This article was downloaded by: *[Porro, Giuseppe]*
On: *7 September 2009*
Access details: *Access Details: [subscription number 914467149]*
Publisher *Routledge*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Education Economics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713415403

Teachers' evaluations and students' achievement: a 'deviation from the reference' analysis

Stefano M. Iacus [a]; Giuseppe Porro [b]
[a] Department of Economics, Business and Statistics, University of Milan, 20122 Milan, Italy [b] Department of Economics and Statistics, University of Trieste, 34127 Trieste, Italy

First Published on: 04 September 2009

To cite this Article Iacus, Stefano M. and Porro, Giuseppe(2009)'Teachers' evaluations and students' achievement: a 'deviation from the reference' analysis',Education Economics,99999:1,
To link to this Article: DOI: 10.1080/09645290903105277
URL: http://dx.doi.org/10.1080/09645290903105277

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Education Economics
2009, 1–21, iFirst Article

Routledge
Taylor & Francis Group

Teachers' evaluations and students' achievement: a 'deviation from the reference' analysis

Stefano M. Iacus[a] and Giuseppe Porro[b]*

[a]*Department of Economics, Business and Statistics, University of Milan, Via Conservatorio 7, 20122 Milan, Italy;* [b]*Department of Economics and Statistics, University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy*

Several studies show that teachers make use of grading practices to affect students' effort and achievement. Generally linearity is assumed in the grading equation, while it is everyone's experience that grading practices are frequently non-linear. Representing grading practices as linear can be misleading both from a descriptive and a prescriptive viewpoint. Here we propose to identify grading practices as 'deviations from a reference', which is a fully non-parametric criterion, and measure their effects on achievement based on this classification. To show the effectiveness of our approach, we apply the methodology to a data-set on Italian lower secondary school.

Keywords: evaluation; grading practice; students' achievement; classification techniques

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

1. Introduction

Strong evidence has been provided in the economic literature about the positive relationship between education and earnings (Card 1999). A large debate is open, instead, on the role of quantity and quality of schooling in determining the education results, in terms of students' achievement, job opportunities, and earnings: is it worth investing more resources in schooling, with the aim to improve the quality of the education process? To what extent are these resources a substitute for an additional year of school? Hanushek (2002, 2003), for example, criticizes schooling policies simply based on input increasing, concluding that the increase in the cost of education does not yield remarkable improvement in students' achievement and, consequently, in their earnings perspectives. Krueger (2002, 2003), on the other hand, analyzes the same empirical evidence, drawing the opposite conclusion. Interestingly, Hanushek (2003) indicates, as a fallacy of the studies on the impact of schooling policies, the lack of attention to the incentives mechanisms within schools: when little feedback to good performance is provided, it should not be surprising that added resources do not yield significant improvement to achievement. One of the devices that can affect the achievement of the students, given the cost of the education policy, is the grading procedures chosen by the teachers. In slightly different terms, the criterion used by the teacher to assign grades conditional to the observed achievement may be considered

*Corresponding author. Email: giuseppe.porro@econ.units.it

ISSN 0964-5292 print/ISSN 1469-5782 online
© 2009 Taylor & Francis
DOI: 10.1080/09645290903105277
http://www.informaworld.com

a (fairly) costless input of the education process, and greater attention should be devoted to choosing proper grading practices.

Moreover, an erroneous distribution of incentives (both to teachers and students) in designing grading standards is at the basis of the widespread and worrisome phenomenon of *grade inflation* (see, for example, Johnes 2004). As far as school grades are considered, a reliable indicator of students' ability by the labor market and, at the same time, an indicator of teachers' quality by the school managing authorities, a well-known free-rider problem, can be observed (Marks 2002): students put pressure on teachers to get good marks with low effort, and teachers have an incentive to assign good marks to low achievement in order to improve their quality signal. The result is a progressive fall of grading standards, known as grade inflation. As a consequence, grades can no longer be assumed as indicators of performance of the education institutions, and they miss a great part of their potential role in sorting more able and less able students and in signaling their actual achievement.

The issue is discussed by an increasing stream of papers, which suggests that the strategic use of grading practices by teachers may affect students' effort and achievement at any level of the education process. Through grade assignment, in fact, teachers can do much more than simply register the knowledge level achieved by their students: they can give a reward to student improvement, encourage further effort, stimulate students' potential ability, and punish indolence. In particular, according to several empirical studies, if a teacher makes use of high grade standards (i.e. gives good marks for high performance only), his/her students may achieve, *ceteris paribus*, higher knowledge.

In empirical studies, researchers measure the stringency of grading standards assuming a linear *grading equation*; that is, a linear relationship between achievement (usually measured by a test score) and grades, and estimating the two parameters (intercept and slope) of this function. The estimated coefficients should characterize the stringency of the grading practice. But frequently, one of the two estimated parameters results in non-statistical significance: hence, one is ignored and the other is used to univocally identify and order the grading practices. The coefficients of the linear grading equation are included as regressors into an *education production function* (EPF) and are used to estimate the impact of grading standards on achievement.

It is common-sense that in order to be able to evaluate the effect of grading practices on students achievements it is, at first, necessary to identify these practices. Surely, linearity and ordering are convenient shortcuts at an explanatory level, but unfortunately, the actual relationship between achievement and grades may be non-linear: sometimes teachers show a preference for using extreme grades or, conversely, median grades only; sometimes they over-reward (under-reward) high performances so that their grading profiles show increasing (decreasing) growth rates, and so on. In these cases, imposing a linear relationship between achievement and grades may neglect these aspects of the grading scheme and may lead to incorrect evaluation of the grading standards.

This paper proposes a new approach based on the definition of a *reference grading practice* and an indicator of *deviation from this reference*: the aim is to properly capture, in a fully non-linear and non-parametric way, all kinds of grading strategies. The indicator of the 'Deviation from the Reference' (DfR) can be further included into the EPF and used to evaluate the effects of the different grading practices on the achievement of the students. The indicator can be used to implement education

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

policies aimed at controlling grade inflation and inducing compliance towards a specific grading standard.

The paper is structured as follows. Section 2 reviews the instruments used by the economic literature to measure the effects of grading practice on student achievements. Section 3 motivates the DfR approach. Section 4 introduces a data-set used to discuss the fallacies of the assumption of linearity on grading practices both at class and school levels. Section 5 explains how to construct the DfR indicator in details using real data. Section 6 contains the empirical analysis of the data introduced in Section 4. The results confirm that, in line with the literature, higher grading standards favor students' performance, but also show that grading practice cannot be ordered. Compared with the linear approach, the DfR indicator is able to capture the variety of grading practices properly and their effects on achievement. Some policy suggestions are proposed in Section 7, and possible limitations of the DfR approach are discussed in the last section.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

2. Grading equations and student achievements: a review

Several empirical studies about the effect of grading practices on students' achievement find their theoretical background in Correa and Gruver (1987) and Costrell (1994). Both of these models provide justifications for higher grade standards affecting students' effort and achievement.

All of the empirical studies descending from these frameworks adopt a *linear grading equation*,

$$grade = \beta_0 + \beta_1 achievement,$$

to characterize the grading standards: in particular, estimating the parameters of the grading equation, Bonesronning (1999, 2004a, 2004b, 2008) finds that the slope is seldom significant, while the intercept is.

The estimated intercept $\hat{\beta}_0$ of the linear grading equation is usually included as a regressor into the EPF, and shows the positive effect of higher grading standards on students' achievement.

The effect of higher standards on average achievement and their potential distributional consequences are examined by Betts (1997, 1998) and Betts and Grogger (2003). These authors indicate that higher standards improve average students' performance, affecting the achievement test scores of abler students more than those of the less able. In these papers the stringency of grading standards is measured by a linear regression of the achievement test scores on the school grades,[1] allowing for school fixed effects. The estimates of the fixed effects are included into an education production function as indicators of the grading practices, and show their positive effect.

A positive effect of higher grading standard on students' achievement is also estimated by Figlio and Lucas (2004), using a data-set on elementary school pupils. They also agree with Betts and Grogger (2003) about the existence of considerable distributional effects. Figlio and Lucas (2004) represent a partial exception to the adoption of a linear grading equation. In their paper, in fact, three different measures of grading standards are proposed – and a couple of them are non-parametric and, to some extent, related to the present proposal.

3. The DfR approach

This section introduces an indicator that can be used to describe grading practices, whose application is postponed to Sections 5 and 6.

The idea of DfR is a basic and common concept to describe variability in the social sciences. The best known example is the notion of variance, which measures the average deviance from a reference value (the mean). Variability does not necessarily induce an ordering on behaviors, and this is particularly true for social phenomena (see, for example, the concept of 'deviating behaviour' in criminology).

We deploy the DfR intuition to capture the variability of grading practices. Assume the availability of an *objective* measure of competence (e.g. the score of a multiple choice test) and of a *subjective* one (e.g. the teacher's grades) for each student. The DfR procedure is based on the following three steps:

(i) Define an arbitrary, *ex ante*, relationship that maps scores to grades: for example, it maps the intervals of normalized scores[2] [0.0,0.6), [0.6,0.7), [0.7,0.8), [0.8,0.9), [0.9,1.0] to grades *F,D,C,B,A*, respectively.
This constitutes the 'reference' grading practice, defined in order to evaluate the deviations of all the observed grading standards from it.

(ii) Measure the deviation of teacher's grade from the reference grade for any given student score. Due to the fact that the DfR approach does not provide a single quantity but rather a deviance profile, teachers with similar DfR profiles are grouped into classes that constitutes the items of a categorical variable: the *DfR* variable. The values of the *DfR* variable naturally identify the grading practices.

(iii) The *DfR* variable can be included into an education production function in order to estimate the effect of these grading practices on students' competence:

$$achievement = epf(S, X, Y, DfR)$$

where S, X and Y are, respectively, vectors of student's, teacher's and school's characteristics.

How to apply in practice the steps of the DfR procedure will be illustrated in details in Sections 5 and 6 on empirical data.

4. On the linearity assumption

It is common-sense that the relationship between achievement and grades is hardly linear. At the same time, when non-linear grading practices are aggregated, say, by school, district, and so forth, they seem to be linear – and this is taken as evidence of a linear school policies when, on the contrary, what is observed is merely the effect of averaging.

In this section we present some evidence about both issues, analyzing single teacher grading practices and showing the fallacy of the linearity assumption on empirical cases. As a driving example, we will present a data-set that will be used throughout the rest of the paper.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

4.1 The data-set

The survey has been carried out on a sample of 20 lower secondary schools[3] in Lombardy (Italy) during the period 2003–2005. Three multiple-choice achievement tests (Italian language, mathematics and science) were administered to first-year students in March 2003. Similar tests (for Italian language and mathematics only) were submitted to the same students subsequently (second-year students in May 2004, and third-year students in May 2005). In May 2005 a questionnaire was administered both to students (about their school carrier and school climate) and to teachers (about their professional features and school environment). A longitudinal archive has been set up in 2005, containing the records of 1243 (for Italian language) and 1259 (for mathematics) students, belonging to 77 classes, who took part in all the three waves of the survey. We use the data-set on Italian language in the following application. The same analysis on mathematics has been made and is not reported here, but the results obtained on the Italian language data-set are confirmed.

Every year, the final grade of the students is registered (*ITA*03, *ITA*04). The third-year grade corresponds to the teacher's evaluation at the end of the first semester (*ITA*05*FST*). All grades are coded by *A*, *B*, *C*, *D*, *F*. The normalized test scores are represented by variables *SCITA*03, *SCITA*04, *SCITA*05, respectively, and have been converted into Rasch measures *M*03*ITA*, *M*04*ITA*, *M*05*ITA* to make scores comparable from year to year when used together in the same regression model. This conversion does not affect the definition of *DfR*, which will be discussed in the next section.

The control variables are: *GENDER*, the student's gender; *M*03*ITA*, the initial achievement level; *NCLASS*, the number of students in the class; *NPROFITA*, a discrete variable indicating whether the student changed the teacher once ore more during the lower secondary school; *CHANGECL*, a dummy indicating whether the student changed class during the lower secondary school; and *BOOKS*, a proxy of the family background (number of books at home).

4.2 Evidence of non-linearity

Figure 1 reports the results for 77 lower secondary school classes in the data-set: for each class, the relationship between the scores of achievement test (*SCITA05*) and the school grades (*ITA05FST*) on Italian language in 2005 is shown. For each class we estimate the following linear grading equation:

$$ITA05FST_i = \beta_{0i} + \beta_{1i} SCITA05_i, \qquad i \in C_j, \tag{1}$$

where C_j is the set of indexes of the observations in class j, and $j = 1, \ldots, 77$, and also a non-parametric regression.[4] Both estimated models are represented in Figure 1. As one can notice, in several cases the linear assumption seems to be inadequate to describe the phenomenon: sometimes grades grow more than proportionally with respect to achievement (e.g. Classes 54, 60, 67, 2, 11), sometimes they grow less than proportionally (e.g. Classes 72, 68, 44), and in other classes the grading practice are simply non-monotonic with respect to the achievement scores (e.g. Classes 58, 46, 4). So, the linear assumption of Equation (1) is not satisfied.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

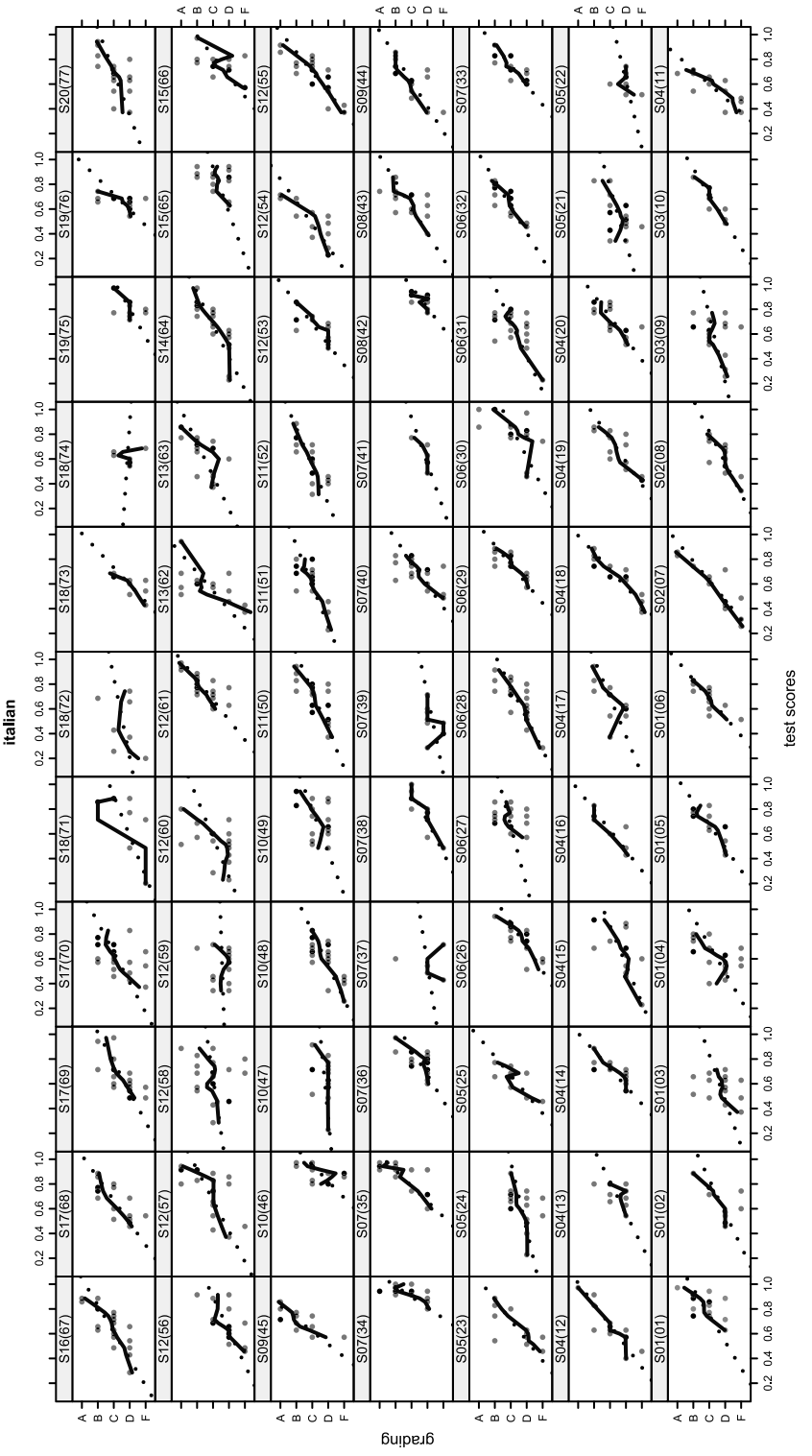

Figure 1. The non-linear behavior of grading practices in the 77 classes.

4.3 Do school-level grading standards really exist?

Here we focus on the second issue: sometimes, when working with aggregated data, linear behaviors may appear. This is used as empirical evidence to justify the existence of a school grading policy. We think that the evidence of such a policy is in general debatable, and sometimes the assumption seems to be motivated more by the availability of the data rather than by true evidence of schools' policies. Dardanoni, Modica, and Pennisi (2009a, 2009b), for example, are forced to assume the existence of a school policy because they are carrying out a comparative study on grading practices based on the OECD PISA 2003 data-set (OECD 2005), which is aggregated at the school level. They motivate the adoption of a linear grading practice, arguing that, if the teachers are constrained by students' perception of unfairness in assigning grades, the grading policy should be linear. But in our data, this assumption is violated in several classes (see again Figure 1): as we noticed, it may even happen that teachers assign lower grades to higher achievement, and *vice versa*.

If we aggregate the 77 classes of our data-set at the school level, the grading behavior we observe is, to some extent, more linear than we observe at the single-class level (see Figure 2). Is this the evidence of a school grading policy or are we simply averaging out the different practices of the teachers belonging to each school?

Let us estimate the linear grading equations at the *school level*:

$$ITA05FST_i = \beta'_{0i} + \beta'_{1i} SCITA05_i, \quad i \in S_k, \tag{2}$$

where S_k is the set of indexes of the observations in class k, and $k = 1, \ldots, 20$. Results are presented in Table 1.

Consider School S07, whose grading equation seems to be not so far from linearity: if the teachers belonging to School S07 were following a school grading policy, we would observe some kind of compliance with this policy and hence their behaviors should be quite similar to each other and to the general standard. On the contrary, what we observe in Figure 3, where the single-class grading equations of School S07 are shown, does not seem to confirm a compliance with a general policy – see, for instance, Classes S07(37) and S07(39) versus Classes S07(34), S07(35) and S07(40).

The consequences of the linearity assumption at the school level might be relevant if we want to provide incentives to schools whose high grading standards – according to the literature – seem to positively affect students' achievement. So, for instance, using a policy based on linear modeling, one would like to penalize School S07 because its slope $\hat{\beta}_1$ is one of the highest (see Table 1) and indicates a low grading standard. But, as previously discussed, the single-teacher standards inside School S07 are quite diversified (see again Figure 3). So, if we decide to penalize School S07 because of its generous grading standard, we are, at the same time, punishing both generous and severe standards, giving a biased incentive to the teachers.

5. Definition of the reference

We now apply the DfR approach. As said, a quick look at Figure 1 shows high heterogeneity in teacher's behavior. Each of these behaviors corresponds to a grading practice that we try to identify properly via the DfR approach. To this aim, we need to introduce the *reference* in this experiment. Define as a reference a class[5] where grades are given according to the arbitrary scale of achievement scores described in Table 2.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

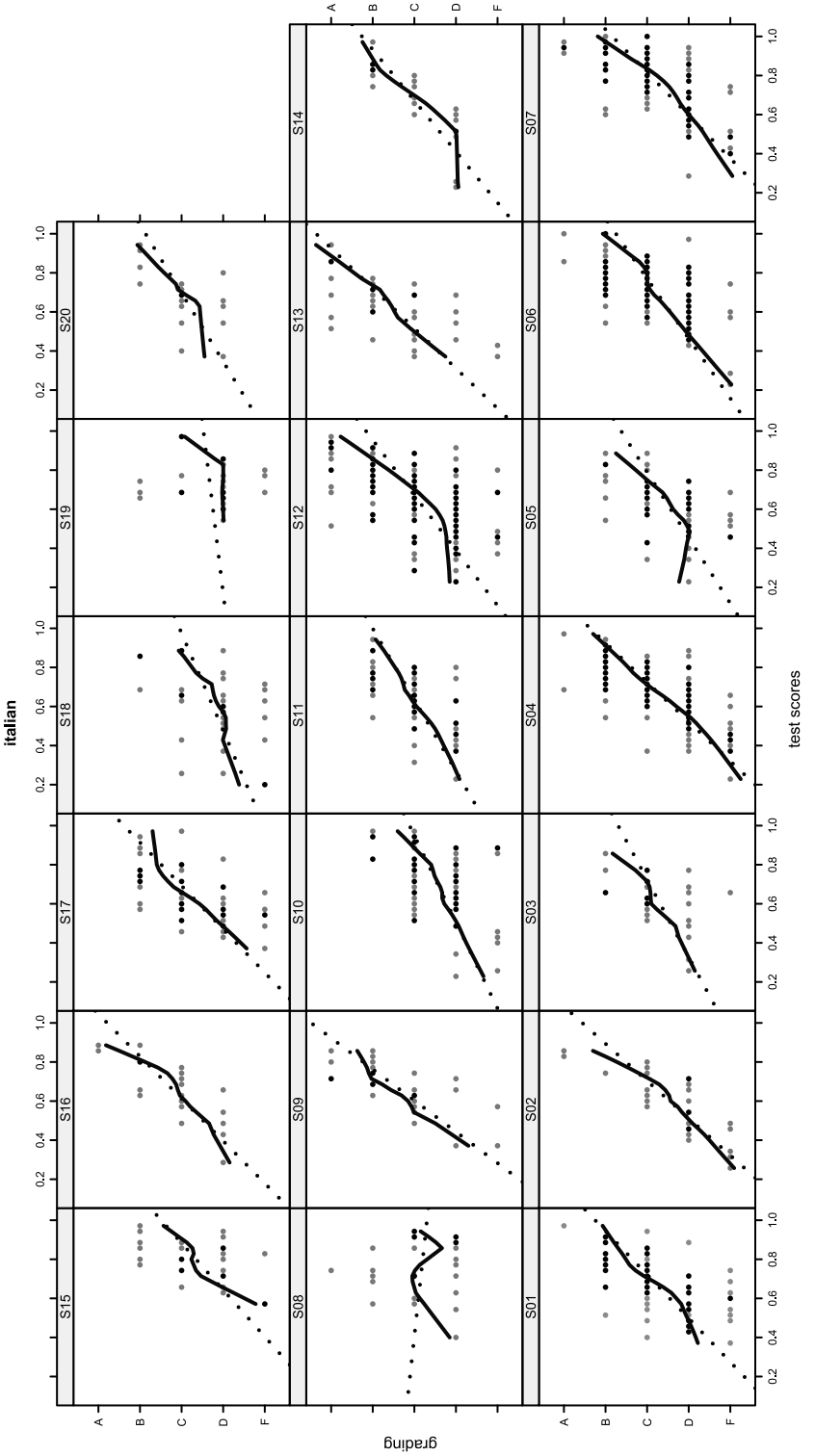

Figure 2. Grading practices aggregated by school.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

Table 1. Linear grading equations estimated at school level.

	School									
	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10
$\hat{\beta}_0$	−0.229	−0.689	1.174	−0.587	0.649	0.485	−0.708*	3.207***	−0.766	0.854*
	(0.405)	(0.488)	(0.623)	(0.302)	(0.421)	(0.366)	(0.333)	(0.854)	(0.828)	(0.360)
$\hat{\beta}_1$	4.478***	5.248***	2.522*	4.955***	2.975***	3.264***	4.544***	−0.510	6.248***	2.262***
	(0.587)	(0.808)	(0.979)	(0.445)	(0.678)	(0.501)	(0.432)	(1.071)	(1.212)	(0.493)
R^2	0.396	0.628	0.210	0.490	0.231	0.254	0.476	0.008	0.505	0.231
Adj. R^2	0.389	0.613	0.178	0.487	0.219	0.248	0.471	−0.026	0.486	0.220
AIC	215.828	59.021	60.730	289.817	149.717	293.062	282.456	81.420	70.236	158.961
n	91	27	27	131	66	127	124	31	28	72
	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20
$\hat{\beta}_0$	1.263***	0.620*	0.482	0.427	−0.688	0.173	−0.116	1.040*	1.903	1.013
	(0.301)	(0.275)	(0.712)	(0.393)	(1.224)	(0.599)	(0.498)	(0.480)	(1.220)	(0.691)
$\hat{\beta}_1$	2.740***	3.543***	4.880***	3.833***	4.178*	4.616***	4.503***	2.011*	0.571	2.854*
	(0.463)	(0.409)	(1.104)	(0.564)	(1.539)	(0.889)	(0.763)	(0.766)	(1.653)	(0.999)
R^2	0.358	0.314	0.387	0.698	0.221	0.600	0.375	0.182	0.005	0.353
Adj. R^2	0.347	0.310	0.367	0.683	0.191	0.577	0.364	0.155	−0.038	0.309
AIC	111.444	430.906	93.146	35.081	74.934	41.435	142.542	85.237	67.815	35.932
n	65	166	33	22	28	20	60	33	25	17

Note: Significance at ***0.001, **0.01, *0.05.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

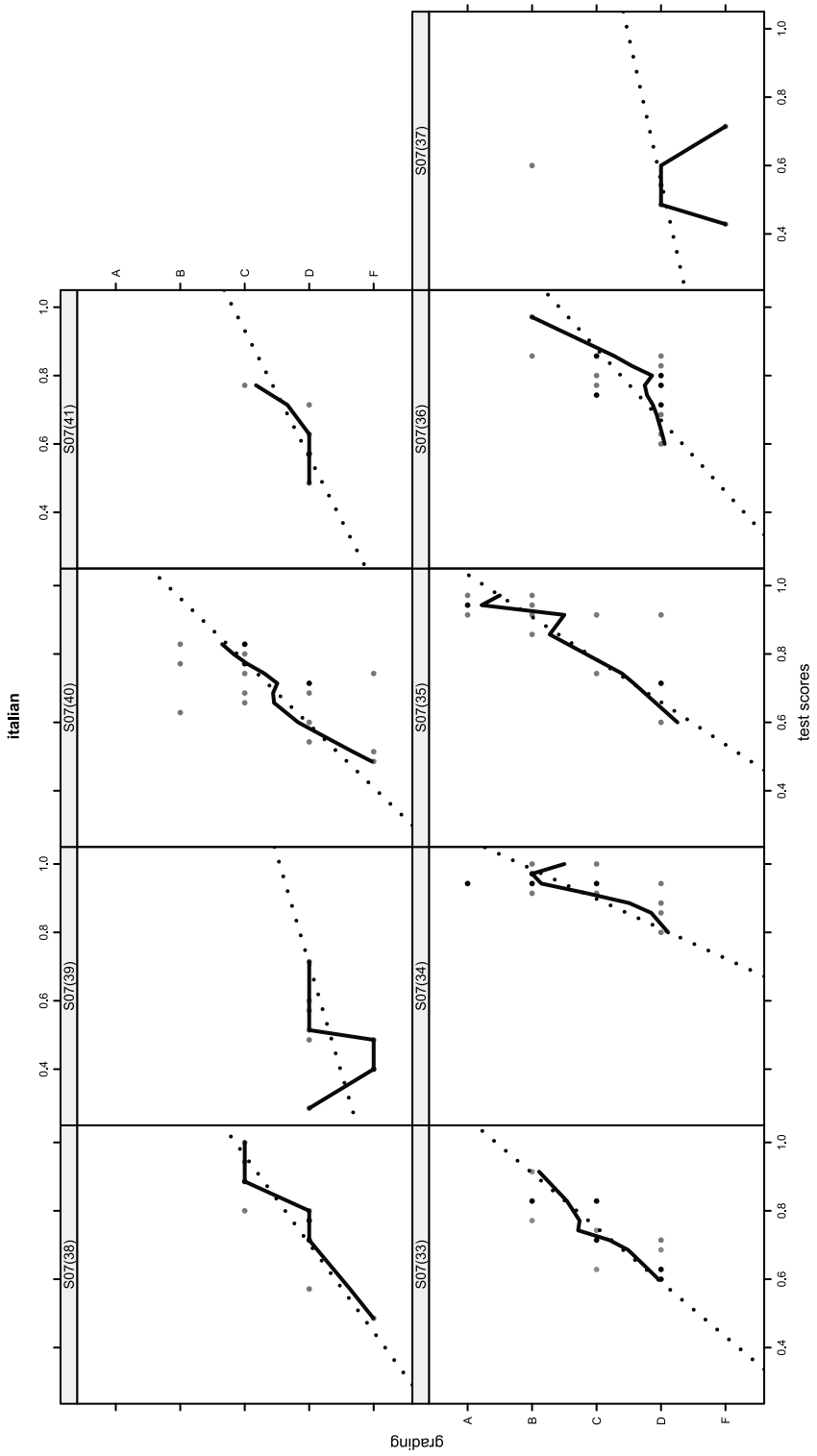

Figure 3. Grading practices of School S07 by class.

Table 2. 'Reference' grading function of normalized test scores.

Score	Grade	Original Italian grade
[0.0,0.6)	F	'Insufficiente'
[0.6,0.7)	D	'Sufficiente'
[0.7,0.8)	C	'Buono'
[0.8,0.9)	B	'Distinto'
[0.9,1.0]	A	'Ottimo'

5.1 The DfR *variable*

Let us go back to our current application and let us reclassify each grading practice according to the deviation of the teacher's behavior from the reference as identified in Table 2. For each class, the average test score corresponding to each grade is evaluated; then, an integer value is associated to each grade according to the discrepancy between the effective scores of the class and the theoretical scores of the reference class (see Table 3).

For instance, consider the average test score $\overline{S}_C$ corresponding to Grade C. We proceed as follows: if $\overline{S}_C \in [0,0.6)$, a value of +2 is assigned (strong over-evaluation); if $\overline{S}_C \in [0.6,0.7)$, a value +1 is assigned (mild over-evaluation); if $\overline{S}_C \in [0.7,0.8)$, a value of 0 is assigned (no deviation); if $\overline{S}_C \in [0.8,0.9)$, a value of –1 is assigned (mild under-evaluation); if $\overline{S}_C \in [0.9,1]$, a value of –2 is assigned (strong under-evaluation). Similarly for the other grades. When a grade is never used by a teacher, then that particular grade is considered as missing in his grading practice.

In such a way, a predominance of negative values indicates higher grading standard, whilst a predominance of positive values indicates lower standards. Each class or, more precisely, each teacher's grading practice is now identified by a profile containing the five variables *F, D, …, A*. The profiles represent the deviation of the teacher's average grades from the reference. Each variable *F, D, …, A* can assume integer values in the interval [–4; +4]. Next, teachers are grouped (e.g. by means of cluster analysis), and each group constitutes an item of the new categorical variable *DfR* describing the different grading practices applied by the teachers. This new variable *DfR* is then included into an education production function, in order to evaluate the effect of grading standards on the students' achievement.

5.2 Identification of the grading practices in our data

We have chosen to identify the grading practices adopted in 2005 and estimate their impact on achievement in the same year. This is because, while teachers are not

Table 3. Classification criterion of grading practices.

Score	Grade F	Grade D	Grade C	Grade B	Grade A
[0.0,0.6)	0	+1	+2	+3	+4
[0.6,0.7)	–1	0	+1	+2	+3
[0.7,0.8)	–2	–1	0	+1	+2
[0.8,0.9)	–3	–2	–1	0	+1
[0.9,1.0]	–4	–3	–2	–1	0

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

supposed to change class during the school year, they might have changed throughout the period 2003–2005: therefore, the restriction to year 2005 ensures a one-to-one correspondence between teachers and classes.

Nevertheless, in the two previous years (2003 and 2004) teachers and/or students might have changed class. Therefore, the achievement may be the result of different teachers' grading practices. The number of students who remain with the same teacher amounts to more than 66% of the sample size. The proportion of students who changed class is 2.5%. Further, to control for potential effects of this changes on achievement, we introduce into the EPF the dummy variables *NPROFITA* and *CHANGECL*. As will be seen in the empirical analysis, *CHANGECL* is usually not significant, while *NPROFITA* has a negative impact on achievement, but only level *NPROFITA* ≥ 3 is sometimes significant.

The grading practices of each teacher in 2005 have been profiled according to Table 3. The profiles are gathered by means of cluster analysis[6] and, as a result, 15 groups have been identified. These groups are described in Table 4.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

Table 4. Fifteen groups of grading practices identified by our method and corresponding to the items of the categorical variable *DfR*.

Class	Grade F	Grade D	Grade C	Grade B	Grade A
Group 1					
8	0	0	0		
13		0	0	0	
15	0	0	0	0	
19	0	0	0	0	
28	0	0	0	0	
29		0	0	0	
33		0	0	0	
47		0	0		
48	0	0	0		
56	0	0	0	–1	
Group 2					
2	–1	1	0	0	
10		1	0	0	
20	–1	1	0	0	
24	–1	1	0		
64		1	0	0	
69	0	1	0	0	
Group 3					
3	0	1	1	3	
11	0	1	1	2	3
70	0	1	1	2	
Group 4					
4	–1	1	1	1	
58	–2	1	1	2	1
74	–1	1	1		
76	–1	0	1	2	

Table 4. (*Continued.*)

Class	Grade F	Grade D	Grade C	Grade B	Grade A
Group 5					
5		1	0	1	
23	0	1	–1	1	
41		1	0		
53		1	0	1	
55	0	1	0	1	1
Group 6					
6	0	0	1	0	
18	0	0	1	0	
22	0	0	1		
49		0	1	0	
57	0	0	1	0	1
61		–1	1	0	0
77		0	1	0	
Group 7					
7	0	1	1	1	1
12	0	1	1	1	0
16		1	1	1	
25	0	1	1	1	
32		1	1	1	
39	0	1			
51		1	1	1	
67		1	1	1	1
68	0	1	1	1	
73	0	1	1		
Group 8					
1		–1	–1	0	0
27		–1	0	1	
65		–2	–1	0	
66	–1	–1	0	0	
71	0	–1	–1	0	
Group 9					
9	–1	1	2	2	
37	0	1		2	
59		1	2	2	
60		1	2	2	3
62	0	1	2	2	3
72	0	1	2	2	
Group 10					
14		0	0	1	
31	0	0	0	2	
40	0	0	0	1	

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

Table 4. (*Continued.*).

Class	Grade F	Grade D	Grade C	Grade B	Grade A
Group 11					
21	0	1	2	0	
44	0	1	2	1	
52		1	2	1	
54		1	2	2	2
63		0	2	2	1
Group 12					
17		1	1	0	
50		1	1	0	
Group 13					
26	0	−1	−1	−1	
30	−2	−1	−1	−1	0
35		−1	−1	−1	0
36		−1	0	−1	
46	−3	−2	−1	−1	
Group 14					
34		−2	−2	−1	0
38	0	−1	−2		
42		−2	−2		
75	−2	−1	−2		
Group 15					
43		1	1	1	2
45	0		1	1	2

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

The 15 groups correspond to the items of the categorical variable *DfR* and represent the variety of grading standards applied in the 77 classes of the survey in 2005.

6. Results on empirical data

This section illustrates the empirical analysis of the impact of grading practices on students' competence. We compare the results obtained using the DfR approach with those obtained assuming linearity of the grading equations.

6.1 Can grading practice be ordered?

Consider Table 4: the first group in *DfR* shows the smallest deviance from the reference grading practice; all of the other groups exhibit different levels and kinds of deviation, which are clearly difficult to order. In fact – while teachers in Group 13 or Group 14 surely have higher standards, compared with teachers in Group 9 or Group 12 – the comparability of Group 4 and Group 8 is more debatable: teachers in Group 4 tend to emphasize the differences among students and, therefore, apply high standard to the lowest achievement levels but are more generous, in terms of grades, to students

with medium-high performances; teachers in Group 8, on the contrary, seem to penalize intermediate performances.

This is the main contribution of variable *DfR*, compared with the indicators of grading standards usually adopted in the literature: *DfR* respects the heterogeneity of grading practices and does not force an order of grading standards. It simply groups standards that are similar according to the criterion announced in Table 3. Therefore, the impact we are going to estimate on students' achievement is not forced to be the impact of *higher* or *lower* standards, but – more realistically – it will be the effect of *different kinds* of grading practice.

6.2 The education production function with DfR

We estimate the following EPF:

$$M05ITA_i = \alpha_0 + \alpha_1 GENDER_i + \alpha_2 M03ITA_i + +\alpha_3 NCLASS_i + \alpha_4 NPROFITA_i + \alpha_5 CHANGECL_i + \alpha_6 BOOKS_i + \alpha_7 DfR_i \quad \text{(Mod1)}$$

where *i* runs in the set of indexes of all students, and the variables *M*05*ITA*, *GENDER*, *M*03*ITA*, *NCLASS*, *NPROFITA*, *CHANGECL* and *BOOKS* are defined in Section 4.1 while *DfR* is defined in Section 5.

Table 5 contains the results for Mod1.[7] The Italian language ability of students in 2005 clearly depends on the initial achievement level (*M*03*ITA*). Female students gain, *ceteris paribus*, higher achievement. Family background exhibits a positive impact whose value and significance increase with the quality of the background itself.

Several groups of classes in *DfR* have a significant impact on the achievement level. In particular, the highest positive effect with respect to the reference grading practice is shown by Groups 13 and 14. As we mentioned, teachers belonging to these groups have quite selective grading standards. Their students rarely receive a Grade A, but some of them would have if they had been graded using the reference grading practice: indeed, all the students in these classes with Grade B have test scores belonging to the interval [0.9, 1.0]. Moreover, the students with Grade F often would get a more than positive evaluation in the reference class. On the other side, the strongest negative effect comes from Groups 3, 9 and 12. In these classes, grading practices are clearly less severe: students with the highest grades often had a quite poor performance in the achievement test; frequently, students who would fail in the reference class obtain a Grade D or even a Grade C. To confirm that the relationship between grading practices and achievement cannot easily be reduced to a monotonic curve, we should also notice the composition of Group 6: in fact, despite the evidence of grading practices with null or slightly positive deviation, the impact on the test scores is significantly positive.

6.3 A comparison: **DfR** *versus linear regression*

We now compare the use of *DfR* with the use of linear grading equations in the EPF. Consider again the linear grading Equation (1) for each class:

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

Table 5. Results for Mod1, Mod2 and Mod3.

	Mod1	Mod2	Mod3
Intercept	0.558*	1.696***	0.659*
	(0.260)	(0.288)	(0.305)
GENDER: female	0.223***	0.260***	0.225***
	(0.046)	(0.049)	(0.046)
M03ITA	0.365***	0.351***	0.352***
	(0.024)	(0.025)	(0.024)
NCLASS	–0.006	0.004	0.004
	(0.007)	(0.007)	(0.007)
NPROFITA: two/one	–0.062	–0.055	–0.090
	(0.061)	(0.062)	(0.061)
NPROFITA: three or more/one	–0.106	–0.187*	–0.109
	(0.076)	(0.077)	(0.076)
CHANGECL: no/yes	0.473**	0.372*	0.443**
	(0.151)	(0.159)	(0.150)
BOOKS: 11–25/0–10	0.170	0.230	0.193
	(0.161)	(0.169)	(0.160)
BOOKS: 26–100/0–10	0.365*	0.387*	0.374*
	(0.152)	(0.159)	(0.151)
BOOKS: 101–200/0–10	0.465**	0.533***	0.493**
	(0.153)	(0.161)	(0.152)
BOOKS: >200/0–10	0.578***	0.672***	0.606***
	(0.152)	(0.160)	(0.151)
DfR: 2/1	–0.162		–0.109
	(0.100)		(0.101)
DfR: 3/1	–0.499***		–0.414**
	(0.130)		(0.138)
DfR: 4/1	–0.186		–0.096
	(0.122)		(0.124)
DfR: 5/1	–0.025		–0.022
	(0.115)		(0.120)
DfR: 6/1	0.357***		0.390***
	(0.098)		(0.102)
DfR: 7/1	–0.252**		–0.160
	(0.091)		(0.102)
DfR: 8/1	0.733***		0.799***
	(0.117)		(0.118)
DfR: 9/1	–0.637***		–0.453***
	(0.113)		(0.123)
DfR: 10/1	–0.067		–0.072
	(0.127)		(0.128)
DfR: 11/1	0.097		0.209
	(0.146)		(0.149)
DfR: 12/1	–0.491***		–0.272*

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

Table 5. (*Continued*).

	Mod1	Mod2	Mod3
	(0.108)		(0.131)
DfR: 13/1	0.904***		0.720***
	(0.108)		(0.118)
DfR: 14/1	1.254***		0.885***
	(0.123)		(0.158)
DfR: 15/1	–0.123		–0.029
	(0.171)		(0.189)
$\hat{\beta}_0$		–0.603***	–0.208**
		(0.041)	(0.074)
$\hat{\beta}_1$		–0.333***	–0.092
		(0.031)	(0.049)
R^2	0.451	0.384	0.459
Adj. R^2	0.440	0.378	0.447
AIC	2952.075	3068.101	2938.402
n	1219	1219	1219

Note: Mod1, EPF as in Equation (Mod1) with *DfR* only; Mod2, EPF as in Equation (Mod2) with β_0 and β_1 only; Mod3, combination of Mod1 and Mod2. Significance at ***0.001, **0.01, *0.05.

$$ITA05FST_i = \beta_{0i} + \beta_{1i}SCITA05_i, \qquad i \in C_j$$

where C_j is the set of indexes of the observations in class j, and $j = 1, \ldots, 77$, and insert the estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ as grading standard indicators into the following EPF:

$$M05ITA_i = \alpha_0 + \alpha_1 GENDER_i + \alpha_2 M03ITA_i + +\alpha_3 NCLASS_i + \alpha_4 NPROFITA_i + \alpha_5 CHANGECL_i + \alpha_6 BOOKS_i + \alpha_7 \hat{\beta}_{0i} + \alpha_8 \hat{\beta}_{1i} \qquad \text{(Mod2)}$$

At the 5% level, 92% of the estimated values of β_0 (respectively 97% at 1%) and 28% of the estimated values of β_1 (respectively 52% at 1%) are not significant: in other words, in our case the grading standards seem to be better summarized by the slope of the grading equation. Table 5 also presents the result of the EPF estimation for Mod2. The coefficients associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ are negative and significant, which is intuitive but not as informative as the results concerning the variable *DfR*. The overall goodness of the model measured by the adjusted R^2 and by the AIC[8] index confirms that our variable *DfR* better captures the relationship between students' achievements and grading practices.

Incidentally, Mod1 and Mod2 provide similar evidence: harder grading standards are associated with higher achievement levels. Our main point is: Do the linear grading equations and the *DfR* variable also provide the same information about the grading practices in the 77 classes? The answer seems to be negative. In fact, when *DfR* and the linear grading equation are used together in the EPF, as in Mod3:

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

$$M05ITA_i = \alpha_0 + \alpha_1 GENDER_i + \alpha_2 M03ITA_i + +\alpha_3 NCLASS_i + \alpha_4 NPROFITA_i + \alpha_5 CHANGECL_i + \alpha_6 BOOKS_i + \alpha_7 \hat{\beta}_{0i} + \alpha_8 \hat{\beta}_{1i} + \alpha_9 DfR_i \quad \text{(Mod3)}$$

what we notice is that all values of *DfR* that are significant in Mod1 are still significant in Mod3 with the exception of *DfR*7, and $\hat{\beta}_1$ is not significant. This shows how *DfR* captures both linearities and non-linearities. At the same time, $\hat{\beta}_0$ is significant and *DfR*7 is not but, looking at the group identified by *DfR*7, it can be seen that the corresponding grading practices are simply a translation of the reference practice (which is, in turn, linear). Therefore, the effect of *DfR*7 is captured by $\hat{\beta}_0$.

As mentioned, we have estimated Mod1–Mod3 with a two-stage least squares approach to take into account heteroscedasticity among classes that may potentially affect estimates of β_0 and β_1 due to different class sizes.[9]

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

7. Policy implications

Contrary to the majority of the literature, we have shown with an empirical analysis that the relationship between students' achievement and teachers' grades is frequently non-linear. We have also shown how unreliable it is to use data aggregated at school level to prove that policies exist at that level. On the other hand, our results confirm the accepted knowledge that higher grading standards can increase achievement. Therefore, the present paper puts its emphasis on basically two complementary facts: a correct analysis needs to identify *individual* teachers' grading practices and only after that can one aggregate at higher levels; and grading practices are usually non-linear, hence a DfR approach can properly identify individual grading practices (both linear and non-linear).

An underlying issue, which may prevent the adoption of the DfR approach, is that an *objective* measure of achievement has to be available (e.g. the scores of a multiple choice test). That is why we share with other studies (see Dardanoni, Modica, and Pennisi 2009a, 2009b) the recommendation for a centralized monitoring of students' competence, still absent in Italy: it is a pre-condition to compare educational institutions nationwide and to take degenerative phenomena (e.g. grade inflation, excessive heterogeneity in achievement levels) under control.

If the aim of centralized monitoring is to induce higher achievement levels, limit grade inflation and reduce heterogeneity in teachers' practices, the incentives provided by the educational authorities should be correctly designed. As we have shown, grading practices are teacher specific: hence, one can think that incentives should be conditioned to the single teacher performance, rather than to a hypothetical school-based grading policy. On the other hand, the policy-maker may wish to assign resources at some level of aggregation (school, district, etc.), because this pursues, among others, redistributive aims.

Therefore, the elements of an adequate policy (given the availability of an objective measure of competence at the single-teacher level)[10] could be the following: determine a baseline grading practice at some central level; evaluate the DfR indicator at the single-teacher level; aggregate the results at the school/district level by, for example, measuring the proportion of teachers that are compliant or use grading standards stricter than the reference (i.e. DfR profile contains zeros or negative

numbers); and condition the resources assignment to the proportion of compliant (or more demanding) practices.

The policy, as it can be seen, tolerates a portion of 'deviant' behaviors (less demanding grading procedures): this is to allow some teachers use strategically their grading practices to address all of the school aims, including the social mission of the educational system (the 'no child left behind' issue). As suggested by the literature (Lillard and DeCicca 2001), in fact, the adoption of severe standards may increase the incidence of dropouts.

8. What can go wrong?

This section aims at discussing a few issues that may prevent or limit the use of our approach in specific situations.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

8.1 What if a reference cannot be easily found?

In our particular experiment a reference like the one proposed in Table 2 can be reasonably designed and accepted. In some contexts a reference is sometimes exogenous: in measuring health status of individuals, some reference values for biological tests (blood pressure, etc.) are usually provided by authorities (e.g. World Health Organization); in quality of life analysis, several indicators (earnings, family structure, town size, level of pollution, etc.) are used to identify a reference level of welfare; and so forth. In other cases, the identification of a reference may not be so obvious. In these situations a reference can still be identified by some automatic rule. To simplify the exposition and still make it general, assume both the objective quantitative measurement X (e.g. the scores) and the subjective qualitative evaluation Y (e.g. the gradings) are available. Assume Y has k values and partition the support of X in exactly k intervals. Then, the following two conditions must be realized in order to properly create a *DfR* variable that is able to discriminate deviations from the automatic reference: each interval of X is uniquely associated (mapped) to a single value of Y (and *vice versa*); and the intervals must contain a sufficient number of observations. Therefore, for example, a straightforward way to construct an automatic reference is to partition the support of X in equi-frequent intervals and associate the first interval of X with the first value of Y (even if Y is not necessarily ordered), associate the second interval of X with the second value of Y, and so forth. Of course, when a natural or meaningful reference can be identified – like in our application – its use is advisable because it helps with the interpretation of the results.

8.2 Subjective versus objective

The DfR approach clearly works on the basis of the existence of an objective measure of achievement. This is reasonable to obtain in situations where achievement is measured by an exogenous (to the teacher) test (e.g. multiple choice, comprehension test) in which marking is not controversial. In this case, given the maximum potential score of the test, normalization of students' scores are easy and reasonably interpretable. In examinations where the measure of achievement cannot be objectively quantified (e.g. examinations by essays), some bias in the value of the final score is induced by the correctors, no matter how strict are the guiding rules given to them. In these

situations, our approach is hardly justifiable in that DfR assumes that an accepted reference can be given and deviation from it can be measured.

The test, as usual, has to be well calibrated (not too easy, not too hard), in order to avoid that the scores are squeezed at the top or at the bottom of the score range: in that case, in fact, a teacher might appear as non-compliant with the reference grading practice, because the test is not able to discriminate the different achievement levels of students. Clearly, it is easy to detect a non-calibrated test *ex post* and, possibly, re-calibrate it before applying the DfR procedure.

8.3 School level versus teacher level

If the dimension of the classes is too small, both linear modeling and DfR may suffer from very small sample size. In particular, the *DfR* variable may be difficult to identify because cluster analysis may fail to find significantly different grading profiles. In these cases, aggregation at school level is the only information that can be extracted from the data. Still, DfR remains competitive compared with linear modeling.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009

Acknowledgements

The authors wish to thank IReR Lombardia and IRRE Lombardia for providing the data from the Research Project IReR:2005B018, and in particular Guido Gay. They are grateful to Massimiliano Bratti, Gary King and two anonymous referees for their helpful comments.

Notes

1. Formally, it is an inverse grading equation.
2. For example, each score is divided by the maximal attainable score of the complete test.
3. In the Italian educational system, 'elementary' corresponds to Grades 1–5, 'lower secondary' to Grades 6–8, and 'higher secondary' up to Grade 13.
4. It is a polynomial local regression (*loess*): see Cleveland, Grosse, and Shyu (1992).
5. Which may or may not exist for a given data-set.
6. In our application, we apply the Random Recursive Partitioning method (see Iacus and Porro 2007, 2009) to derive a proximity measure between profiles, and then apply the hierarchical cluster analysis. Any other clustering method or distance between profiles would be applicable.
7. In order to take into account for variables estimated/defined at class/cluster level, all models (Mod1, Mod2 and Mod3) are estimated using a two-stage Weighted Least Square (WLS) regression with appropriate weights.
8. Indeed, the AIC statistic is given by $AIC = -2\log \text{likelihood}(\theta) + 2\dim(\theta)$, where θ is the vector of coefficients and $\dim(\theta)$ is the number of parameters to be estimated. In Mod1 we have, due to *DfR*, 13 parameters more than in Mod2: this notwithstanding, the AIC has a lower value.
9. To control for potential unobservables we have also regressed $M05ITA_i - M03ITA_i$ against all other variables of Mod1–Mod3, with either *DfR* or $(\hat{\beta}_0, \hat{\beta}_1)$ or both. In all of these modifications, the empirical evidence remains basically the same as in Mod1–Mod3.
10. This is, for instance, still not available in the OECD PISA 2003 survey.

References

Betts, J.R. 1997. Do grading standards affect the incentive to learn? Discussion Paper 97–22, Department of Economics, University of California at San Diego.

Betts, J.R. 1998. The impact of educational standards on the level and distribution of earnings. *American Economic Review* 88: 266–75.

Betts, J.R., and J. Grogger. 2003. The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review* 22: 343–52.

Bonesronning, H. 1999. The variation in teachers' grading practices: Causes and consequences. *Economics of Education Review* 18: 89–105.

Bonesronning, H. 2004a. Do the teachers' grading practices affect student achievement? *Education Economics* 12: 151–67.

Bonesronning, H. 2004b. Can effective teacher behavior be identified? *Economics of Education Review* 23: 237–47.

Bonesronning, H. 2008. The effect of grading practices on gender differences in academic performance. *Bulletin of Economic Research* 60: 245–64.

Card, D. 1999. The causal effect of schooling on earnings. In *Handbook of labor economics,* ed. O. Ashenfelter and D. Card, 1801–63. Amsterdam: North Holland.

Cleveland, W.S., E. Grosse, and W.M. Shyu. 1992. Local regression models. In *Statistical models in S,* ed. J.M. Chambers and T.J. Hastie, 309–76. Monterey: Wadsworth and Brooks-Cole.

Correa, H., and G.W. Gruver. 1987. Teacher–student interaction: A game theoretic extension of the economic theory of education. *Mathematical Social Science* 13: 19–47.

Costrell, R.M. 1994. A simple model of educational standards. *American Economic Review* 84: 956–71.

Dardanoni, V., S. Modica, and A. Pennisi. 2009a. Grading across schools. *The B.E. Journal of Economic Analysis and Policy* 9: no. 1.

Dardanoni, V., S. Modica, and A. Pennisi. 2009b. School grading and institutional contexts. Mimeo. Università di Palermo, Palermo.

Figlio, D.N., and M.E. Lucas. 2004. Do high grading standards affect student performance? *Journal of Public Economics* 88: 1815–34.

Hanushek, E.A. 2002. Evidence, politics, and the class size debate. In *The class size debate,* ed. L. Mishel and R. Rothstein, 37–65. Washington, DC: Economic Policy Institute.

Hanushek, E.A. 2003. The failure of input-based schooling policies. *Economic Journal* 113: F64–98.

Iacus, S.M., and G. Porro. 2007. Missing data imputation, matching and other applications of random recursive partitioning. *Computational Statistics and Data Analysis* 52, no. 2: 773–89.

Iacus, S.M., and G. Porro. 2009. Random Recursive Partitioning: A matching method for the estimation of the average treatment effect. *Journal of Applied Econometrics* 24: 163–85.

Johnes, G. 2004. Standards and grade inflation. In *International handbook on the economics of education,* ed. G. Johnes and J. Johnes, 462–83. Cheltenham: Edward Elgar.

Krueger, A.B. 2002. Understanding the magnitude and effect of class size on student achievement. In *The class size debate,* ed. L. Mishel and R. Rothstein, 7–35. Washington, DC: Economic Policy Institute.

Krueger, A.B. 2003. Economic considerations and class size. *Economic Journal* 113: F34–63.

Lillard, D.R., and P.P. DeCicca. 2001. Higher standards, more dropouts? Evidence within and across time. *Economics of Education Review* 20: 459–73.

Marks, D. 2002. Academic standards as public goods and varieties of free-rider behaviour. *Education Economics* 10: 145–63.

OECD. 2005. *PISA 2003 technical report.* Paris: OECD.

Downloaded By: [Porro, Giuseppe] At: 20:37 7 September 2009