# Teachers' evaluations and students' achievement: how to identify grading standards and measure their effects.

Stefano M. Iacus,
Department of Economics, Business and Statistics, University of Milan
Via Conservatorio 7, 20122 Milan, Italy

Giuseppe Porro*,
Department of Economics and Statistics, University of Trieste,
Piazzale Europa 1, 34127 Trieste, Italy

August 26, 2008

**Abstract**

A new procedure to identify grading practice is proposed. In our approach, grading practice are given in terms of a categorical variable whilst usually in the literature, coefficients of the regression line which models school grades as a function of students' achievement are taken as indicators of grading standards. The procedure, which is essentially nonparametric, allows to identify clearly a variety of grading practices and their effect on students' performance. It also shows that ordering grading standards is not possible: hence the usual approach based on regression coefficients is unlikely to be satisfactory.

The new methodology is easy to implement and widely applicable. As an example, we consider data from a survey on Italian lower secondary school students. The evidence, which essentially confirms the generic result provided by the literature, suggests that higher grading standards improve students' achievement but, in our case, grading practices are more easily interpretable.

**J.E.L. classification**: I2
**Key words**: grading practice; students' achievement; classification techniques

*Corresponding author. E-mail giuseppe.porro@econ.units.it

# 1 Introduction

As an increasing stream of papers suggests, the strategic behavior of teachers - besides their education, professional experience, gender - may affect students' effort and achievement at any level of the education process. Among the behavioral features which can have an impact on students' performance, a relevant role seems to be played by their grading practices: according to several empirical studies, in fact, 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.

The theoretical rationale for this relationship is set out in the student-teacher interaction model proposed by Correa and Gruver (1987), where the hypothesis is made that students care about the teacher's evaluation, both because this evaluation is, in itself, a reward of their effort in learning and because high grades are an important pre-requisite for admission in higher education institutions or for finding good jobs. Formally, school grades appear among the arguments of the utility function of the student: they are an increasing function of the student's real achievement which, in its turn, is an increasing function of the learning effort of the student. Therefore, the teacher may affect the student's effort and achievement, making a strategic use of the grading practice.

In empirical studies, researchers measure the stringency of grading standards assuming a linear *grading equation*, i.e. a linear relationship between achievement (usually measured by a test score) and grades and estimating the parameters of this function. Frequently, one of the coefficients is not sig-

nificant, hence the other coefficient is used to summarize and order grading practices. Further, the coefficients (one or both) are included as regressors into an education production function, and used to estimate the impact of grading standards on achievement.

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 (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 uncorrect evaluation of the grading standards.

In this paper a new methodology to identify the nonlinear relationship between achievement and grades is proposed. The procedure is based on the following steps: suppose the normalized score of a test and teacher's grading for each student are available. An arbitrary, *a priori* and pre-test relationship between scores and gradings can be formulated, i.e. scores in $[0, .6)$ correspond to grade $F$, scores in $[0.6, 0.7)$ to grade $D$, etc. Then, it is possible to measure the discrepancy between the teacher's grade and the *a priori* grade for each given student score. Further, standard classification techniques (cluster analysis) allow to identify and gather teachers who adopt similar grading standards. So, we can define a categorical variable, which identifies the different grading practices adopted by the teachers. This variable can be included into the education production function in order to estimate the effect of grading standards on students' performance.

As an example, we apply the procedure to a survey on the achievement of lower secondary school students in 77 classes in Lombardy (Italy) in 2003-2005. The results confirm that higher grading standards favor students' performance. For completeness, a comparison with the traditional linear analysis is shown.

The paper is structured as follows: Section 2 reviews the findings available in the literature. Section 3 explains the new approach in details and Section 4 contains an application of the method. In the Appendix we present, in brief, another application, just to show that the method is robust and widely applicable to different situations.

# 2 The literature

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 these models provide justifications for higher grade standards affecting students' effort and achievement. Still, the former is mainly focused on the incentive grade-effort-achievement relationship, the latter also consider the potential distributional aspects of the problem in a principal-agent theoretical framework.

In particular, in Correa and Gruver (1987) the students care about the teacher's *perceived* achievement, which is represented by the teacher's evaluation. This implies that school grades appear as an argument of the students' utility function. Grades are supposed to increase with the *actual* achievement of the students (*grading equation*). The educational achievement is,

as well, an increasing function of the students' effort. Therefore, students' effort and achievement are affected by the teachers' grading strategies: more precisely, higher grading standards induce higher effort and improve the students' performance in the achievement test.

Crucially, the model assumes a linear grading equation and, consequently, the empirical studies that move from this theoretical framework (Bonesronning, 1999, 2004a, 2004b, 2004c) make the same linearity assumption. Estimating the parameters of the grading equation, Bonesronning finds that the slope is seldom significant, while the intercept is[1]: therefore, the estimated intercept of the grading equation is usually included as a regressor into the education production function, in order to evaluate the effect of grading standards on students' achievement.

Costrell (1994) sets up an asymmetric information model where firms cannot observe the actual productivity of workers and rely on the possession of a high-school diploma to evaluate future productivity of new hires. If productivity is an increasing function of effort and the minimum knowledge to obtain a diploma increases (i.e. if the grading standard becomes more stringent), the students who decide to get a diploma will make more effort and be more productive. On the other side, marginal students might abandon high-school when the evaluation standards become more severe[2]. One of the consequences is that an egalitarian policy reduces standards, in order to increase the graduation rate. The result is generalized by Betts (1998),

[1] An increase in the intercept of the grading equation is equivalent to income effect and hence, if leisure is a normal good, the prediction will be a decrease in learning effort. On the other hand, no clear-cut prediction can be made for changes in the slope of the equation, where income and substitution effect work jointly.

[2] See Lillard and DeCicca (2001).

who shows that, if students differ in ability, "an egalitarian policy maker might prefer higher standards than would a policy maker whose goal was to maximize the sum of earnings" (p.266).

The effect of higher standards on average achievement and their potential distributional consequences still remain an empirical matter. Betts (1997) and Betts and Grogger (2003) both provide evidence on this point. The first paper indicates that higher standards improve average students' performance, affecting the achievement test scores of abler students more than those of the less able; Betts and Grogger (2003) confirm the results, pointing out that disadvantaged groups (e.g. minority students) might register a reduction in graduation rates due to more severe requirements and to the increase in the achievement gap with respect to top-level students.

In both papers the stringency of grading standards is measured by a linear regression of the achievement test scores on the school grades[3], 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: initially abler students seem to experience higher benefits from higher standards. Nevertheless, the relationship between initial ability and improvement is more complex - and linked to a peer-group effect - in Figlio and Lucas (2004): in fact, they find that

[3]Formally, it is an inverse grading equation.

"initially low-ability students benefit most from high standards when their classmates are high-ability, while initially high-ability students benefit most from high standards when their classmates are low-ability" (p.1817).

Three different measures of grading standards are proposed in Figlio and Lucas (2004). One of them is an inverse linear grading equation including teacher-level fixed effects. The second is the average difference between the test scores and the grades obtained by the students of a specific teacher: the higher the difference, the higher the teacher's standard. The third is the average score obtained by students who received by a specific teacher a grade of B. The second measure proves to be the most conservative and therefore is adopted in the study. Both the second and the third measure are nonparametric and go in the direction of our proposal.

# 3 The procedure

As discussed in the previous section, the grading equations generally aim to estimate a single-quantity indicator of grading standard, allowing for ordering teachers' behavior according to their stringency: that seems to be the reason why usually the linearity assumption is made and the estimated intercept of the equation is used as a regressor into the education production functions.

Unfortunately, it is common sense that the relationship between achievement and grades is hardly linear. In Figure 1 the result of a survey on 77 lower-secondary-school classes in Lombardy (Italy) is represented [4]: for each class, the relationship between the scores of achievement test ($SCITA05$)

[4]The data set is described and analyzed in the following section.

and the school grades ($ITA05FST$) on Italian language in 2005 is shown. Both the linear regression and a nonparametric regression [5] are represented. 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 (see, e.g., classes 612, 618, 696, 52, 205), sometimes they grow less than proportionally (see, e.g., classes 765, 753, 462), in other classes the grading practice are simply non monotonic with respect to the achievement scores (see, e.g. classes 616, 510, 54). Each of these behaviors correspond to a grading practice which we try to identify properly with our new approach. In particular, we summarize the grading practices as deviations from a *reference behavior* (that conventionally we name the "fair" behavior). Define as a reference a class, which may or may not exist for a given data set, where grades are given according to the following *a priori* and arbitrary[6] scale of achievement scores:

| Scores | 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" |

Table 1: "Reference" grading function of normalized test scores

Now, each grading practice can be reclassified according to the deviation of the teacher's behavior from the standard depicted in Table 1. For each class, the average test score corresponding to each grade is evaluated; then,

[5] It is a polynomial local regression (*loess*): see Cleveland *et al.* (1992).

[6] But not unreasonable.

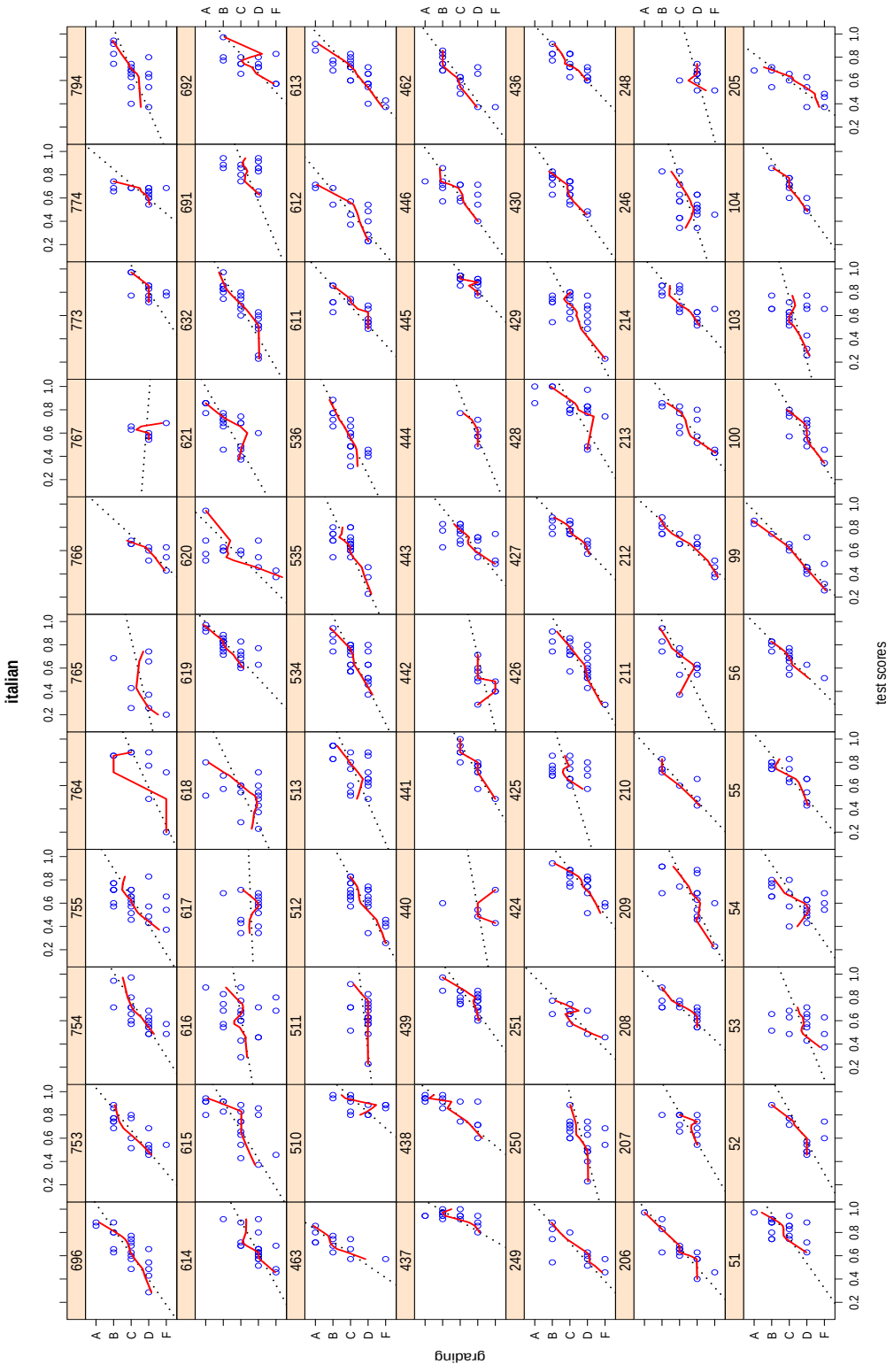


Figure 1: The non linear behaviour of grading practices (data set “Italian”)

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 2).

| Scores | F | D | C | B | 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 |

Table 2: Classification criterion of grading practices

For instance, if the average test score corresponding to grade C:

- is in [0, 0.6), a value +2 is assigned (strong over-evaluation),
- is in [0.6, 0.7), a value +1 is assigned (mild over-evaluation),
- is in [0.7, 0.8), a value 0 is assigned (no discrepancy),
- is in [0.8, 0.9), a value -1 is assigned (mild under-evaluation),
- is in [0.9, 1], a value -2 is assigned (strong under-evaluation)

and so forth. When a grade is never used by a teacher, 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 five new variables $F, D, \ldots, A$, representing the discrepancy of the teacher's average grades from the theoretical ones. Each variable can assume

integer values in the interval [-4; +4]. Next, teachers are grouped by means of cluster analysis and each group is an item of the new categorical variable $CLDEV$ describing the different grading practices applied by the teachers. We choose to use the five variables to define a proximity measure among classes and to generate a *proximity matrix*[7] used further in the cluster analysis algorithm to obtain $CLDEV$. Any other clustering method would be applicable.

This new variable $CLDEV$ can be included into an education production function, in order to evaluate the effect of grading standards on the students' achievement.

# 4 The application

We test our approach on a survey on lower secondary school students in Lombardy (Italy). The grading practices are represented both by the new variable $CLDEV$ and via the linear grading equation. The two representations are included alternatively into the education production function, in order to estimate the effects of grading practices on students' achievement, and the results are compared.

[7]The proximity used is based on the RRP method defined in Iacus and Porro (2006, 2007). Any other proximity measure being equal to one when the units are identical, to a value in the interval (0,1) when they differ by at least one variable and is equal to zero when the units are totally different can be applied.

## 4.1 The data set

The survey has been carried out on a sample of 20 lower secondary schools in Lombardy (Italy) during the period 2003-2005. Three multiple-choice achievement tests (Italian language, Mathematics and Science) have been administered to first-years students in March 2003. Similar tests (for Italian language and Mathematics only) have been submitted to the same students subsequently (second-year students in May 2004; third-year students in May 2005). In May 2005 a questionnaire has been administered both to students (about their school carrier and school climate) and 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 is reported and briefly commented in the Appendix: all the results obtained from the Italian language data set are confirmed.

Every year, the final grade of the students is registered ($ITA03$, $ITA04$). The third-year grade corresponds to the teacher's evaluation at the end of the first semester ($ITA05FST$). The normalized test scores are represented by variables $SCITA03$, $SCITA04$, $SCITA05$ respectively, and have been converted into Rasch measures $M03ITA$, $M04ITA$, $M05ITA$ to make scores comparable from year to year when used together in the same regression model. This conversion does not affect the definition of $CLDEV$.

## 4.2 The grading practices

We evaluate the grading practice adopted in 2005 and estimate their impact on achievement in the same year. This is because, while teachers are not supposed to change class during the school year, they might have changed throughout the period 2003-2005: therefore, the restriction assures a one to one correspondence between teachers and classes.

The grading practices of each teacher in 2005 have been identified according to variable $CLDEV$ of the previous section. The grading practices are gathered as in the dendrogram produced by the cluster analysis and reported in Figure 2. The dendrogram suggests a classification into fifteen groups, described in Table 3. The fifteen groups constitute the items of the categorical variable $CLDEV$ and represents the variety of grading standards applied in the 77 classes of the survey in 2005.

## 4.3 Can grading standards be ordered?

The first group in $CLDEV$ shows the highest amount of "fairness" or "lowest discrepancy" from the reference grading practice; all the other groups exhibit different levels and kinds of "non-fairness" or deviation, which are clearly difficult to order. In fact - while teachers in group 13 or 14 surely have higher standards, compared to teachers in group 9 or 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

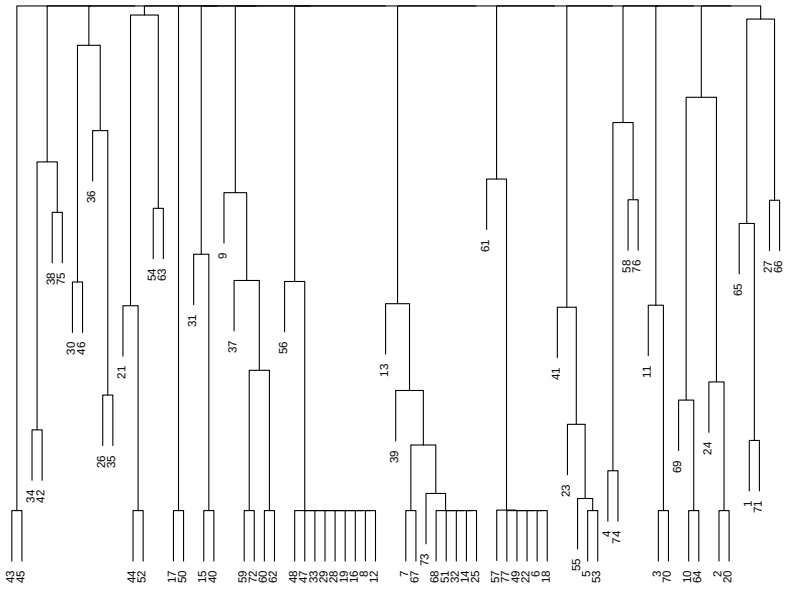

Figure 2: Dendrogram of the 77 teacher's grading practices (data set "Italian")

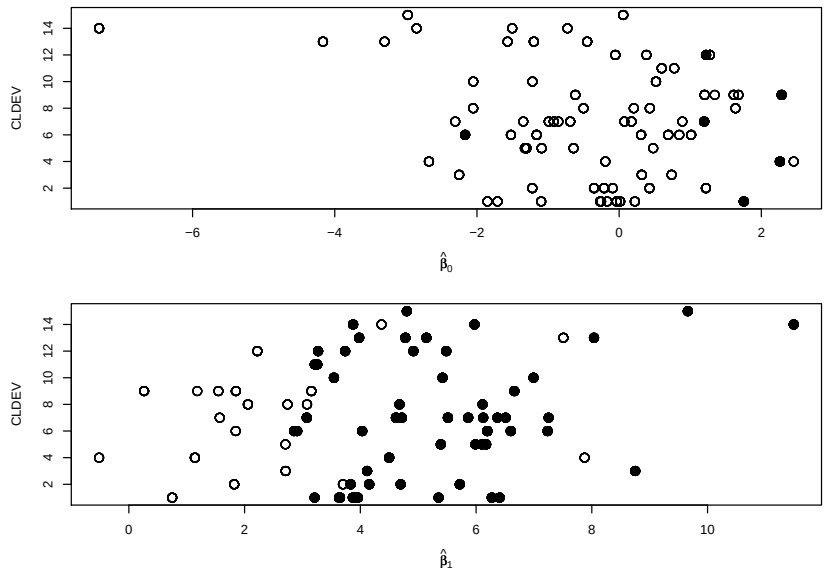


Figure 3: *CLDEV* against $\hat{\beta}_0$ and $\hat{\beta}_1$. Filled dots correspond to coefficients significative at 5% level (data set "Italian")

| | F | D | C | B | A |
|---|---|---|---|---|---|
| Group 1 | | | | | |
| 100 | 0 | 0 | 0 | | |
| 207 | | 0 | 0 | 0 | |
| 209 | 0 | 0 | 0 | 0 | |
| 213 | 0 | 0 | 0 | 0 | |
| 426 | 0 | 0 | 0 | 0 | |
| 427 | | 0 | 0 | 0 | |
| 436 | | 0 | 0 | 0 | |
| 511 | | 0 | 0 | | |
| 512 | 0 | 0 | 0 | | |
| 614 | 0 | 0 | 0 | −1 | |
| Group 2 | | | | | |
| 52 | −1 | 1 | 0 | 0 | |
| 104 | | 1 | 0 | 0 | |
| 214 | −1 | 1 | 0 | 0 | |
| 250 | −1 | 1 | 0 | | |
| 632 | | 1 | 0 | 0 | |
| 754 | 0 | 1 | 0 | 0 | |
| Group 3 | | | | | |
| 53 | 0 | 1 | 1 | 3 | |
| 205 | 0 | 1 | 1 | 2 | 3 |
| 755 | 0 | 1 | 1 | 2 | |
| Group 4 | | | | | |
| 54 | −1 | 1 | 1 | 1 | |
| 616 | −2 | 1 | 1 | 2 | 1 |
| 767 | −1 | 1 | 1 | | |
| 774 | −1 | 0 | 1 | 2 | |
| Group 5 | | | | | |
| 55 | | 1 | 0 | 1 | |
| 249 | 0 | 1 | −1 | 1 | |
| 444 | | 1 | 0 | | |
| 611 | | 1 | 0 | 1 | |
| 613 | 0 | 1 | 0 | 1 | 1 |
| Group 6 | | | | | |
| 56 | 0 | 0 | 1 | 0 | |
| 212 | 0 | 0 | 1 | 0 | |
| 248 | 0 | 0 | 1 | | |
| 513 | | 0 | 1 | 0 | |
| 615 | 0 | 0 | 1 | 0 | 1 |
| 619 | | −1 | 1 | 0 | 0 |
| 794 | | 0 | 1 | 0 | |
| Group 7 | | | | | |
| 99 | 0 | 1 | 1 | 1 | 1 |
| 206 | 0 | 1 | 1 | 1 | 0 |
| 210 | | 1 | 1 | 1 | |
| 251 | 0 | 1 | 1 | 1 | |
| 430 | | 1 | 1 | 1 | |
| 442 | 0 | 1 | | | |
| 535 | | 1 | 1 | 1 | |
| 696 | | 1 | 1 | 1 | 1 |
| 753 | 0 | 1 | 1 | 1 | |
| 766 | 0 | 1 | 1 | | |

| | F | D | C | B | A |
|---|---|---|---|---|---|
| Group 8 | | | | | |
| 51 | | −1 | −1 | 0 | 0 |
| 425 | | −1 | 0 | 1 | |
| 691 | | −2 | −1 | 0 | |
| 692 | −1 | −1 | 0 | 0 | |
| 764 | 0 | −1 | −1 | 0 | |
| Group 9 | | | | | |
| 103 | −1 | 1 | 2 | 2 | |
| 440 | 0 | 1 | | 2 | |
| 617 | | 1 | 2 | 2 | |
| 618 | | 1 | 2 | 2 | 3 |
| 620 | 0 | 1 | 2 | 2 | 3 |
| 765 | 0 | 1 | 2 | 2 | |
| Group 10 | | | | | |
| 208 | | 0 | 0 | 1 | |
| 429 | 0 | 0 | 0 | 2 | |
| 443 | 0 | 0 | 0 | 1 | |
| Group 11 | | | | | |
| 246 | 0 | 1 | 2 | 0 | |
| 462 | 0 | 1 | 2 | 1 | |
| 536 | | 1 | 2 | 1 | |
| 612 | | 1 | 2 | 2 | 2 |
| 621 | | 0 | 2 | 2 | 1 |
| Group 12 | | | | | |
| 211 | | 1 | 1 | 0 | |
| 534 | | 1 | 1 | 0 | |
| Group 13 | | | | | |
| 424 | 0 | −1 | −1 | −1 | |
| 428 | −2 | −1 | −1 | −1 | 0 |
| 438 | | −1 | −1 | −1 | 0 |
| 439 | | −1 | 0 | −1 | |
| 510 | −3 | −2 | −1 | −1 | |
| Group 14 | | | | | |
| 437 | | −2 | −2 | −1 | 0 |
| 441 | 0 | −1 | −2 | | |
| 445 | | −2 | −2 | | |
| 773 | −2 | −1 | −2 | | |
| Group 15 | | | | | |
| 446 | | 1 | 1 | 1 | 2 |
| 463 | 0 | | 1 | 1 | 2 |

Table 3: The 15 groups of grading practices identified by our method and corresponding to the items of the categorical variable $CLDEV$ (data set "Italian")

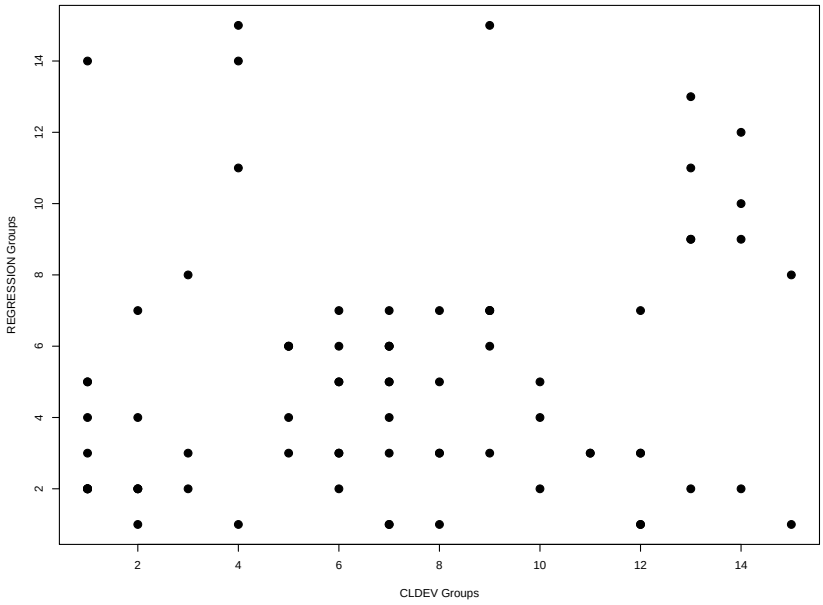


Figure 4: Classification of grading practices obtained using $CLDEV$ and the regression coefficient $(\hat{\beta}_0, \hat{\beta}_1)$. The plot shows no relationships between the two sets of classes (data set "Italian")

to penalize intermediate performances.

This is the main contribution of variable $CLDEV$, compared to the indicators of grading standards usually adopted in the literature: $CLDEV$ respects the heterogeneity of grading practices and does not force an order of grading standards. It simply groups standards which are similar according to the criterion announced in Table 2. 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. As we will see, tracing the effect of more (or less) severe teachers on achievement will not be impossible even in this more realistic context, but it requires to be more careful.

## 4.4 The education production function with $CLDEV$

We estimate the following education production function (EPF):

$$\begin{aligned} M05ITA_i = &\alpha_0 + \alpha_1 GENDER_i + \alpha_2 M03ITA_i + +\alpha_2 NCLASS_i \\ &+ \alpha_3 NPROFITA_i + \alpha_4 CHANGECL_i + \alpha_5 BOOKS_i \\ &+ \alpha_6 CLDEV_i \end{aligned} \quad (1)$$

where $i$ runs in the set of indexes of all the students, $GENDER$ is the student's gender, $M03ITA$ is the initial achievement level, $NCLASS$ is the number of students in the class, $NPROFITA$ is a discrete variable indicating whether the student changed the teacher once ore more during the lower secondary school, $CHANGECL$ is a dummy indicating whether the student changed class during the lower secondary school, $BOOKS$ is an indicator of the family background (number of books at home).

Table 4 contains the results. The Italian language ability of students in 2005 clearly depends on the initial achievement level ($M03ITA$). 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 $CLDEV$ 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, i.e. by the "fair

teacher": 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 "fair" class obtain a "D" or even a "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 fair or medium-low grading standards, the impact on the test scores is significantly positive.

## 4.5 A comparison: $CLDEV$ vs linear regression

We now estimate the linear grading equation for each class:

$$ITA05FST_i = \beta_{0i} + \beta_{1i} SCITA05_i, \quad i \in C_j, \tag{2}$$

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

$$\begin{aligned} M05ITA_i = &\alpha_0 + \alpha_1 GENDER_i + \alpha_2 M03ITA_i + +\alpha_2 NCLASS_i \\ &+ \alpha_3 NPROFITA_i + \alpha_4 CHANGECL_i + \alpha_5 BOOKS_i \\ &+ \alpha_6 \hat{\beta}_{0i} + \alpha_7 \hat{\beta}_{1i} \end{aligned} \tag{3}$$

| | Estimate | Std. Error | t value | Pr(>\|t\|) | Signif. |
|---|---|---|---|---|---|
| (Intercept) | 0.5583 | 0.2599 | 2.15 | 0.0319 | * |
| GENDER (female) | 0.2227 | 0.0465 | 4.79 | 0.0000 | *** |
| M03ITA | 0.3652 | 0.0239 | 15.29 | 0.0000 | *** |
| NCLASS | −0.0056 | 0.0066 | −0.85 | 0.3961 | |
| NPROFITA (2) | −0.0623 | 0.0610 | −1.02 | 0.3079 | |
| NPROFITA ($\geq 3$) | −0.1064 | 0.0764 | −1.39 | 0.1637 | |
| CHANGECL (no) | 0.4732 | 0.1513 | 3.13 | 0.0018 | ** |
| BOOKS (11-25) | 0.1698 | 0.1605 | 1.06 | 0.2905 | |
| BOOKS (26-100) | 0.3649 | 0.1516 | 2.41 | 0.0162 | * |
| BOOKS (101-200) | 0.4651 | 0.1529 | 3.04 | 0.0024 | ** |
| BOOKS ($> 200$) | 0.5781 | 0.1522 | 3.80 | 0.0002 | *** |
| CLDEV 2 | −0.1616 | 0.1004 | −1.61 | 0.1077 | |
| CLDEV 3 | −0.4992 | 0.1301 | −3.84 | 0.0001 | *** |
| CLDEV 4 | −0.1860 | 0.1221 | −1.52 | 0.1280 | |
| CLDEV 5 | −0.0250 | 0.1153 | −0.22 | 0.8285 | |
| CLDEV 6 | 0.3569 | 0.0982 | 3.63 | 0.0003 | *** |
| CLDEV 7 | −0.2518 | 0.0906 | −2.78 | 0.0055 | ** |
| CLDEV 8 | 0.7330 | 0.1174 | 6.24 | 0.0000 | *** |
| CLDEV 9 | −0.6372 | 0.1130 | −5.64 | 0.0000 | *** |
| CLDEV 10 | −0.0672 | 0.1273 | −0.53 | 0.5976 | |
| CLDEV 11 | 0.0974 | 0.1461 | 0.67 | 0.5052 | |
| CLDEV 12 | −0.4913 | 0.1083 | −4.54 | 0.0000 | *** |
| CLDEV 13 | 0.9042 | 0.1080 | 8.37 | 0.0000 | *** |
| CLDEV 14 | 1.2545 | 0.1234 | 10.17 | 0.0000 | *** |
| CLDEV 15 | −0.1233 | 0.1714 | −0.72 | 0.4718 | |

Signif. codes: *** 0.001 ** 0.01 * 0.05 . 0.1
Multiple R-Squared: 0.4512, Adjusted R-squared: 0.4401, AIC = 2952.075

Table 4: Education production function estimated as in equation (1) for scores in "Italian" test

At 5% level, 92% of the estimated values of $\beta_0$ (respectively 97% at 1%) and 28% of the estimated values of $\beta_1$ (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 shows the result of the EPF estimation. The coefficients associated to $\hat{\beta}_0$ and $\hat{\beta}_1$ are negative and significant, which is intuitive but not as informative as the results concerning the variable $CLDEV$. In particular, it is hardly convincing that the higher coefficient is associated to $\hat{\beta}_0$ which, in turn, is rarely significant. The overall goodness of the model measured by $R^2$ and by the AIC index confirm that our variable $CLDEV$ better captures the relationship between students' achievements and grading practices. Indeed, the AIC statistics is

| | Estimate | Std. Error | t value | Pr(>\|t\|) | Signif. |
|---|---|---|---|---|---|
| (Intercept) | 1.6960 | 0.2877 | 5.90 | 0.0000 | *** |
| GENDER (female) | 0.2597 | 0.0487 | 5.34 | 0.0000 | *** |
| M03ITA | 0.3511 | 0.0252 | 13.91 | 0.0000 | *** |
| NCLASS | 0.0043 | 0.0069 | 0.62 | 0.5336 | |
| NPROFITA (2) | −0.0549 | 0.0616 | −0.89 | 0.3727 | |
| NPROFITA ($\geq 3$) | −0.1867 | 0.0775 | −2.41 | 0.0161 | * |
| CHANGECL (no) | 0.3718 | 0.1588 | 2.34 | 0.0194 | * |
| BOOKS (11-25) | 0.2298 | 0.1691 | 1.36 | 0.1745 | |
| BOOKS (26-100) | 0.3866 | 0.1594 | 2.43 | 0.0154 | * |
| BOOKS (101-200) | 0.5327 | 0.1605 | 3.32 | 0.0009 | *** |
| BOOKS (> 200) | 0.6717 | 0.1596 | 4.21 | 0.0000 | *** |
| $\hat{\beta}_0$ | −0.6031 | 0.0413 | −14.59 | 0.0000 | *** |
| $\hat{\beta}_1$ | −0.3331 | 0.0313 | −10.64 | 0.0000 | *** |

Signif. codes: *** 0.001 ** 0.01 * 0.05 . 0.1

Multiple R-Squared: 0.3844, Adjusted R-squared: 0.3782, AIC = 3068.101

Table 5: Education production function estimated as in equation (3) for scores in "Italian" test

given by

$$AIC = -2 \log \text{likelihood}(\theta) + 2 \dim(\theta)$$

where $\theta$ is the vector of coefficients and $\dim(\theta)$ is the number of parameters to be estimated. In model (1) we have, due to $CLDEV$, 13 parameters more than in model (3): this notwithstanding, the AIC has a lower value.

Incidentally, model (1) and model (3) provide similar results: harder grading standards are associated to higher achievement levels. Our main point is: do the linear grading equations and the $CLDEV$ variable also provide the same information about the grading practices in the 77 classes? Consider Figure 3: in the top plot the estimated intercepts $\hat{\beta}_0$ (horizontal axis) and the 15 $CLDEV$ groups (vertical axis) are plotted. Were the two variables giving the same information, we should observe some kind of correlation: more precisely, classes belonging to the same $CLDEV$ group should appear in the same subinterval of the $\hat{\beta}_0$ axis. This is clearly false and due to the linear approximation of the grading equation. Consider, for instance, group

9 in $CLDEV$: the grading standards are clearly low and the parameter associated to the group in model (1) is negative. The estimated values of $\hat{\beta}_0$ for the six classes belonging to group 9 go from -0.615 to 2.285 and only the highest value is statistically significant. A similar representation is given in the bottom plot of Figure 3: the estimated slopes $\hat{\beta}_1$ are plotted against $CLDEV$: again no correlation can be noted and the same paradoxical evidence can be pointed out: the estimated $\hat{\beta}_1$ for classes in group 9 go from 0.263 to 6.661 and only this last value is significant.

Actually, $\beta_0$ and $\beta_1$ jointly represent the stringency of the grading practice: in fact, a teacher can be more selective both assigning lower grades to each achievement level (lower $\beta_0$) and increasing grades at a lower rate when achievement increases (lower $\beta_1$). Hence $\beta_0$ and $\beta_1$ should be used jointly as grading practice indicators. To take it into account, a cluster algorithm has been applied to the 77 classes, using the estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ values as clustering variables: we have obtained 15 groups of classes, as similar as possible according to the estimates intercept and slope of their linear grading equation. In Figure 4 $CLDEV$ is plotted against groups generated by cluster analysis on the coefficients $(\hat{\beta}_0, \hat{\beta}_1)$: also in this case, no kind of correlation between the two variables can be found, letting us conclude that the grading standard representations obtained via the linear grading equations or via $CLDEV$ cannot be considered substitutes.

# 5 Conclusion

Our analysis confirms that teachers' grading practices do have an impact on students' achievement: higher grading standards are usually associated to better achievements.

At the same time, we have shown that the relationship between achievement and grades is frequently non linear, while the literature preferably uses linear grading equations to estimate how a grading practice is severe.

We suggest a non parametrical criterion to describe the stringency of a grading practice and, through simple classification techniques, give a representation of the variety of grading practices adopted by the teachers and evaluate their effect on students' achievement. The comparison with the linear regression analysis allow us to point out some misleading results of the linear regression.

Therefore, even if our results basically agree with the previous studies, we propose this method as more reliable for at least two reasons:

i) grades are not a linear function of the achievement. Hence, when we force the relationship to be linear, we cannot be confident that the impact we estimate on students' achievement is not a simple statistic effect;

ii) a variety of grading strategies exists, which cannot be easily ordered. Our method allows for a representation of all the grading practices, and examines the effect of each of them, without forcing them in an order.

# References


[1] Betts J.R. (1997), *Do grading standards affect the incentive to learn?*, University of California at San Diego, Department of Economics, Discussion Paper 97-22.

[2] Betts J.R.(1998), "The impact of educational standards on the level and distribution of earnings", *American Economic Review*, 88, 266-275.

[3] Betts J.R. - Grogger J. (2003), "The impact of grading standards on student achievement, educational attainment, and entry-level earnings", *Economics of Education Review*, 22, 343-352.

[4] Bonesronning H. (1999), "The variation in teachers' grading practices: causes and consequences", *Economics of Education Review*, 18, 89-105.

[5] Bonesronning H. (2004a), "Do the teachers' grading practices affect student achievement?", *Education Economics*, 12, 151-167.

[6] Bonesronning H. (2004b), "Can effective teacher behavior be identified?", *Economics of Education Review*, 23, 237-247.

[7] Bonesronning H. (2004c), *Do teachers favor girls?*, paper presented at First Network Workshop of the RTN "Economics of Education and Education Policy in Europe", Amsterdam, 7-8 November 2004.

[8] Cleveland W.S. - Grosse E. - Shyu W.M. (1992), "Local regression models", in J.M. Chambers - T.J. Hastie (eds), *Statistical Models in S*, Monterey: Wadsworth and Brooks-Cole.

[9] Correa H. - Gruver G.W. (1987), "Teacher-student interaction: a game theoretic extension of the economic theory of education", *Mathematical Social Science*, 13, 19-47.

[10] Costrell R.M. (1994), "A simple model of educational standards", *American Economic Review*, 84, 956-971.

[11] Figlio D.N. - Lucas M.E. (2004), "Do high grading standards affect student performance?", *Journal of Public Economics*, 88, 1815-1834.

[12] Iacus S.M. - Porro G. (2006), "Random Recursive Partitioning: a matching method for the estimation of the average treatment effect", to appear in *Journal of Applied Econometrics*. Preliminary version at http://services.bepress.com/unimi/economics/art9.

[13] Iacus S.M. - Porro G. (2007), "Missing data imputation, matching and other applications of Random Recursive Partitioning", to appear in *Computational Statistics & Data Analysis*. Preliminary version at http://services.bepress.com/unimi/statistics/art7.

[14] Lillard D.R. - DeCicca P.P. (2001), "Higher standards, more dropouts? Evidence within and across time", *Economics of Education Review*, 20, 459-473.

# Appendix

In order to show that identification of grading practices and their effects on student's achievement is independent of the definition of the *a priori* reference grading practice, we run the same analysis of Section 4 for grades and scores in Mathematics on the same 77 classes of students. We will not go into details but just outline the steps of the analysis. Figure 5 shows that grading practices are non linear, as it was for the Italian data set (see Figure 1). Hence, for each class we calculate the discrepancy with respect to the reference class, in order to identify grading practices and run a cluster analysis on them. The analysis of the dendrogram (omitted here) suggests to gather the grading practices in 14 groups shown in Table 6. Notice that for this data set no actual class has zero discrepancy with respect to the standard reference class, but this fact does not affect the results of the analysis that follows: the only requirement is, in fact, to have a reference grading practice. We also estimate the following linear grading equation for each class in order to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$MAT05FST_i = \beta_{0i} + \beta_{1i} SCMAT05_i, \quad i \in C_j, \tag{4}$$

where $MAT05FST_i$ are the grades in Mathematics for the students in class $C_j$, $SCMAT05_i$ are the test scores in Mathematics for the same students. Clustering of classes by $(\hat{\beta}_0, \hat{\beta}_1)$ and $CLDEV$ defines different and non correlated groups as for the Italian data set. We then proceed with the estimation

of the two models:

$$\begin{aligned} M05MAT_i = &\alpha_0 + \alpha_1 GENDER_i + \alpha_2 M03MAT_i + +\alpha_2 NCLASS_i \\ &+ \alpha_3 NPROFITA_i + \alpha_4 CHANGECL_i + \alpha_5 BOOKS_i \\ &+ \alpha_6 CLDEV_i \end{aligned} \tag{5}$$

and

$$\begin{aligned} M05MAT_i = &\alpha_0 + \alpha_1 GENDER_i + \alpha_2 M03MAT_i + +\alpha_2 NCLASS_i \\ &+ \alpha_3 NPROFITA_i + \alpha_4 CHANGECL_i + \alpha_5 BOOKS_i \\ &+ \alpha_6 \hat{\beta}_{0i} + \alpha_7 \hat{\beta}_{1i} \end{aligned} \tag{6}$$

where $M05MAT$ and $M03MAT$ are the Rash converted scores for the test in Mathematics. The summary of model (5) is contained in Table 7 and that of model (6) is in Table 8. The evidence from the econometric analysis confirms the conclusions of Section 4 on the Italian data set.

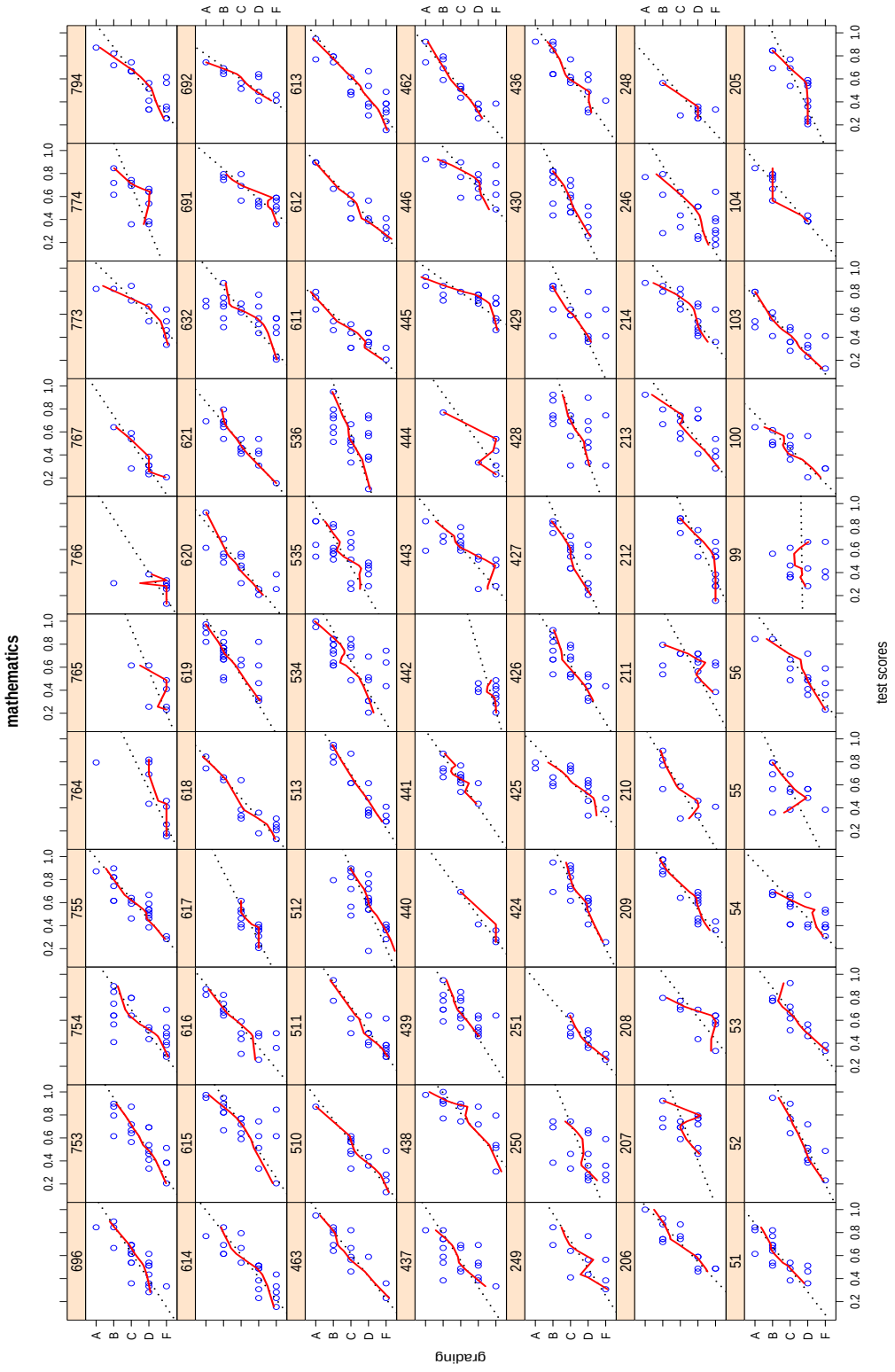


Figure 5: The non linear behaviour of grading practices (data set "Mathematics")

| | F | D | C | B | A |
|---|---|---|---|---|---|
| Group 1 | | | | | |
| 206 | 0 | 1 | 0 | 0 | 0 |
| 214 | 0 | 1 | 0 | 0 | 1 |
| 424 | 0 | 1 | 0 | 0 | |
| 513 | 0 | 1 | 0 | 0 | |
| Group 2 | | | | | |
| 52 | 0 | 1 | 0 | −1 | |
| 209 | 0 | 1 | 1 | −1 | |
| 438 | 0 | −1 | −1 | −1 | 0 |
| Group 3 | | | | | |
| 53 | 0 | 1 | 1 | 1 | |
| 205 | | 1 | 1 | 1 | |
| 426 | 0 | 1 | 1 | 1 | |
| 428 | 0 | 1 | 1 | 1 | |
| 429 | 0 | 1 | 1 | 1 | |
| 436 | 0 | 1 | 1 | 1 | 0 |
| 440 | 0 | 1 | 1 | | |
| 441 | | 1 | 1 | 1 | |
| 463 | 0 | 1 | 1 | 1 | 0 |
| 615 | 0 | 1 | 1 | 0 | 0 |
| 619 | | 1 | 1 | 1 | 0 |
| 691 | 0 | 1 | 1 | 1 | |
| 753 | 0 | 1 | 1 | 1 | |
| 765 | 0 | 1 | 1 | | |
| 774 | | 1 | 1 | 1 | |
| Group 4 | | | | | |
| 56 | 0 | 1 | 2 | 0 | 1 |
| 427 | | 1 | 2 | 0 | |
| 511 | 0 | 1 | 2 | 0 | |
| 696 | 0 | 1 | 2 | 0 | 1 |
| Group 5 | | | | | |
| 99 | 0 | 1 | 2 | 3 | |
| 246 | 0 | 1 | 2 | 3 | 2 |
| 248 | 0 | 1 | 2 | 3 | |
| 611 | 0 | 1 | 2 | 3 | 2 |
| 620 | 0 | 1 | 2 | 3 | 2 |
| 766 | 0 | 1 | | 3 | |
| Group 6 | | | | | |
| 51 | | 1 | 2 | 2 | 2 |
| 54 | 0 | 1 | 2 | 2 | |
| 55 | 0 | 1 | 2 | 2 | |
| 250 | 0 | 1 | 2 | 2 | |
| 535 | | 1 | 2 | 2 | 2 |
| 618 | 0 | 1 | 2 | 2 | 2 |
| 692 | 0 | 1 | 2 | 2 | 2 |
| 767 | 0 | 1 | 2 | 2 | |

| | F | D | C | B | A |
|---|---|---|---|---|---|
| Group 7 | | | | | |
| 104 | | 1 | | 1 | 1 |
| 210 | 0 | 1 | 2 | 1 | |
| 249 | 0 | 1 | 2 | 1 | |
| 251 | 0 | 1 | 2 | | |
| 430 | | 1 | 2 | 1 | |
| 442 | 0 | 1 | | | |
| 444 | 0 | 1 | | 1 | |
| 462 | 0 | 1 | 2 | 1 | 1 |
| 510 | 0 | 1 | 2 | | 1 |
| 536 | | 1 | 2 | 1 | |
| 613 | 0 | 1 | 2 | 1 | 1 |
| 616 | 0 | 1 | 2 | 1 | 1 |
| 617 | | 1 | 2 | | |
| 755 | 0 | 1 | 2 | 1 | 1 |
| 794 | 0 | 1 | 1 | 1 | 1 |
| Group 8 | | | | | |
| 207 | | 0 | 1 | 1 | |
| 213 | 0 | 0 | 1 | 1 | 0 |
| Group 9 | | | | | |
| 100 | 0 | 1 | 2 | 3 | 3 |
| 103 | 0 | 1 | 2 | 3 | 3 |
| 621 | 0 | 1 | 2 | 2 | 3 |
| Group 10 | | | | | |
| 208 | 0 | 1 | 0 | 1 | |
| 425 | 0 | 1 | 0 | 2 | 2 |
| 754 | 0 | 1 | 0 | 2 | |
| Group 11 | | | | | |
| 211 | 0 | 0 | 0 | 2 | |
| 212 | 0 | 0 | −1 | | |
| 512 | 0 | 0 | 0 | 1 | |
| 764 | 0 | 0 | | | 2 |
| 773 | 0 | 0 | 0 | 0 | 1 |
| Group 12 | | | | | |
| 437 | 0 | 1 | 2 | 2 | 1 |
| 612 | 0 | 1 | 2 | 2 | 1 |
| Group 13 | | | | | |
| 439 | −1 | 1 | 1 | 1 | |
| 445 | −1 | −1 | 0 | 1 | 1 |
| 446 | −1 | −1 | 0 | 0 | 0 |
| 534 | −1 | 1 | 0 | 1 | 0 |
| Group 14 | | | | | |
| 443 | 0 | 1 | 1 | 2 | 2 |
| 614 | 0 | 1 | 1 | 2 | 2 |
| 632 | 0 | 1 | 1 | 2 | 3 |

Table 6: The 14 groups of grading practices identified by our method and corresponding to the items of the categorical variable $CLDEV$ (data set "Mathematics")

| | Estimate | Std. Error | t value | Pr(>\|t\|) | Signif. |
|---|---|---|---|---|---|
| (Intercept) | 1.0345 | 0.3318 | 3.12 | 0.0019 | ** |
| GENDER (female) | 0.0443 | 0.0554 | 0.80 | 0.4242 | |
| M03MAT | 0.5560 | 0.0270 | 20.56 | 0.0000 | *** |
| NCLASS | 0.0172 | 0.0079 | 2.17 | 0.0302 | * |
| NPROFITA (2) | 0.0073 | 0.0773 | 0.09 | 0.9245 | |
| NPROFITA ($\geq$ 3) | −0.0378 | 0.0916 | −0.41 | 0.6802 | |
| CHANGECL (no) | 0.4166 | 0.1889 | 2.21 | 0.0276 | * |
| BOOKS (11-25) | −0.0810 | 0.1888 | −0.43 | 0.6680 | |
| BOOKS (26-100) | 0.1809 | 0.1782 | 1.01 | 0.3103 | |
| BOOKS (101-200) | 0.3318 | 0.1800 | 1.84 | 0.0656 | . |
| BOOKS (> 200) | 0.5575 | 0.1789 | 3.12 | 0.0019 | ** |
| CLDEV 2 | 0.4413 | 0.1880 | 2.35 | 0.0190 | * |
| CLDEV 3 | −0.1320 | 0.1366 | −0.97 | 0.3343 | |
| CLDEV 4 | −0.6257 | 0.1653 | −3.78 | 0.0002 | *** |
| CLDEV 5 | −0.9046 | 0.1634 | −5.54 | 0.0000 | *** |
| CLDEV 6 | −0.6213 | 0.1492 | −4.16 | 0.0000 | *** |
| CLDEV 7 | −0.6499 | 0.1400 | −4.64 | 0.0000 | *** |
| CLDEV 8 | 0.2170 | 0.2249 | 0.96 | 0.3347 | |
| CLDEV 9 | −0.8678 | 0.1862 | −4.66 | 0.0000 | *** |
| CLDEV 10 | −0.4612 | 0.1860 | −2.48 | 0.0133 | * |
| CLDEV 11 | −0.1665 | 0.1692 | −0.98 | 0.3254 | |
| CLDEV 12 | −0.6756 | 0.2066 | −3.27 | 0.0011 | ** |
| CLDEV 13 | 0.3317 | 0.1637 | 2.03 | 0.0430 | * |
| CLDEV 14 | −0.5460 | 0.1752 | −3.12 | 0.0019 | ** |

Signif. codes: *** 0.001 ** 0.01 * 0.05 . 0.1
Multiple R-Squared: 0.4209, Adjusted R-squared: 0.4096, AIC = 3319.769

Table 7: Education production function estimated as in equation (5) for scores in "Mathematics" test

| | Estimate | Std. Error | t value | Pr(>\|t\|) | Signif. |
|---|---|---|---|---|---|
| (Intercept) | 1.9285 | 0.3500 | 5.51 | 0.0000 | *** |
| GENDER (female) | 0.0566 | 0.0571 | 0.99 | 0.3220 | |
| M03MAT | 0.5736 | 0.0276 | 20.78 | 0.0000 | *** |
| NCLASS | 0.0151 | 0.0079 | 1.91 | 0.0569 | . |
| NPROFITA (2) | 0.0073 | 0.0725 | 0.10 | 0.9201 | |
| NPROFITA ($\geq$ 3) | −0.1268 | 0.0911 | −1.39 | 0.1643 | |
| CHANGECL (no) | 0.4221 | 0.1945 | 2.17 | 0.0302 | * |
| BOOKS (11-25) | −0.0390 | 0.1947 | −0.20 | 0.8411 | |
| BOOKS (26-100) | 0.2311 | 0.1829 | 1.26 | 0.2065 | |
| BOOKS (101-200) | 0.4213 | 0.1844 | 2.28 | 0.0225 | * |
| BOOKS (> 200) | 0.6425 | 0.1826 | 3.52 | 0.0004 | *** |
| $\hat{\beta}_0$ | −0.4750 | 0.0529 | −8.97 | 0.0000 | *** |
| $\hat{\beta}_1$ | −0.2717 | 0.0357 | −7.61 | 0.0000 | *** |

Signif. codes: *** 0.001 ** 0.01 * 0.05 . 0.1
Multiple R-Squared: 0.3740, Adjusted R-squared: 0.3677, AIC = 3391.441

Table 8: Education production function estimated as in equation (6) for scores in "Mathematics" test