in this trial (a one month's salary one-off payment) was not likely to prompt this scenario. But it is reasonable and intuitive to believe that collaboration between people who individually commit to self improvement produce better overall results than collaboration between people often unmotivated or encumbered by routine and the lack of any form of recognition of individual merit.

If the experiment was repeated and sooner or later became common practice (see ch. 9), the awarded would not always be the same and the assessment would lose its aspect of a final verdict to become a periodic check-up whose results are to some extent reversible, both positively and negatively.

2. *Those awarded will always be a minority which may create the view that the majority are sub-standard.*

The objection would be valid if this trial remained an isolated event. Consequently the verdict inherent in this situation would be untested.

If, however, the experiment is repeated and, sooner or later, became fully operational, the group of those awarded would be expanded gradually to include a number that can be estimated between 1.5 and 2 times the percentage of those awarded in a single event.

In fact it is reasonable to assume that, while the outstanding will always remain at the top, the wider range of those who are "good" and esteemed, but so far do not excel, will be encouraged to commit themselves to excel. In short, if the awards were to remain at 30%, those driven to improve would be at around 50%: what makes a big difference.

3. *It is possible, by chance, that non-excellent teachers in a school are rewarded because of the low to medium levels of competitors, while in another, where many good teachers are present, some remain excluded.*

This statement is theoretically well founded, but the value of a teacher can not be separated from the context in which they operate. If this was not so, it would be hypothetically possible to draw up a national ranking of more than 700,000 teachers!! But comparison is possible - and it makes sense - all conditions being equal: that is within the context of the individual school.

If the model was in future fully implemented, some compensation would occur by the spontaneous movement of teachers from a school deemed too "competitive" to others. A dynamic that, if put into practice, would help to distribute the best teachers across a wider audience of school, the same who now tend to concentrate on certain schools, those that are attended by better users or have, in general, a better reputation. This would increase the likelihood that all students - at whatever school they are at - have the opportunity to learn from some good teachers.

4. *Many people think that the effectiveness of teaching provided depends on the school's team of teachers much more than on the individual. Why then perform an experiment to identify and reward the individual?*

Valorizza is not an isolated experiment: it works in a parallel process, which aims to explore the effectiveness of schools and their leaders. The two paths are conceptually complementary, in that they address the issue from both points of view.

Incidentally, it is curious that - while more studies show that the quality of teachers is the single variable most correlated with the educational success of students - there is still resistance to evaluate this aspect.

5. *A model that favors the evaluation of teachers with a proven reputation involving only internal evaluators (principal, teachers, families and students) and not other stakeholders outside the school (inspectors, head teachers from other schools, etc.) risks becoming self referential.*

The underlying assumption must be kept in mind: that the teaching profession can not be measured analytically, by the sum of "objective" requirements, observable from the outside. In short, it is not the individual elements that make up a successful formula, the result can be better appreciated as a whole and only in the specific context. Once again - if an external objective evaluation were conceivable - you should be able to create a general ranking throughout the entire country. Which appears, at first glance, like a paradox much evident than the feared risk of self reference.

Teachers identified and rewarded by Valorizza are not the absolute best, they are just the most appreciated in their school. But it is in that school that they teach, not in an abstract national school-type, which does not exist.

6. *The Valorizza model uses non-objective indicators, including self-evaluation of the teacher and user satisfaction.*

This objection does not take into account the overall ratio of the model, which tends to attain objectivity (in the sense of maximum inter-subjective sharing) not through elements that are per se objective (which do not exist), but through the intersection of a plurality of points of views expressed by different stakeholders, all directly related to the candidates and able to express an opinion on their performance. Each individual

subjective point of view is fallible, while it is very unlikely that, taken all together, all are wrong, yet converge on the same result.

As for the so-called "objective" indicators of the professional quality of teachers, the doctrine has never been able to identify them with sufficient certainty. Usually - because they are more readily available and "measurable" - the number of academic qualifications and/or years of professional experience are used as indicators, but there is no empirical evidence that these factors are related with the results of the students.

7. *Families and students can not make a reliable judgment on the professional quality of teachers. In addition, there is a risk that lax teachers are judge more favorably than demanding ones.*

In theory, this claim is credible. It does not take account of two factors: a) the judgment of users is only one element, that gains relevant weight and value only insofar as it converges with that of the three members of the evaluating committee b) none of the evaluation elements in the model has a given "weight": it is up to the evaluators, depending on the context (and therefore also on their knowledge of the user profile) to give weight to those judgments.

In practice, however, the trial has proved the contrary: there has been a remarkable resonance with the judgment of the users and of the evaluation nucleus. Which would tend to show that the users - when their voice is heard voice - tends to express itself more soundly than is usually tended to believe.

8. *Might there be negative consequences in the behavior of households where awards are given to teachers other than those of their children?*

This risk must not be emphasized: this is already the case today when there is pressure from users to enroll their children in certain classes and not in others. However, once the most highly regarded teachers are identified, it will be easier for principals to construct more balanced class teams of teachers, so that all students are afforded the opportunity to make use of at least some of the most respected teachers. And anyway those not awarded are not "bad" teachers, they are just not outstanding.

9. *Rewarding only 30% of the teachers risks demotivating the others.*

This might be true if the model were applied once only. If it become a recurring procedure on a regular basis (for instance, every three years), the percentage of teachers motivated to improve would increase, for two reasons. First, each time a certain number of awards would be given to different people thus enlarging the group of recipients; in addition, even those who are, or at least consider themselves to be merit-worthy, would be motivated to improve, hoping to be rewarded next time. It can be reasonably assumed that the combination of these two dynamics would possibly engage at least 50% of the total.

10. *In the event that the model is universally applied, there is a risk that a "rotation" of the awarded occurs, so as to result in an undifferentiated distribution that tends to reward a large majority of teachers.*

If the selection of the merit-worthy did not happen every year, but, for example, every three years (the reputation of people do not tend to change quickly), it would take many years before such a hypothetical scenario could take place. Meanwhile, many teachers would have changed schools or retired, and so the circle would never close. But even aside from this, it must be considered that the members of the nucleus would

hardly agree to choices clearly contrary to the evidence and so discredit themselves. It is foreseeable that the number of the awarded will increase and this is part of the expected benefits, because it involves a larger number of teachers, but never extends to all.

11. *Valorizza may appear as an attempt to oversimplify a complex problem, a kind of short cut.*

It is true that the method used had among its stated objectives the sleekness and speed of decision process: however, this is not sufficient to call it a “shortcut”, at least in the reductive sense of this term. The problem is not how long and complicated the process is but if it achieves the results and whether these results impinge upon the evidence.

The experiment carried out showed that it is possible to draw up a list of the most appreciated teachers in each school within a few months; more importantly, the surveys carried out during and after the experiment showed a high degree of consensus (on average, greater than two thirds) on the decisions reached, among both the evaluators and the other members of the school community. This level of convergence of judgment is not easily attributable to chance or to a collective error. Who better than those who have direct insight into the daily work of a teacher, is able to tell whether they are suitable?

So, which is preferable and what is more important? How long, bureaucratic (and expensive) a process is, and how the results are then disputed, or a good shortcut that leads to the desired goal?

### ***c. Criticism gathered in the field during and after the experiment***

1. *An experiment conducted in only 33 schools that have voluntarily participated (and which can therefore be considered biased in favor) is not a representative and significant sample for the generalization of the model.*

The argument might make sense if things were to pass directly to a systematic and compulsory application of Valorizza to all schools. If, as seems advisable for several reasons, its extension occurs on a voluntary basis and is gradual and progressive, the perspective is different (see chapter 9).

With this approach, on the other hand, the 33 schools in the experiment constitute a representative sample of schools willing to get involved and to evaluate their teachers. In a country like Italy, which begins from scratch in this matter, it is unrealistic to aim for a sweeping general obligation.

2. *During the observations and interviews conducted by the Foundations, teachers felt that having involved only the last two years of high school students was a limiting factor in the evaluation process and asked to extend the participation to students of five years of high school.*

This is one of the criticisms which need to be considered in depth. On the one hand there is a risk that young and immature students are less “rational” or are more exposed to external psychological pressures. On the other hand, the positive experience of the first trial tempts greater confidence be given to users.

3. *Is it reasonable to ask the opinion of parents for schools in socio-economically deprived areas?*

One should bear in mind that the model tested is set according to a logical framework in which none of the elements of evaluation is decisive alone, but only when it converges with the other.

Moreover, it is the nuclei that “weigh” the single elements, according also to their knowledge of the socio-economic context of the school. But it turned out that, in general, and even in the areas apparently at greater risk, the guidelines of the parents were significantly in tune with those of the evaluators of the nucleus.

4. *Many nuclei have pointed out the difficulty of limiting the list of merit-worthy to 30%, stating that there were several other good teachers who were left out. Why not widen the band of the awarded?*

This choice was made for two reasons. The first: it was necessary to predetermine the cost and operate within the available resources. Second, what ever percentage was adopted, there would always be a number of candidates “tied” with the last on the list.

The analysis of the work of the nuclei and the mode of their choices reveals a fact: in every school in the sample, two-thirds of the merit-worthy were identified unanimously, while the others were deemed worthy, initially, by two out of three evaluators (in some residual cases only by one).

If a conclusion can be drawn, which was not unexpected even before the trial, it is that the band of excellence recognized by all in each school actually ranks lower than 30%. So, if at all, the work of the nuclei would be made easier by reducing the number of awards and not increasing it, because on the excellent all agree, while it is more difficult to agree on those of average worth.

5. *A one-off bonus (one month extra salary) is not incentive enough to ensure permanent striving for professional improvement.*

This is certainly true and in fact the promoters of this method have not indicated this as the choice for the final model.

In the event of subsequent editions of the trial - even more so if it were to become the general system - this is one of the points to rethink very carefully. Of course, there is a problem of resources that will end up influencing policy choices.

6. *There are more than 8,000 schools spread over more than 40,000 physical locations; many of these complexes are small, too small a basis for a comparison on reputation to be built. On the other hand it is unlikely that an evaluation unit will know and be able to properly appreciate the professional reputation of teachers working in other small locations.*

The objection is well founded and will become even more so if the model becomes fully operational. A solution must be sought: for example, a differentiated composition of the evaluation committee.

7. *The adoption of a model of allocation of awards is not accompanied by a model of professional development. The reasons why individuals have been deemed worthy or not are not disclosed. Plans should be made for feed back to be given to both those selected for award and those not, in order to support formative development.*

This is a serious objection, but somewhat off target. The aim of Valorizza is limited to identifying and rewarding the merit-worthy teachers and not to supplying a formative

evaluation for all. The latter constitutes a basic requirement that should be fulfilled, however, with other instruments and other resources. For example, in many school systems that have a long practice of professional evaluation, the task of returning feedback to teachers is separated from the assessment phase and is usually the responsibility of the school head, generally through appropriate interviews.

8. *The model requires the opinion of all the parents and all students in the last two years of high school. Why was it not planned to take into account the opinion of all teachers in the school, rather than delegating this function to only two members elected in the evaluation committee?*

The answer is complex, but can be reduced to a relatively simple consideration: the idea behind the model was to balance many elements of evaluation from different sources, none of which were to be given a prevailing weight over the others.

It is obvious that a form of referendum among all teachers would in practice have limited the choices of the two members of the nuclei, in effect eliminating the contribution of all the other components involved in the process (parents, pupils and principal). In such a case it would be useless to even ask the candidates themselves to contribute to the process, given the obvious prevalence of the judgment of their colleagues.

A dynamics familiar to those who deal with social phenomena must also be taken into account: in a choice given to all (as in a general election) decisions tend to be taken according to political dynamics and, in the end, by endorsing interest groups; in a choice entrusted to a small committee of experts, considerations of a professional nature tend to prevail.

9. *The experiment worked, but what use of it could possibly be foreseen in the future?*

The question is legitimate, but in this first phase the aim of the Ministry was only to test the model and its procedures.

The results seem to show that it is sound and has provided, in a short time and with nominal cost, reliable results (a view shared by a large majority of stakeholders). The eventual follow up, and possible use at the system level, are of course subject to political decision.

For more on this theme, see Chapter 9.

## 8. Valorizza: strengths and weaknesses

*by Dirk van Damme – Head of Centre for Educational Research and Innovation – OECD (article published in the quarterly magazine “Scuola Democratica” – n. 6 October 2012)*

The Valorizza experiment is an original and very valuable experience and a major contribution to the international debate on teacher evaluation. It is widely recognised that a well-performing and balanced system of teacher evaluation is key to the improvement of teachers' professional practice and of the teaching profession. There are many ways to provide reward, feedback, evaluation to teachers. The original contribution of the Valorizza experiment is that it also attempts to provide recognition to those in the profession who are doing an excellent job. Identifying underperforming teachers is not really the purpose of this scheme; neither are providing professional development, coaching, super-

vision or collegial intervention.

The OECD's Teaching and Learning International Survey (TALIS) has extensively documented appraisal and feedback practices for teachers. One of the most surprising findings is that in general teachers welcome feedback and evaluation on their professional practice, especially if it is not really punitive, if it is positive and if it is seen as a contribution to their professional development. Teachers who receive regular feedback from their headmaster and from colleagues tend to take their job more serious, have more job satisfaction and also tend to increase their investment in the professionalization.

The Valorizza experiment adds interesting perspectives to the international practice in teacher evaluation. It would be worthwhile for the experiments and its promoters to showcase its major facets by means of publications in international scholarly journals. I will discuss a few characteristic elements which seem to be inspiring for the international community.

There are three elements in which the Valorizza scheme slightly departs from the mainstream international experience in teacher evaluation. The first is on using reputation, the second is on the peer judgement instead of evidence, and the third one is that the experiment is school based.

I think that the use of reputation is the most questionable element. The use of reputation will lead to a lot of discussion in the international community. "Can we use reputation as a valid source of information?" I would personally say "yes, we can". Certainly, reputation can include distortion and errors, but it is a valid proxy for quality. Perfect objectivity in many cases will be an impossible objective. The real question is whether, from a pragmatic viewpoint, reputation is a trustworthy criterion. In a professional community reputation is based on an aggregation of inter-subjective judgments made by peers over time. It will never be perfect, even when supported by huge packs of evidence. I would say that the opposing a "mere subjective opinion" against "really objective evidence" is a false question. The real question is whether an inter-subjective judgement by peers in a professional community is possible or not. This is the real core of the issue.

Related to this, is the second issue on the evidence base of evaluation decisions. One of the major challenges of any teacher evaluation scheme, which is especially sensitive when rewards are significant, is the evidence base on which decisions are taken. Of course, promoting evidence-based educational policy and practice is a major objective. Because of their impact on the public good, any significant decision in the educational domain should be based on sound evidence. Yet, there is a possible risk of exaggeration here, of bureaucratic overload as well. One of the interesting elements of the Valorizza experiment is that it has demonstrated that it is possible of developing an evaluation scheme which does not involve a huge bureaucracy, a scheme where the decisions are not supposed to be supported by huge bags of documents. What is interesting in the Valorizza experiment is that it puts emphasis on the judgements of peers. And this judgement is based on available evidences where possible, but the decision to make a judgment is the responsibility of the peers. In a professional community teachers are perfectly capable of judging the quality of the professional practice of colleagues. Evidence does not automatically produce a solid judgment and it is not by demanding ever more evidence that the decision will improve correspondingly.

Finally, the Valorizza experiment is school-based: it assesses and rewards the quality of



teachers on a school level and does not pretend to introduce a scheme where absolute measures of quality are used, valid for all schools. I really like this element of the Valorizza experiment. I have to admit that in the initial stage of my encounter with the experiment, I was a bit sceptical because of the inherent inequality introduced in the system. A school based performance assessment, in which a given number of teachers in each school can be rewarded leads to a system in which the chances of an individual teacher to get the reward are not equally distributed over schools. The reward will be dependent of your competitors in the school, on the conditions in which the school has to work, etc. But in the course of the experiment I have changed my mind and the experiment convinced me that is very worthwhile to do it school based. Despite the fact that it leads to a certain level of unequal treatment, it has the important merit of implicitly taking into account the conditions in which school has to work. By emphasizing the school the message is that teaching is not a solitary practice, that teaching happens in a professional community and in a school context. By doing so the Valorizza experiment implicitly supports the development of the school. I think that this is a very important element.

Still, there is a critical element which remains unsolved. Many people in the international community would say that the Valorizza experiment has no yardstick against which teachers are evaluated and that also for me is a little bit a question mark. I think the scheme needs to have professional standards which act as a reference for the evaluation by the peers. For the future development of the experiment it would be necessary to have a more explicit yardstick to which teachers are evaluated: a set of professional standards that could be developed by the Ministry or even better by the professional community itself together with the policy makers. So that there is a recognized definition of what an excellent teacher is.

My conclusion is that the Valorizza experiment is an extremely interesting new approach in teacher evaluation and really deserves to be continued. It contributes to the international experience on teacher evaluation, mainly by its methodological choice for the use of reputation, by emphasizing peer judgement instead of excessive evidence, and by its school-based design. The experiment certainly is not perfect and could improve on a couple of things: the first is by making more explicit reference to agreed professional standards, as discussed. The second potential for improvement is in further developing the feedback towards evaluated teachers. Indeed, the information value of what is happening in a school towards the whole community of the teachers is rather limited. Evaluated teachers deserve feedback and more ample information on why certain teachers are judged as performing on a level of excellence will also drive the behavioural change of other teachers. And the third and final element of improvement concerns the relationship between teacher evaluation and innovation. In general, evaluation has a difficult relationship to innovation: often evaluation rewards conservative behaviour rather than innovative professional practice. In education innovators tend to be difficult people, controversial, maybe not rewarded by colleagues for what they are doing, they are seen as people making unrest. There is a risk in any evaluation scheme to reward conservatism. This risk can be mitigated by including elements of innovation in the professional standards discussed above.

## Part Four

### Can the model be systematically deployed throughout the educational system?

#### 9. A gradual approach.

##### a. *Premise*

In this chapter, the independent Foundations (Associazione Treille e Fondazione per la Scuola) form hypotheses about the possible implementation of the Valorizza method. The following proposals stem from the empirical findings of the research conducted by the Foundations, which suggests an overall suitability of the model of "professional reputation" for the attainment of the desired benefits.

In this regard, it should be noted that the model aspired to function rapidly, to be easy to manage, economic, and have definable fixed costs, to yield results shared in the school community, to ensure the active participation of teachers in the evaluation (through their two elected representatives in the nucleus), and not be open to legal contentious attack.

##### b. *The application of Valorizza throughout the educational system*

The following proposal is intended to respond to three different and legitimate expectations of every merit-worthy teacher; higher remuneration, greater prestige and career development

##### *Higher remuneration for the merit-worthy*

###### *Proposals:*

- the reward for deserving teachers to be raised to at least two extra months salary
- in order to ensure the maximum agreement on the choice, to reduce the number of those awarded to 20% of the candidates in each school
- the award should not be a one-off, but paid for three consecutive years. The triennial bonus would be annulled should the awarded teacher change school of their free will during the course of the three years, because the evaluation of the nucleus is comparative and is linked to the context of a particular school.
- each school should repeat the evaluation procedure every three years in successive consecutive rounds
- the economic award will in no way become permanent.

##### *Greater prestige and career development within the profession*

In this context, the term "professional career" should be intended in the sense of a path within the function of teaching and not that of a migration to different functions.

The list of teachers awarded must be published in the school. In addition, for those who continue to carry out only teaching functions, merit should be further recognized with a "prestige" award (to be made public), conferring the title of "expert teacher" to those evaluated as merit-worthy for at least two times (not necessarily consecutive) and "teacher emeritus" for those evaluated for at least three. The title of "Emeritus" is thus incremental

over the title of “expert”, not an alternative. The titles of “expert” and “emeritus” can be maintained for life (just for prestige and independently from the economic reward).

#### *A career beyond teaching aimed at the middle management*

The term “a career beyond teaching” refers to alternative or additional activities to teaching (vice-principal, coordinators, trainers, tutors, etc.). The Foundations recommend that these middle managers, which play a key role in the effectiveness and efficiency of schools, should be chosen by the principal from among those teachers awarded through the Valorizza procedure (who have, in addition, the advantage of having been validated by the stakeholders).

Of course, teachers can be recognized as merit-worthy for different reasons: some of them are appreciated for the quality of their teaching, regardless of any organizational capability, while others are appreciated only for this ability. Restricting the choice of principals to those awarded must be accompanied by the freedom of decision of the principal in appreciating the nature of the tasks that are best suited to the characteristics of each.

In this way the results of the Valorizza evaluation would be used to structure a career for specific functions beyond teaching. The fact that access to middle management roles is connected to a positive evaluation of their work as teachers, would be a strong incentive for many teachers to constantly improve their professional behavior.

#### *Concluding considerations*

Through

- a) higher remuneration for the merit-worthy (with the three year salary bonus)
- b) greater professional prestige (with the life-long titles of “expert” and “emeritus”)
- c) greater possibilities of career development for middle management roles, with extra salary.

the profession would be rendered more attractive to bright young graduates, who are an essential element of high quality schooling.

#### ***c. For a gradual diffusion of the model throughout the educational system***

To further improve the Valorizza model through experience and also to reduce foreseeable union resistance, the Foundations suggest a path of gradual diffusion and a progressive method: the Ministry should not make it obligatory for all schools to adopt the Valorizza method, but make it possible (and finance it from year to year) only for those schools who want to voluntarily take part.

In this way Valorizza would spread with a bottom-up process (on the initiative of each individual school) rather than top-down, i.e. through a decision of the Ministry. It would develop probably quite slowly, gradually, but after the start, thanks to the publicity that the Ministry will decide to give to the process and the resources that will be allocated each year to the project, it may spread “through imitation”, even faster than expected *on the basis of free and informed choices by schools*.

The Ministry, given the uncertainty about the number of new schools that would actually choose to adopt the Valorizza model, could provide, each year, a specific “Fund for the Promotion of Valorizza” with a maximum amount available to ensure the certainty of spending and, year on year, decide the amount of the fund based on the suc-

cess of the method (measured by the number of participating schools) and the level of available resources. Participation could be on a first come first served basis (without prejudice to the priority of those who had adopted it in the previous rounds). The Ministry, according to the results, could review the size of the fund annually to strengthen it, and extend it to other schools in addition to those already involved, and decide the pace of its development throughout the educational system.

## 10. Is the proposal coherent with the strategic objective of teacher evaluation?

The proposal formulated in the preceding pages tries to answer - from the research data currently available – to the strategic objectives of the model. In Chapter 5, reference was made to five of them: they are resumed here, adding a comment to each of them that highlights how an evaluation with the features of this proposal is likely to facilitate their achievement.

1. To link a financial reward to a merit recognition and not only to years of service;
  - *the progression only through seniority is one of the greatest criticisms of the Italian system. It takes all the enthusiasm and ambition out of young graduates and teachers, since they can not in any way overtake those who are "in line" before them by right of seniority. Why work harder? If, however, without completely abolishing the recognition of years of service, those who have the ability and the desire to progress are able to do so, based on the quality of their work, this would bring into the system many individual energies that are today discouraged and blocked by a system which does not recognize merit;*
2. set in motion a dynamics of positive emulation among teachers, that broadens the area of professional excellence;
  - *this is a prospect that does not require much explanation: the drive that moves people to emerge is not just economic. Indeed, in contexts with high professionalism, it consists heavily of a desire for status and visibility. The possibility of being recognized by your entire work environment as one of the most outstanding teachers of the school would motivate not only the excellent, but also many of the 'normal' teachers. It is foreseeable that, if 20% were to be selected each time, at least as many, if not more than that, would aspire to be. And this would motivate from 40 to 50 percent of the profession towards self-improvement;*
3. to identify the most esteemed personalities in each school for promoting the development of the middle management, essential for a desirable shared leadership;
  - *in the long term, an evaluation system based on the individual school would allow for the formation of a reservoir of professional resources, capable of further development, from which middle management could be drawn. The school can not function effectively by concentrating all development and organizational roles in the hands of the principal: there is a strong need for middle management. In this same reservoir, through natural selection, and then a selec-*

*tion from above, the senior figures should be chosen: the principals and inspectors.*

4. attract, over time, graduates of high quality to the teaching profession, with the prospect of better wages and career development;
  - *one of the characteristics of successful school systems is the ability to attract the most capable and brilliant young graduates, not only for the economic benefits but for the prestige surrounding the teaching profession. Career opportunities and a reward system, that are both conferred to those who receive a wide appreciation by their peers and their environment, could in time prove to be the drive that enables schools to achieve the goal of attracting the best graduates;*
5. encourage all teachers to develop the practice of self-evaluation, as a prerequisite for the general improvement of their performance;
  - *self-assessment in some ways precedes the decision to candidate: people do not voluntarily candidate if they do not think they have at least a reasonable chance of being selected. But, for this exercise of self-analysis to begin, there must be an external opportunity, which could be that of a permanent reputation-based evaluation in the school. If this were true, most or all teachers would be led - at least once every three years - to reflect on themselves and their own way of working, in light of the decision whether to participate or not in the selection;*

Of course, a system that evaluates and rewards the most deserving teachers has a cost, but the proposed system has the advantage of allowing the annual definition of the costs by the Ministry. If all schools and all teachers were to participate – and 20% of the candidates received a bonus of 15% of their annual salary (the two extra months allowance) – the total additional cost per year would not exceed 3% of the general budget for teachers wages: a minor cost compared to the benefits that could be achieved.

At present, since general elections just happened in Italy, it is impossible to say what the future of Valorizza will be. It will depend on the decision of the new Minister of Education, that are still unknown.



"THIS TEXT WAS WRITTEN BY THE TWO FOUNDATIONS LISTED BELOW WHO RETAIN ALL RIGHTS OVER AND RESPONSABILITY FOR ITS CONTENT". THE BOOKLET HAS BEEN DISTRIBUTED FOR THE OCCASION OF THE **INTERNATIONAL TEACHER SUMMIT ON TEACHING PROFESSION** (AMSTERDAM 13–14 MARCH 2013)

The **Associazione TreeLLLe** – for LifeLong Learning – has the improvement of the quality of education as its objective. It is a think tank that through research, analysis, design and the publication of results offers a service to public opinion, to social forces, educational institutions and policy makers on a national and local level. In collaboration with international experts TreeLLLe plays a role of monitoring educational systems and the innovations arising in other countries. TreeLLLe is a non-profit organisation without links to government or political parties. The president is Attilio Oliva.

**Fondazione per la Scuola** working with schools and for schools, promoting a better quality education: this is the mission of Fondazione per la Scuola, one of the operating bodies of Compagnia di San Paolo. For a better future, it is central to allow quantity and quality growth, investing in knowledge, education and training. For these reasons, 10 years ago Fondazione per la Scuola was established, as an operating structure that develops and sustains projects. Receivers are foremost school institutions, thanks to a consistent collaboration with headmasters and teachers, involved and motivated in creating a quality education system, Fondazione per la Scuola is effortlessly in service of schools. The Foundation promotes projects in supporting school autonomy; learning and classroom work; Scientific Education; Civic Education for a European perspective. The president is Anna Maria Poggi.