A Performance-Based Evaluation Model for Rewarding Merit in Italian Schools

Donatella Poliandri, ricercatrice INVALSI, donatella.poliandri@invalsi.it
Paola Muzzioli, ricercatrice INVALSI, paola.muzzioli@invalsi.it
Isabella Quadrelli, ricercatrice INVALSI, isabella.quadrelli@invalsi.it
Sara Romiti, ricercatrice INVALSI, sara.romiti@invalsi.it

Abstract

The pilot study *Evaluation and Development of Quality in schools* (VSQ) aims at introducing a performance measurement system of schools that integrates two different perspectives: merit rewarding and school improvement. Starting from Sen's concepts of merit and meritocracy, a construct of school quality was developed which included incentives for the consequences of actions (students' results at SAT) and action propriety, that is "positive" actions that are considered good in themselves regardless of the results they produce. Evaluation tools were designed to include both such perspectives: value added, that is a measure of the extent to which the educational experience enhances the knowledge, abilities and skills of students, and Scoring rubrics that are tools designed to evaluate the quality of positive actions carried out by schools in four areas: Inclusion, Enrichment and Remediation, Evaluation, Orientation. Both these aspects have contributed to the construction of a final score used for ranking schools and for granting rewards to the schools with higher performance rates. Schools also receive financial support for implementing improvement plans: the amount of such support is higher for schools that were in the bottom distribution of the ranking. Results show that both quantitative and qualitative instruments (added value and Scoring rubrics) contributed to the evaluation of the educational process through the exploration of different aspects. Better performing schools showed a significant positive association between students' evaluation policies and the value added of Italian and Math. Thus, the provision at a school level of a shared and structured system of student evaluation seems to promote students' achievement of key competences.

Tag words: 1) Performance-based Evaluation 2) Mixed Methods Research 3) Scoring rubrics

1. Aims of the study and theoretical framework

The aim of this work is to present the pilot study *Evaluation and Development of Quality in schools* (VSQ), a research project carried out by the Italian Ministry of Education (MIUR) together with the National Institute for the Educational Evaluation of Instruction and Training (INVALSI) and the Sapienza University of Rome. This study, for the first time in Italy, tries to implement a reward-based evaluation system in schools and, provides, at the same time, support for school improvement. VSQ aims at introducing a performance measurement system of schools with the following objectives: to design an autonomous, clear, reliable and shared model for school evaluation; to acknowledge excellence and promote support for school improvement. The project, thus, includes and integrates two different perspectives: merit rewarding and school improvement. Top performing schools receive financial rewards; low performing schools will be supported to start an improvement plan.

This work focuses on the qualitative research tools designed for performance evaluation of schools, and their use within an evaluation programme based on economic incentives and improvement actions.

The theoretical framework is based on Sen's concepts of merit and meritocracy. In *Merit and Justice* (2000) Sen considers meritocracy as an extension "of a general system of rewarding merit and elements of such a system clearly have been present in one form or another throughout human history" (p. 8). There are at least two ways of considering merit and systems of rewarding it:

1. Incentives: actions are rewarded for the good they provide, and a system of remunerating the activities that generates good consequences would tend to produce a better society. The idea of merit in this instrumental perspective relates to the motivation of producing better results. In this view, actions are meritorious in a

2. Action propriety: actions are judged by their propriety, not by their results, and they may be rewarded according to the quality of such actions, judged in a result-independent way. So, actions are considered good for themselves, regardless of the goodness of their consequences.

In one form or another both these views have been invoked in discussions about merit and systems of rewarding merit; but it is fair to say that the incentive approach is dominant today in economics, at least in theory (even though the language used in practice often betrays interest for the other categories). Although the praiseworthiness of "proper action" is not denied in economic reasoning, the economic justification of rewarding merit tends to be grounded in consequences. According to Sen, we can scarcely dispense with rewarding good or right actions, at least for their incentive effect. "The art of developing an incentive system lies in delineating the content of merit in such a way that it helps to generate valued consequences" (Sen 2000. p. 9). For this purpose, a system that rewards merit does not necessarily produce inequality; in fact, equality or equal opportunities could be the leading criteria to define merit.

However, the idea of merit has not been researched extensively: we are not sure about its content until we make some specifications, in particular about the objectives to be pursued in a society in term of merit. In sum, merit can be defined only through the procedures we use to evaluate it.

In the light of the foregoing considerations, we have investigated the possible ways of defining, within a system of rewarding merit, a concept of quality for schools. The research questions were: are students' achievements at SAT sufficient to define the quality of a school? Which criteria determine whether a school deserve a reward or not?

After answering negatively to the first question, we have tried to define a multi-criteria concept of quality. This concept integrates the two different perspectives on merit suggested by Sen: incentives and action property. Thus, the idea of quality is based both on the measurement of students' results at SAT and on the evaluation of actions carried out by schools considered to be relevant *per se*; positive and pro-active actions that can be assessed to be good as they are socially shared and valued by a social group, independently from their impact on results.

Both these aspects have contributed to the construction of a final index used for ranking schools and for granting rewards to the schools with higher performance rates.

2. Method and data survey

Qualitative and quantitative instruments have been used to define the quality's construct.

Specifically, we used standardized instruments to assess the student's learning, such as reading understanding and grammar and math tests administered by INVALSI

The results from the INVALSI tests were used to construct a measure of added value. The added value represents the "measure of quality in terms of the extent to which the educational experience enhances the knowledge, abilities and skills of students" (Harvey and Green, 1993).

Robust evidence shows that students' learning depends on students' characteristics and schools' environment. This means that to improve students' learning we should consider different variables, such as: students' previous learning level and individual characteristics (e.g. gender, socio-economic family status, or nationality), the structural aspects of the school (e.g. teachers' mean age, number of school's buildings, or teachers' turn-over), and environment's characteristics (e.g. number of foreign students, location of the school).

Value added models allows to compare schools from the same starting point, that is not considering the differences in students' achievement due to these individual, structural and contextual variables. In other words, in added value models the variables which are not under the direct control of schools are not considered.

Evaluation of schools' positive actions has been done by Evaluation Teams that visited each school and assessed the quality level in several areas: Inclusion (inclusion of students with disabilities, inclusion of foreign students); Enrichment and Remediation actions; Evaluation (planning and students' evaluation, internal evaluation/self-evaluation) and Orientation.

The instrument used to evaluate schools' positive actions is the *Scoring rubric*. Scoring rubrics are widely used for the authentic assessment of students' performances (Comoglio, 2002; Comoglio, 2007; Wiggins, 1996); during the years its use has been extended to other contexts, such as the evaluation of schools' performance (e.g. grids used by Ofsted's Inspectors in the United Kingdoms).

A rubric contains evaluation criteria and, for each of them, a description of the good and less good performances related to different levels of quality. Rubric can be designed with a more holistic approach when a general judgment is required for several aspects, or with a more analytic one when a judgment is required on a single aspect.

Aim of Scoring rubrics is to help Evaluation Teams to make a judgement based on empirical evidence.

Quality criteria or standards are defined as propositions. Each quality criteria is articulated through a "multifocus" perspective: that is the same quality criteria is assessed considering several dimensions; from each dimension indicators, based on precise elements or empirical evidence, are selected. For example, in the Rubric assessing Remediation actions are identified four dimensions:

- needs analysis: in this dimension are considered how schools collect information about or monitor educational and training needs of students in order to implement specific actions (e.g. how students who need Remediation classes are identified?);
- organizational structure: this dimension is related to organizational aspects of the intervention (e.g. Is there a teacher leader for remediation actions? How students access remediation classes?);
- teaching: refers to the specific teaching strategies and actions activated (e.g. assigning students to homogeneous groups by level, setting different assessments for different levels);
- satisfaction / effectiveness: considers students' opinion about remediation classes and the impact remediation classes attending on students' results (e.g. if students consider useful the remediation classes attended, if students who attended remediation classes have improved their results during the school year).

For each quality criteria a four step evaluation scale is defined; each step corresponds to a level of judgement (inadequate, acceptable, good, excellent). Each level is described in a detailed way; the description also contains all the elements or the empirical evidences that is necessary to consider for expressing the judgement (Rubric score).

Seven Scoring rubrics were constructed in order to evaluate the four areas of positive actions.

Table 1

The scoring rubrics

AREA	SCORING RUBRIC
Inclusion	Inclusion of students with disabilities Inclusion of foreign students
Enrichment & Remediation	Enrichment Remediation
Evaluation	Planning and students' evaluation Internal evaluation / self-evaluation
Orientation	Orientation

In order to support the Evaluation Team in expressing the judgement (Rubric score), for each Scoring rubric it was provided a checklist that contains a list of indicators that were considered relevant and that needed to be observed for reaching a good level of quality. Before expressing the Rubric score, teams' components filled individually the checklist and then shared it within the group in order to negotiate a common starting point (that is, a consensus within the group about the indicators that have been observed).

School visits have been conducted by Evaluation Teams, made up of three evaluators, each of them with a specific qualification and professional background (the coordinator was a Ministerial inspector, one component had a teaching background and the other had a research background).

Evaluators collected documents and other materials such as official school documents, projects, individual education plans, reports, meeting minutes, etc., and conducted interviews (with Principal, coordinators, students, teachers and parents). Research protocols included guidelines about people to interview for each area of evaluation and advice about relevant questions to ask. Evidence from document analysis and interviewees' opinions constitute the empirical basis for expressing the judgement.

The final score – that defined each school's position in the final ranking and the possibility to obtain the economic reward – was made up from the added value score[1] and the Mean rubric score (a synthetic score resulting from the mean of the individual scores assigned to each rubric). These two scores have different weights in the calculation of the Final score:

- the weight of the value added score is 60%: 35% for Italian and 25% for Math. The higher weight given to Italian arises from the relation between competences in Italian

[1] The Added Value Model utilized in this Project is implemented by Patrizia Falzetti and Roberto Ricci (INVALSI).

and competences in other subjects, for example the link between scientific processes comprehension and Italian literacy.

- the weight of the Mean rubric score is 40%.

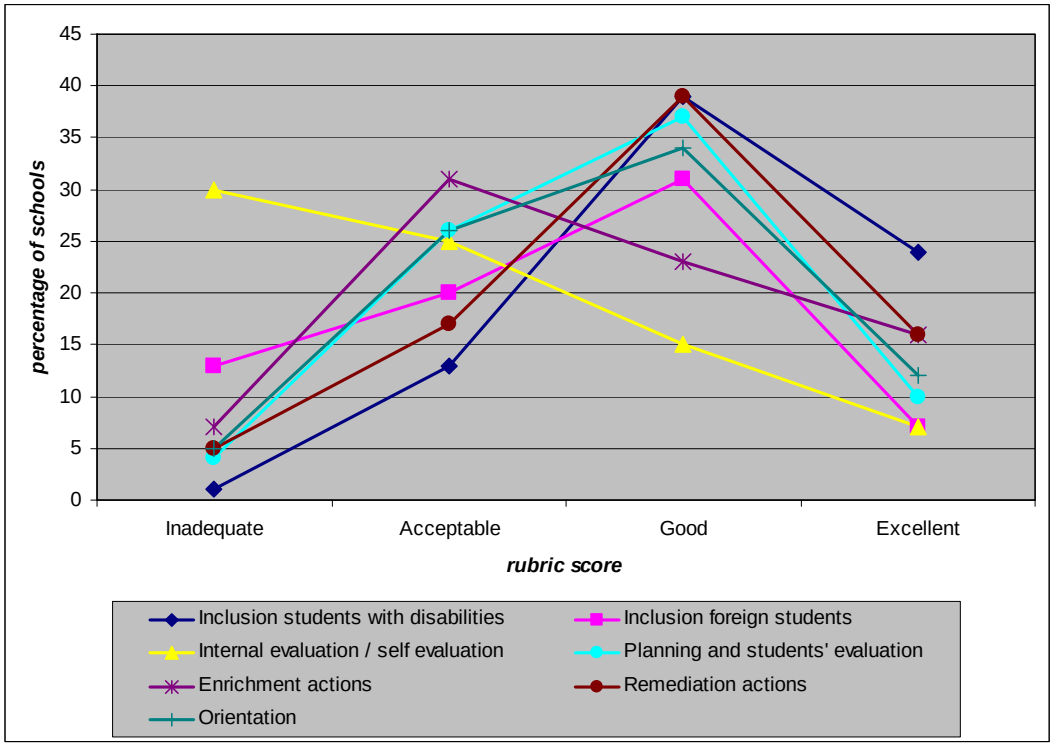
3. Results and discussion

77 primary/middle schools from 4 different provinces (Pavia, Mantova, Arezzo, and Siracusa) in Italy have participated in this study.

A first result shows that for most rubrics, the majority of schools have received the score “good” (the third level).

Graph 1

Percentage distribution of schools by Rubric score



Internal evaluation / self evaluation shows a different trend, the percentage of schools decreases with increasing score, so most schools have lower score, 40% was assessed as inadequate (the first level), only 10% obtains an excellent score (the fourth level).

With regard to the area with the higher level of quality, only 1% of schools receive inadequate score for activities related to the inclusion of students with disabilities.

Overall Mean rubric score is 2,63 with a standard deviation of 0,60, this confirm that most schools have received acceptable and good scores.

In order to analyze if the number of indicators included in the final checklist filled by the Team is related to the rubric score assigned, correlation between the percentage of filling out and score has been calculated.

Table 2

Pearson correlation between Rubric score and percentage of checklist filling out

	Rubric score
Inclusion students with disabilities	,796**
Inclusion foreign students	,770**
Internal evaluation/self-evaluation	,734**
Planning and students' evaluation	,752**
Enrichment actions	,775**
Remediation actions	,753**
Orientation	,802**

**. Correlation is significant at the 0,01 level (2-tailed).

We supposed that a high number of checks (that is high number of indicators observed by the Team) could be a good indicator of the total quality reached by the school in the specific area investigated by the rubric. Correlations are quite high for all Scoring rubrics; this confirms the fact that the higher the number of indicators included in the checklist, the higher the level of quality reached by a school in a specific area.

The distribution of schools by the value added levels of Italian and Maths show that the majority of schools are in the medium level.

Table 3

Numerical e percentage distribution of schools by value added levels

Added value levels	Added value ITALIAN	%	Added value MATH	%
Low	19	24,66	23	29,87
Medium-low	2	2,60	3	3,90
Medium	38	49,35	35	45,45
Medium-high	4	5,19	3	3,90
High	14	18,18	13	16,88
Total schools	77	100	77	100

Almost 50% of schools have an intermediate level of added value both in Italian and Maths, the number of schools with lower levels of value added is higher than schools with higher ones.

Correlation between Rubric scores and value added confirm that both quantitative and qualitative instruments (added value and Scoring rubrics) contributed to the evaluation of the educational process through the exploration of different aspects.

In fact, we found positive and low significant associations between the value added and the evaluation Rubrics scores, in the range of .10 to .16.

Table 4

Pearson correlation between Added value, Mean rubric score and Final score

	MEAN rubric score	Added value ITALIAN	Added value MATH	Final score
MEAN rubrics score	1			
Added value ITALIAN	,10	1		
Added value MATH	,16	,61**	1	
Final score	,51**	,85**	,81**	1

**. Correlation is significant at the 0,01 level (2-tailed)

A positive relation means that the two areas move in the same direction, while the low association indicates that the two areas are not measuring the same aspects. This means that the two instruments give different but complementary information about schools.

Correlations showed a significant positive association between students' evaluation policies and the value added of Italian and Math. Thus, the provision at a school level of a shared and structured system of student evaluation seems to promote students' achievement of key competences.

In addition, we found a positive significant relation between value added scores in Math and effective Enrichment actions.

Table 5

Pearson correlation between Rubric score and Added value

	Inclusion students with disabilities	Inclusion foreign students	Internal evaluatation/self-evaluation	Planning and students' evaluation	Enrichment actions	Remediation actions	Orientation	Added value ITALIAN	Added value MATH
Inclusion students with disabilities	1								
Inclusion foreign students	,12	1							
Internal evaluation/self-evaluation	,45**	,24*	1						
Planning and students' evaluation	,61**	,31**	,58**	1					
Enrichment actions	,49**	,20	,42**	,58**	1				
Remediation actions	,43**	,40**	,31**	,56**	,61**	1			
Orientation	,33**	,37**	,41**	,52**	,39**	,42**	1		
Added value ITALIAN	-,10	,16	-,06	,21	,16	,14	,03	1	
Added value MATH	,06	,16	,01	,23*	,25*	,11	-,01	,61**	1

**. Correlation is significant at the 0,01 level (2-tailed).

*. Correlation is significant at the 0,05 level (2-tailed).

The Teams' scores among the different observed aspects are highly interrelated. Specifically, our results showed a significant association between the Student evaluation

system, Enrichment and Remediation actions. In other words, these results suggest the importance of a good student evaluation system in order to plan effective individualisation interventions (enrichment or remediation).

Schools that performed better were rewarded with money token; regional ranking have been produced and 25% of schools placed in the top distribution of the ranking have been rewarded. Overall, 20 schools received a money incentive.

All participating schools have received an individual Evaluation report that highlights strengths and difficulties in the areas under investigation and contains suggestions about some improvement actions that each school may decide to implement. Schools also receive financial support for implementing improvement plans: the amount of such support is higher for schools that were in the bottom distribution of the ranking.

In 2013, participating schools will be evaluated again, using the same instruments (value added and Scoring rubrics) in order to assess the effect of the competitive mechanism and of the improvement plan on school performance, and to distribute the final part of the money incentive.

Conclusion

To conclude, within the scope of the VSQ project it has been possible to design qualitative research tools starting from a construct of merit defined by evaluable criteria. As a consequence the idea of merit was also made explicit and clear to those involved in the research, including schools. Moreover, the adoption of a multi-criteria approach has allowed assessing schools not by a sole performance indicator (students' results at SAT) but to construct a more complex index of quality that integrates the two rewarding systems suggested by Sen, that is results and propriety actions. However, more effort is still needed

both to construct a really shared and acknowledged idea of quality for schools and to improve qualitative and quantitative instruments for measuring it.

From the point of view of the programme evaluation design, VSQ aimed at integrating a reward system and a positive discrimination system for school evaluation, in order to mediate a competitive approach with one of equal opportunity. At the empirical level, it is necessary to assess the impact of a rewarding system on the promotion of a real, generalised tension to improvement generated in schools by the competitive mechanism.

References

Amrein-Beardsley, A.. "Methodological Concerns About the Education Value-Added Assessment System". *Educational Researcher* 37, 2 (2008): 65-75.

Arrow, K., Bowels, S., Durlauf, S.. *Meritocracy and economic inequality*. Princeton: University Press, 2000.

Bondioli A., Ferrari M.. ed. *Manuale di valutazione del contesto educativo: teorie, modelli, studi per la rilevazione della qualità nella scuola.* Milano: F. Angeli, 2000.

Bosker, R., Witziers, R.. "School Effects: Problems solutions and a meta analysis" (paper presented at the Eighth Annual International Congress for School Effectiveness and Improvement, CHN, Leeuwarden, The Netherlands, January, 1995).

Bottani, N., Cenerini, A.. ed. *Una pagella per la scuola. La valutazione tra autonomia e equità*. Trento: Erikson, 2003.

Cissel, G.. "Kentucky and Education Reform: The Issue of Pay-for-Performance". *Journal Of Law & Education* 39, 1 (2010): 119-127.

Comoglio, M.. "La valutazione autentica". *Orientamenti Pedagogici* 49, 1 (2002): 93-112.

Duru-Bellatt, M.. *L'inflation scolaire. Les disillusion de la mèritocratie*. Paris: Edition du Seuil, 2006.

Fukuyama, F.. *Trust: the social virtues and the creation of prosperity*. New York: The Free Press, 1996.

Goldstein, H.. "Using Pupil Performance Data for Judging Schools and Teachers: scope and limitations". *British Educational Research Journal*, 27, 4 (2001): 391-405.

Martini A., Ricci R.. "Un esperimento di misurazione del valore aggiunto delle scuole sulla base dei dati PISA 2006 del Veneto". *Rivista di Economia e Statistica del territorio*, 3 (2010), 80-107.

Mc Namee, S., J., Miller, R., K.. *The Meritocracy Myth*. Lanham: Rowan & Lietfiled, 2004.

Moss P.. “Defining Quality: Values, Stakeholders and Processes”. In P. Moss & A. Pence, *Valuing Quality in Early Childhood Services*. Londra: Paul Chapman Publishing Ltd, 1994.

OECD. “Establishing a Framework for Evalutation and Teacher Incentives. Considerations for Mexico”. OECD Publishing, 2011.

Raudenbush, S., W., Bryk, A., S.. *Hierarchical Linear Models. Applications and Data Analysis Methods, Second Edition*. Newbury Park CA: Sage Publications, 2002.

Sanders, W., L., Horn, S., P.. *The Tennessee value-added assessment system (TVAAS): mixed methodology in educational assessment*. In Shinkfield, A., J., Stufflebeam, D.,L. ed., *Teacher evaluation: guide to effective practice*. Boston: Kluwer Academic Publishers (1995): 337-376.

Sen, A.. *Merit and Justice*. In Arrow., K., Bowels, S., Durlauf, S.,. *Meritocracy and economic inequality*. Princeton: University Press, 2000.

Teese, R., Lamb, S., Duru-Bellat, M.. “International Studies in Educational inequality, Theory and Policy”. In *Volume One. Educational Inequality: Persistence and Change*. Dordrecht: Springer, 2007.

U.S. Department of Education. “No Child Left Behind. Evaluation of the Comprehensive School Reform Program Implementation and Outcomes”. Third-Year Report, 2008.

Wiggins, G.. “What is a rubric? A dialogue on design and use”. In Blum, R., E., Arter, J., A. ed. *A handbook of student performance assessment in an era of reconstructing*. Alexandria Va : Association for Supervision and Curriculum Development, 1996.

Young, M.. *The Rise of Meritocracy, 1870-2033: An Essay on Education and Equality*. London: Thames and Hudson, 1958.