GRADING IN HETEROGENEOUS SCHOOLS

VALENTINO DARDANONI, SALVATORE MODICA, AND ALINE PENNISI

Abstract. This paper studies the relationship between students' cognitive ability and their school grades; and in particular, how the institutional context (e.g. nation-wide external exams) influences the informative value of grades as signals of cognitive competence.

In a simple abstract model of students' valuation we show that unless competence standards are set at above-school level or variation of competence across schools is low, students' competence valuation will be heterogeneous, with weaker schools inflating grades or flattening their dependence on competence, therefore reducing the information content and comparability of school grades.

The theoretical model is applied to data from the PISA 2003 survey in a sample of 5 countries, namely Australia, Germany, Italy, The Netherlands and the USA. According to our estimates, in Australia and the USA schools heterogeneity does not affect grading practices; in the other countries grades are inflated in weaker schools, uniformly in Germany and The Netherlands, to a larger extent for weaker students in Italy.

JEL Classification Number I21

Keywords Grades-vs-Competence, Schools' Heterogeneity, External Exams, PISA 2003.

1. Introduction

Evaluating students' cognitive achievements is key to support decisions not only of future employers, but also of parents, school and college boards, and policy makers. Indeed, the emphasis on educational skills as signals has shifted over

Date: November 6, 2007.

time, from the job market context where it was first recognized in the seventies (e.g. in the classic works by Arrow [2], Spence [19] and Stiglitz [20]), to the education policy context, where the central issue is whether educational services are performing well and whether the producers are accountable for it ([9, 7, 18, 23]), i.e. whether there are "transparent, consistent measures of progress towards the objectives" ([18] p.4).[1]

The *measurement* of achievements however, typically by cognitive tests, raises thorny problems, since no test is perfect, and repeated tests carry the risk that "only what gets measured gets done". These issues are at the center of a lively debate these days, especially in the light of the widespread interest raised by the OECD Programme for International Student Assessment (PISA) study comparing students' achievements in various countries (see `www.pisa.oecd.org`). However, as we learn from [23], the idea of measuring students' abilities, knowledge and competence is almost 150 years old. British legislation for school funding included a test-based system of 'payments for results' in 1862; in 1890 the system was dismantled precisely because the opponents argued that owing to the two problems above the system's disadvantages outweighed its benefits. Most of current work on test design is devoted to counter these objections.

School *grades*, on the other hand, are costless, abundant, frequent, and population-wide; but to be useful they should accurately reflect underlying competence,[2] since the lower their information content, the higher the signaling noise generated by the sender and the de-codification costs incurred by the receiver.

[1] Evaluating students' cognitive achievement is also key to asses the impact of education on economic growth, since it has been argued that students' knowledge and not simply years of education is the relevant variable that influences growth, see e.g. Barro [3] and Hanushek and Wößmann [8].

[2] In addition to skills, grades often incorporate information on student effort and behaviour.

Grade variation over *time* has received much attention, under the broad heading of 'grade inflation'. Grade inflation typically refers to the increase over time in the grades given to students at any given level of achievement. See e.g. Jacob [10], Bas and Van Der Ploeg [9] and RAND Education [13].

The present paper explores how a country's educational system affects the way grading policy varies across *schools*, at a given time. In particular, we first develop a simple theoretical model which investigates how a school's grading policy may depend on the distribution of competence of its own students, such as for example when teachers "grade on a curve", so that weaker schools tend to grant higher grades for given level of achievement. The theoretical model investigates the relationship between school's evaluations and actual competence as a function of the distribution of within-school competence, and yields four possible classes of grading policies, each corresponding to a possible institutional scenario which holds above school level (educational system).

The econometric model derived from theory is then estimated for five countries, namely Australia, Germany, Italy, The Netherlands and USA, using the OECD PISA 2003 test scores and the information reported in the students' questionnaire on school grades. Using a probit model we obtain, for each country, an estimate of the relationship between the teachers' evaluation of students' competence and its level as measured by PISA test scores, and how this relationship depends on the mean and variance of students' competence within each school.

We find that each country (with the possible exception of the USA, see below) corresponds rather closely to one of the four institutional settings that we identify. In particular, in the Australian educational system grading policies do not seem to vary significantly at the school level, while in Germany, The Netherlands and Italy there seem to be a rather substantial dependence, for a given level of students' competence, between school grades and school characteristics. However this dependence, as it will be explained later, seems quite different in Germany and The Netherlands as compared to Italy; while in Germany and the Netherlands both strong and weak students coming from weaker schools achieve uniformly higher grades, in Italy the difference in grades between schools is wider for weak students. Finally, although in the USA differences in grading policies across schools are not significant at the usual statistical levels, there seems to be enough difference in the estimates which calls for further attention.

The PISA dataset gives us a unique opportunity to get information on both students' competence and actual grades. To the best of our knowledge, investigating how the relationship between grading and actual competence varies at the school level —and how it depends on above-schools institutional settings— is rather novel, and we believe that our model and empirical results may help to shed light on the connections between school heterogeneity, institutional settings and the informational content of school grades.

In the next section we present the theoretical model and results. Section 3 contains the empirical results, and section 4 concludes with some policy reflections.

2. Theory

Our goal is to study the relationship between cognitive competence and school grades, in different institutional contexts. To this end we must take into account the fact that teachers' evaluations may differ across schools within each country.

In a given institutional context $c = 1, \ldots, C$ (we will use 'institutional context' and 'country' interchangeably) there are S_c schools, and in each school $s_c = 1, \ldots, S_c$ there are $n(s_c)$ students with competence levels $x_1, \ldots, x_{n(s_c)}$, which we assume to be independent real random variables extracted from some school-dependent cdf F_{s_c}. Competence levels at a point in time are the result of underlying characteristics such as innate ability, effort, discipline and health of the students.

The central assumption of the paper is that, in terms of grading policies, within each institutional context schools differ *only* in the distribution of students' competence. Thus a school s is identified with F_s.

Teachers in school s must choose a (possibly s-dependent) valuation $y_i \in \mathbb{R}$ to students with competence x_i, where of course better students should get higher valuations.[3] Thus, the teachers' problem in a given school in a given country is to choose an increasing map $y = v_s(x)$ which is to be used in their school.

Remark. Assuming that s is identified with F_s implies that, within each institutional context, v_s depends on s only through F_s. In concrete terms, we are assuming that in any given country, a school with a given distribution of students' competence uses the same valuation function regardless of whether, for example, it is public or private, religious or non-confessional, urban or rural, or more or less effective as producer of educational service.

2.1. **Heterogeneity within Countries.** We start by considering how grading policies vary between schools operating in the same institutional context. In this discussion the subscript c from s_c is omitted for simplicity.

To model the fact that usually the main issue in grading within schools is what to do with the weak and the strong students, we assume the existence of external constraints which take the form of two reference competence levels $x_c^-(s) < x_c^+(s)$, a low and a high one —which may depend both on the given school and on the country where it operates— for which grades must be fixed at $y_c^- < y_c^+$.

[3] Valuations y_i's are ultimately mapped into an actual grade which, depending on the country in which the school operates, typically belongs to a set of ordered categories.

Assumption 1. *School-specific valuations v_s must satisfy the following constraints:*

$$v_s(x_c^-(s)) = y_c^- \,, \quad v_s(x_c^+(s)) = y_c^+ \,. \tag{1}$$

For example, y^- may denote the minimum valuation required for the pass grade, while y^+ may be the minimum valuation required for some higher grade. The two reference competence levels $x_c^-(s), x_c^+(s)$ may be for example quantiles, or may be fixed independently of school parameters.

We normalize students' competence level so that in each country it has zero mean and unit standard deviation, and assume that in each school s the low reference competence is below the national average (i.e. $x_c^-(s) < 0$) and the high reference competence is above it ($x_c^+(s) > 0$).[4]

Regarding the choice of v_s we wish to formalize the idea that teachers, when choosing v_s, are constrained by students' perception of unfairness on their part, so that students' relative evaluations must be related to their relative competence. This can be modeled as the requirement that given any two students with competence levels x and x', the difference in their valuations $v_s(x) - v_s(x')$ must be nondecreasing in $x - x'$:

Assumption 2. *$v_s(x) - v_s(x')$ is nondecreasing in $x - x'$ for all $x, x' \in \Re$.*

Our first result is the following:

Proposition 1. *Under Assumptions 1 and 2, there are school-dependent intercept $\alpha(s)$ and slope $\beta(s) > 0$ such that*

$$v_s(x) = \alpha(s) + \beta(s)\, x \tag{2}$$

[4] That the lower and upper tails of the competence distribution have independent non-negligible effects on economic growth has been recently discussed by Hanushek and Wößmann [8].

where $\alpha(s)$ and $\beta(s)$ are given by

$$\alpha(s) = \frac{x_c^+(s)y_c^- - x_c^-(s)y_c^+}{x_c^+(s) - x_c^-(s)}, \ \beta(s) = \frac{y_c^+ - y_c^-}{x_c^+(s) - x_c^-(s)} \,. \tag{3}$$

Proof. For arbitrary $x, x' \in \Re$, since $x - x' = (x - x') - 0$ the assumption implies

$$v_s(x - x') - v_s(0) = v_s(x) - v_s(x') \,.$$

Let now $z = -x'$ and use the above equation twice to obtain $v_s(x + z) - v_s(0) = v_s(x) - v_s(0 - z) = v_s(x) + v_s(z) - 2v_s(0)$, that is

$$v_s(x + z) + v_s(0) = v_s(x) + v_s(z) \,.$$

Letting $f(x) = v_s(x) - v_s(0)$, one then has

$$f(x + y) = f(x) + f(y) \,.$$

This is a Cauchy equation, whose only increasing solution is $f(x) = cx$ for some $c > 0$ (Aczel [1], Theorem 1 page 34). Hence $v_s(x)$ is linear as claimed. The constraints in equation (1) can now be used to give a system of two linear equations into two unknowns which can be solved as claimed. □

Remark. It may be interesting to notice that the linearity of the valuation function can be interpreted directly as the result of the teachers' minimization of students' perception of unfairness on their part subject to the external constraints given by Assumption 1. To this end, notice that at each x 'unfairness' can be taken as increasing in the distance between v_s' and 1 (prime denoting derivative), since when $v_s'(x) \neq 1$ differences in ability are not matched by differences in valuations. Taking a quadratic loss for simplicity, and assuming that teachers worry about valuations in the $[x^-(s), x^+(s)]$ interval, the teacher's problem is then

$$\min_{v_s} \int_{x^-(s)}^{x^+(s)} (v_s'(x) - 1)^2 \, dx \tag{4}$$

$$\text{subject to} \quad v_s(x^-(s)) = y_c^-, \; v_s(x^+(s)) = y_c^+ \,. \tag{5}$$

It easily follows that the solution to this problem is linear.[5]

Linearity of the valuation function in students' competence levels implies that, within each country, schools' heterogeneity affects the valuation process only through the intercept and slope parameters $\alpha(s)$ and $\beta(s)$. The latter depend on c via x_c^- and x_c^+, so the next step is to investigate how valuation depends on s in different types of institutional contexts.

2.2. **Heterogeneity of Countries.** An institutional context is characterized by the constraints which determine the two reference points $(x_c^-(s), y_c^-)$ and $(x_c^+(s), y_c^+)$ for each school. Recall that the values y_c^- and y_c^+ have been assumed school-independent; on the other hand, even within the same country, each school may be characterized by quite different distribution of students' competence, so $x^-(s)$ and $x^+(s)$ in principle may vary across schools. Thus, in our model there are four possibilites which describe different institutional scenarios: i) $x^-(s)$ and $x^+(s)$ are both s-independent; ii) $x^-(s)$ and $x^+(s)$ are both s-dependent; iii) $x^+(s)$ is s-dependent; iv) $x^-(s)$ is s-dependent. In details, the four institutional settings can be described as follows:

[5] It is an elementary problem in the calculus of variations, with Euler equation $v_s'' = 0$. See e.g. Kamien-Schwartz [11].

[A] *Absolute Valuation.* Grades follow common procedures at above-school level. In this case, constraints on school-level grading amount to setting a common scale, that is, $x_c^-(s)$ and $x_c^+(s)$ are fixed independently of school, at $x_c^- < 0$ and $x_c^+ > 0$.

[R] *Relative Valuation.* The proportion of students below y^- and above y^+ –determined by probability levels p^-, p^+ respectively– is fixed above the school level. In terms of the constraints (5), this amounts to having $x_c^-(s)$ and $x_c^+(s)$ determined as the p^--th and p^+-th quantiles. This is equivalent to scenario A if competence distribution is invariant across schools; if on the other hand school populations are heterogeneous, quantiles will generally be lower the weaker the school population.

[AL] *Absolute Lower Bound.* In this case there is a minimum absolute acceptable level of competence required for the valuation y^-; on the other hand, the upper tail (valuations above y^+) is determined in relative terms within each school by p^+. The formal translation of this case implies x_c^- being school-independent, and $x_c^+(s)$ as being the p^+-th quantile in school s.

[RL] *Relative Lower Bound.* In this case in each school s there is a maximum acceptable fraction of failed students, but the high competence level is fixed in absolute terms. This implies that $x_c^-(s)$ is the p^--th quantile in school s while x_c^+ is fixed.

These specifications need not be necessarily determined by written rules; as we shall see, they may be inferred implicitly from analysis of teachers' behavior.

We come to the main purpose of this section, that is to study how the valuation function v_s varies across schools when c belongs to one of these institutional contexts. Given linearity this amounts to studying how, in each different setting, the intercept $\alpha(s)$ and slope $\beta(s)$ vary depending on the distribution of competence levels F_s.

We concentrate on the mean μ_s and standard deviation σ_s of F_s, assuming that higher moments have a negligible effect on the valuation function.[6] At this point it is convenient to simplify notation further: given identification of s with F_s and the latter with its first two moments (μ_s, σ_s), a school is effectively identified with a pair (μ, σ). In the sequel we shall then write $s = (\mu, \sigma)$.

Using now subscripts for partial derivatives we proceed under the following

Assumption 3. *Let $q^-(\mu,\sigma) < 0 < \; q^+(\mu,\sigma)$ be the p^--th and p^+-th quantiles of $F_{(\mu,\sigma)}$. Then*

$$(i)\ q^+_\mu = q^-_\mu > 0\,, \quad (ii)\ q^-_\sigma \le 0,\ q^+_\sigma \ge 0\,.$$

Recall that $q^-(s) < 0 < q^+(s)$ follows from our assumption that, in all schools, low and high reference competence levels are not above/below the national average, which has been normalized to zero. Assumption 3 says that an increase in average competence in a given school implies a uniform upward shift of the two reference quantiles; and the low (high) reference quantile does not increase (decrease) when the dispersion of competence levels in the school increases. This assumption holds for example in the special case where F_s belongs to a family of location-scale distributions with μ and σ as location and scale parameters.[7]

The implications of Assumption 3 are described in the next proposition.

[6] In fact, in our application even the second moment is usually not significant.

[7] Indeed, when F_s belongs to a family of location-scale distributions there is a fixed c.d.f. H such that the competence variable, for any school, is distributed as $H((x-\mu)/\sigma)$, so that if z is the p-th quantile, then $z = \mu + H^{-1}(p)\,\sigma$; since $q^-(s) < 0 < \; q^+(s)$ one has $H^{-1}(p^-) < 0 < H^{-1}(p^+)$, and the claim follows.

Proposition 2. *Under assumption 3, in the four scenarios* [A], [R], [RL], [AL] *the coefficients* α, β *defined in Proposition 1 satisfy:*

$$\begin{array}{ll} [\mathrm{A}]\ \alpha_\mu = \beta_\mu = \alpha_\sigma = \beta_\sigma = 0 & [\mathrm{R}]\ \alpha_\mu < 0,\ \beta_\mu = 0,\ \beta_\sigma \leq 0 \\ [\mathrm{AL}]\ \alpha_\mu < 0,\ \beta_\mu < 0,\ \alpha_\sigma \leq 0,\ \beta_\sigma \leq 0 & [\mathrm{RL}]\ \alpha_\mu < 0,\ \beta_\mu > 0,\ \alpha_\sigma \geq 0,\ \beta_\sigma \leq 0\,. \end{array}$$

Proof. We omit the c subscript, and write for example x^+_σ for $\partial x^+_c(\mu, \sigma)/\partial\sigma$. In case [A], α, β are independent of (μ, σ) since x^-, x^+ are. In case [R], x^- and x^+ are the p^--th and p^+-th quantiles of F_s, so $(x^+ - x^-)_\mu = 0$ by assumption 3 (i), whence $\beta_\mu = 0$; and $\alpha = y_+ - \beta x^+$, whence $\alpha_\mu = -\beta x^+_\mu < 0$. Finally, $\beta_\sigma = -\beta(x^+ - x^-)_\sigma/(x^+ - x^-) \leq 0$ by assumption 3 (ii). In case [AL], x^- is fixed and x^+ is the p^+-th quantile, so one easily checks that $\beta_\mu < 0$; and $\alpha_\mu = -\beta_\mu x^- < 0$ from $x^- < 0$. Also, $\beta_\sigma = -\beta x^+_\sigma/(x^+ - x^-) \leq 0$ from $x^+_\sigma \geq 0$; and then $\alpha_\sigma = -\beta_\sigma x^- \leq 0$ from $x^- < 0$. For case [RL] the argument is analogous to the one just given. □

The proposition implies that in institutional settings where [A] holds there is a homogeneous valuation across different school types. This is the benchmark, undistorted system. Its simplest implementation is identifiable with country-level curriculum-based external exit examinations, but we shall see this is not strictly necessary for nationwide standards to emerge. In the other cases, if variation in competence across schools is non-negligible, departure from absolute valuations implies that, within the same country, both the intercept and the slope of the valuation function may become school-specific; in these cases not only valuations in some schools may be uniformly inflated (intercept effect), but also in some schools the less capable students are over-evaluated and the strong ones penalized (slope effect). If these effects are substantial, the grading signal may become much less informative of the students' underlying ability. Quantitative estimates are given in section 3.3 (figure 2 page 19 illustrates).

3. Application to the PISA 2003 Survey

3.1. **From Theory to Estimable Equation.** The theory developed in the previous section provides the framework for empirical estimation. Proposition 1 shows that, under our assumptions, evaluation v_{is} of student i in school s is a linear function of her competence x_{is}, with school-specific slope and intercept. Taking a first order approximation of α and β with respect to school's mean and standard deviation μ_s and σ_s, and ignoring higher order moments, using again subscripts for partial derivatives for α and β, in each country the valuation of student i in school s in can be written as

$$v_{is} = a + bx_{is} + \alpha_\mu \mu_s + \alpha_\sigma \sigma_s + \beta_\mu \mu_s \, x_{is} + \beta_\sigma \sigma_s \, x_{is} \tag{6}$$

with $b > 0$, while the signs of the other coefficients depend on the institutional setting where schools operate as spelled out in Proposition 2.

Students evaluations may also depend on a vector of student-specific covariates, some of which may be observable by the econometrician while some other may represent residual unobservable heterogeneity. We then augment equation (6) as:

$$v_{is} = a + bx_{is} + \alpha_\mu \mu_s + \alpha_\sigma \sigma_s + \beta_\mu \mu_s \, x_{is} + \beta_\sigma \sigma_s \, x_{is} + \boldsymbol{\gamma}^{'} \boldsymbol{z}_{is} + \epsilon_{is} \tag{7}$$

where $\boldsymbol{z}_i$ is a vector of observable covariates and ϵ_{is} denotes an idiosyncratic error term. In practice, students true evaluations v_{is} are not observed, but can be considered as a continuous latent representation of the observed binary variable p_{is}, which takes value 1 if the student obtains a pass grade (that is, $v_{is} > \bar{v}$, where $\bar{v}$ denotes the school-independent valuation level necessary to pass), and 0 otherwise.

Under the assumption that ϵ_{is} has a standard normal distribution, it follows that the following probit equation holds

$$\Pr(p_{is} = 1 \mid x_{is}, \mu_s, \sigma_s, \boldsymbol{z}_{is}) = \Phi(a_0 + b x_{is} + \alpha_\mu \mu_s + \alpha_\sigma \sigma_s + \beta_\mu \mu_s \, x_{is} + \beta_\sigma \sigma_s \, x_{is} + \boldsymbol{\gamma}' \boldsymbol{z}_{is}), \tag{8}$$

where Φ denotes the standard normal link.[8] In our application, p denotes whether the student has obtained a pass grade in mathematics, x is the PISA measured mathematical competence score, μ and σ respectively measure the mean and standard deviation of PISA mathematical competence in each school, and $\boldsymbol{z}$ contains two covariates, namely student's gender and his/her socio-economic family background. With these specifications, the probit equation (8) is what we estimate. For estimation purposes the competence variable x is standardized in each country, as described in the theoretical model.

3.2. **Data Description and Discussion.** We now describe the data in more detail. We refer to the results of the 2003 survey of the OECD Programme for International Student Assessment (PISA), which focused on mathematics. Carried out every three years since 2000 in over 40 countries, PISA surveys 15-year old students' knowledge, skills and study environment.

In our empirical analysis competence is measured by (a rescaling of) the PISA scores in mathematics, which are originally scaled to an average of 500 points with 100 points standard deviation across OECD countries. As shown in table 1 the five countries present significant differences. In terms of average competence The Netherlands and Australia rank above the others; variance is higher than OECD average in all of them except The Netherlands.

[8] Note that the intercept a in equation (7) is not identifiable since $\bar{v}$ is not observed.

Table 1. Scores

Country	Average Score	Variance in % of OECD Average	Between-School Var., % of Total	% Students Above Pass
AUS	524 (2,1)	104.9	21,1	83,2 (0,7)
DEU	503 (3,3)	105.1	51,7	92,3 (0,6)
ITA	466 (3,1)	108.3	52,2	62,0 (1,0)
NLD	538 (3,1)	91.9	58	72,2 (1,2)
USA	483 (2,9)	106.5	25,7	87,9 (0,6)
OECD	500 (0,6)	100		

Source OECD. Estimates are provided with the corresponding standard error in brackets. For details on score assignment see the Technical Report [17].

Of more direct interest for our study is the between-schools variance —in terms of our model the variability of μ—, as opposed to the within school variance (between-school variance is a measure of how much the higher performing students are grouped together in the same schools and separated from the lower performing students). In particular, when differences between schools are small, Relative Valuation is closer to Absolute; as remarked on page 9, if in the limit between schools variance is zero, models A, R and RL coincide. Notice then that in the five countries we are considering there are large differences, with between-school variance being around 20-25% in Australia and the USA while in Germany, Italy and The Netherlands being over 50%.

It is instructive to look at the whole distribution of school mean μ besides its variance, and at its relation with within-school variability σ. Indeed, the upper and lower panels of Figure 1, referring to Italy and the Netherlands respectively, reveal two different pictures. The first presents a ‘normal’ bell-shaped distribution of μ with within-school variance increasing with school quality: there are relatively few good students in weak schools, but good and poor students alike populate the high performing ones. The same story emerges for the USA (figures not shown).

In the lower panel on the other hand, the bimodality of the mean distribution describes a system partitioned in two performance-based school clusters, a story reinforced by the fact that competence variability in the better schools is lower. Germany is similar to The Netherlands, and as we shall see in these two countries the difference between strong and weak schools in terms of grades are the most pronounced; this fact may have its roots in this 'duality' of their school systems. Finally, in Australia the histogram is bell-shaped and the regression line is slightly downward sloping.[9]

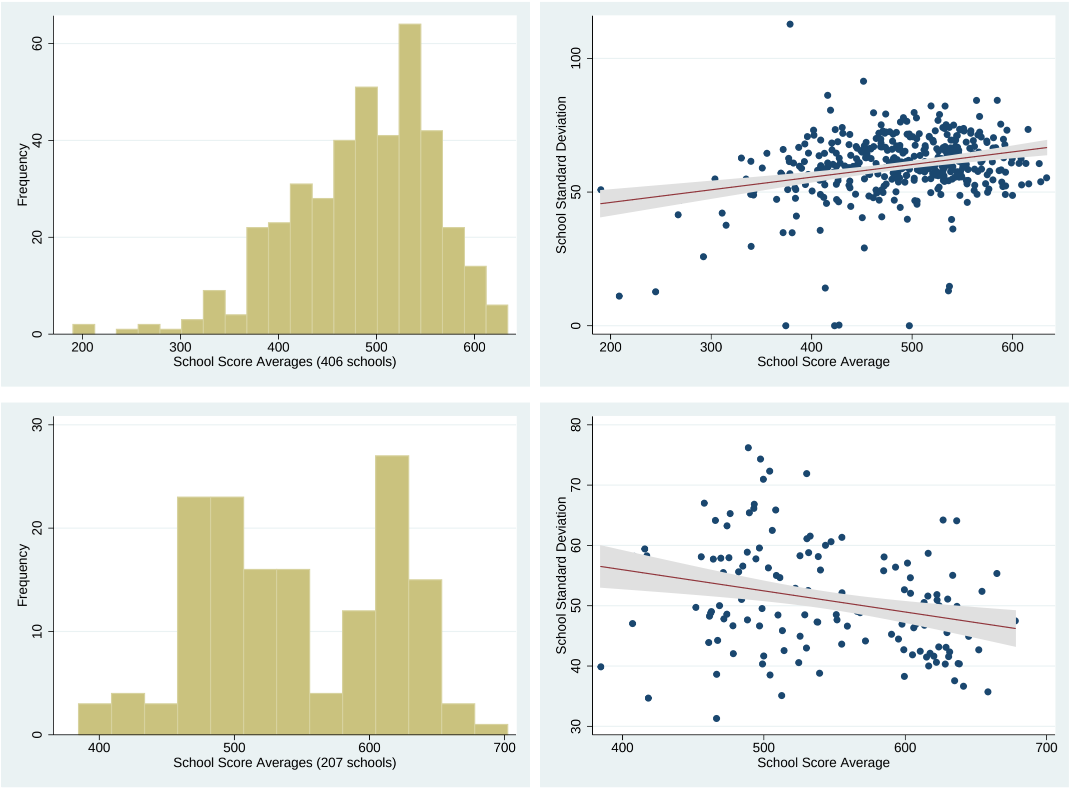

Figure 1. Schools Means and Std: Italy above, The Netherlands below

[9] The slope coefficients of the linear regressions of school standard deviation on school mean shown in the figure, and the others referring to Australia, Germany and the USA, are all significantly different from zero.

Regarding data on grading, countries could choose to administer several optional PISA questionnaires, among which the student's educational career questionnaire. We look at five countries which adopted the educational career questionnaire and asked students to answer the following question: "In last school report, how did your mark in mathematics compare with the pass mark?";[10] this is the dependent variable of our probit equations. As we see from table 1 also in this case there are non-trivial differences: in AUS, DEU and USA around 90% of the students are above the pass grade, while in ITA and NLD around 65%. This difference may be due to the different "grades' message space" as we may call it: in ITA and NLD the grade scale is between 1 and 10, with 1-5 being below pass; the others have a grading scale typically made of 5 or 6 different grades with typically the first two grades being below pass. Since the survey is conducted in April-May in all countries so that the 'last mark' is before the final quarter, teachers with a greater choice of below pass grades can send richer warning, work-stimulating messages.

Finally, the PISA socio-economic and cultural background index (SE) combines information on the occupational, educational and cultural environment of students' household. For details see e.g. [16].

3.3. **Estimation Results.** Our estimation of equation (8), whose results are contained in the table below, is carried out using the sample weights information given in the PISA study, and adjusting the standard errors of the estimates to take into account the cluster structure induced by the school level sampling. In particular, the reported estimates are obtained using STATA's survey probit weighted ML routine with robust linearized SE.

[10] Question Q7, variable EC07Q02

Table 2. Dependent Variable: Probability of Pass

		USA		**AUS**	
Variable	Coef.	Value	t	Value	t
x		.587	3.27	.597	4.61
μ	α_μ	.116	1.70	-.137	-2.22
$\mu\, x$	β_μ	.111	1.83	-.037	-0.74
σ	α_σ	-.112	-0.50	-.394	-1.64
$\sigma\, x$	β_σ	-.206	-1.03	-.197	-1.39
male		-.095	-1.98	-.033	-0.99
s-e backgr.		.116	3.88	.058	2.13
const.		1.380	7.18	1.389	6.53

		DEU		**NLD**		**ITA**	
Variable	Coef.	Value	t	Value	t	Value	t
x		.589	2.85	.997	4.18	.565	3.08
μ	α_μ	-.517	-5.91	-.642	-8.72	-.455	-10.23
$\mu\, x$	β_μ	.013	0.27	-.016	-0.32	.0730	2.56
σ	α_σ	-.178	-0.46	-.308	-0.73	1.000	3.95
$\sigma\, x$	β_σ	.011	0.04	-.634	-1.69	.147	0.55
male		.010	0.17	.149	2.90	-.422	-10.85
s-e backgr.		.070	2.04	-.018	-0.58	.108	5.22
const.		1.639	6.24	.739	2.79	-.165	-0.94

We first carried out a Wald test, in each country, for the hypothesis that $\alpha_\mu = \beta_\mu = \alpha_\sigma = \beta_\sigma = 0$, i.e. that grading conforms to the hypothesis [A]. This hypothesis is not rejected for USA and AUS with p-values respectively equal to 0.152 and 0.195. For the other countries, a glance at the table above reveals that in DEU and NLD, among school-specific competence parameters only α_μ has a significant (negative) sign, suggesting that in below-average-competence schools grades may be uniformly inflated. Therefore, grading practices in DEU and NLD are compatible with hypothesis [R]. On the other hand, in Italy there are strongly significant intercept and slope effects as a function of schools' competence heterogeneity, with signs (when significant) compatible with hypothesis [RL].

We now evaluate the quantitative impact of these distortions. To this end we may compare, for each country, the difference in valuation between a good and a poor school, at a low and at a high level of student performance. Thus, for our purpose, we have to identify the distribution of competence (μ and σ) in a typical strong and weak school, and two appropriate levels of competence which may be considered representative of good and poor students.

Given benchmark school chosen with μ_s and σ_s equal to their country average, strong and weak schools are taken with μ at the 75th and 25th percentiles respectively, with corresponding standard deviations adjusted along the regression line of σ on μ (cfr. figure 1).[11] Coming to performance, there are six PISA levels in Mathematics, as described in [16] (p.48); we have taken the threshold between the first and the second level —score 420— as low performance, and that between the fifth and sixth —score 670— as high.

We are now ready to formulate the following question: given the estimated valuation of a student scoring 420 in an average school, what is the competence score of a student who receives the same valuation in a good [resp. poor] school? The same question is then repeated for the 670 score. Intuitively, if schools evaluation are relative, in a good school it should take a higher score for any given valuation students have higher competence, so the good school line lies below the other. The results are in figure 2 (the average school lines are not shown), where the lines are drawn on the basis of the coefficients of the probit regression presented in table 2.

A glance at figure 2 reveals that in Germany and The Netherlands the difference in school grading is substantial; given the same teachers' evaluation, there is a full PISA-level difference in competence between good and bad schools, both at the

[11] Quantiles are calculated with the weighted `pctile` STATA command.

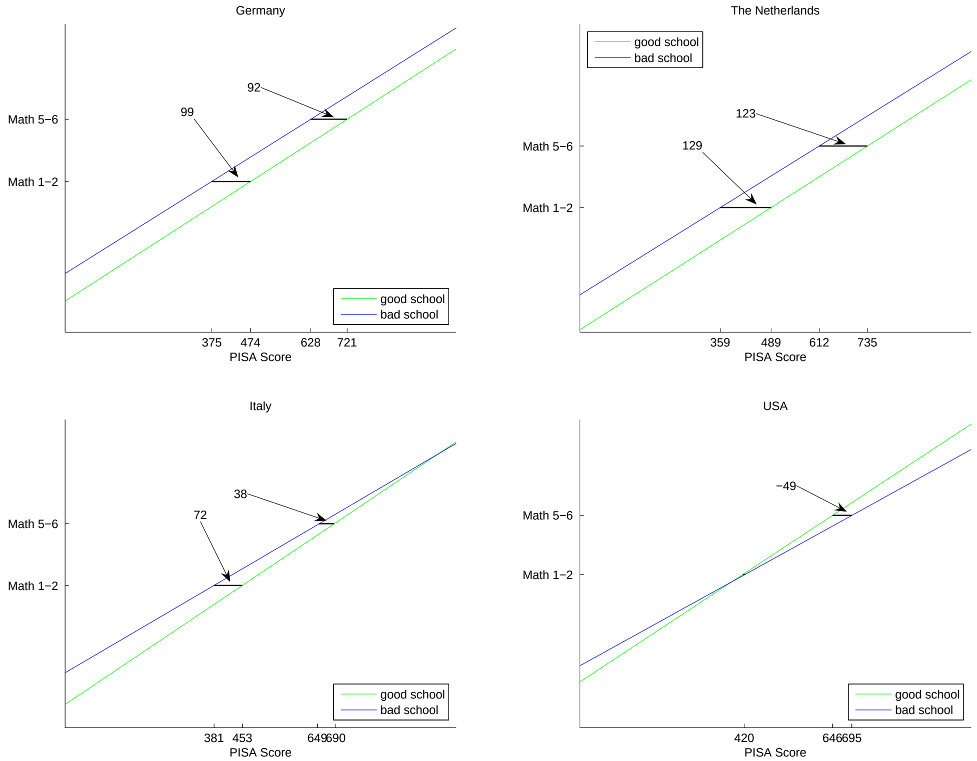

Figure 2. PISA Scores needed for valuation corresponding to thresholds between Math Levels 1-2 and 5-6 (420 and 670 in average school)

.

low and high end of the spectrum. We discussed these cases when describing figure 1 above.

In Italy, confirming the predictions of case [RL], the difference is more marked for low than for high levels, and in fact it is substantial (one PISA level) only for poorly performing students. In the case of Australia (not shown) the two lines essentially coincide.

The US system does not seem to fit our model very well, as it is clear from the figure. Technically this is due to the ‘wrong’ sign of the coefficient of μ (second line of table 2). Thus substantially there may be something our assumptions do not capture, and further investigation of this case would be needed.

We close this section by mentioning the gender and socio-economic background effects. In the PISA 2003 survey males tend to perform better than females in mathematics. Somewhat surprisingly, the results in table 2) say that given performance, in some countries male seem to be penalized in terms of grades, and sometimes substantially so. On the other hand, except in The Netherlands, students coming form higher socio-economic background apparently tend to receive higher grades for given level of competence.

3.4. **Theory and Facts.** We finally interpret these results in the light of the model predictions, which are that countries with absolute standards' requirements imposed at above-school level or low variance of average competence across schools should be in class [A], while under different constraints at above-school level (like a fixed percentage of failed students in each school) grading policy should depend on school parameters.

In the USA there are differences in funding, curricula, grading, and difficulty of secondary schools in the various States, and mandatory exit exams are present in 22 States (see Kober et al. [12]). Moreover, colleges and universities require applicants to submit scores from a Scholastic Achievement Test (SAT) which was introduced in 1901, where SAT scores are intended to supplement the secondary school record and help college admission officers to put local data –such as course work, grades, and class rank– in a national perspective. These consolidated elements of centralization, together with the low variance of school quality (variance of μ), produce a system where there seems to be no statistical evidence of school-dependent grading policy distortions.

In Australia the situation is similar to the USA, with exit exams standardized at state level in various degrees in the nine States (see Masters et al. [15]),[12] and variance of μ relatively low. This produces an institutional system where schools' heterogeneity do not seem to affect grading practices.

In The Netherlands the exit certification is based in equal parts on students' in-school performance and on students' result in the externally conducted semi-independent agency.[13] On the other hand the variability in schools' quality is very high and this appears to be the dominant effect, resulting in grades in weaker schools uniformly higher than those given in the better ones.

In Germany, as reported in [22], 7 of the 16 Landers have external exit exams at the end of secondary school, while the others exams are designed and graded on a local basis; and with one exception, conditional on socio-economic background all states with central exams outperform those without. Thus our results suggest that states without central exams tend to give higher grades.

Italy is the only country in our sample which falls in [RL]. In Italy exams at the end of secondary schools are designed at the national level, but grading is on a local basis, so the system is effectively decentralized (source in footnote 13). Weaker schools are located especially in the South, and the 'political' need not too fail too high a fraction of students from poorer areas accounts for fixation of a school-dependent $x^{-}(s)$, producing higher grades at the bottom end of the distribution. On the other hand there is a strong national cultural tradition, which apparently induces teachers to require high standards from the best students, who are then graded uniformly over the country. The resulting picture is [RL], with

[12] Some effort is being put towards national centralization, see Masters et al. [14].

[13] Our main source of information on school systems in Europe is Eurydice, an institutional information network focused on education systems and policies, established by the European Commission and Member States. See `http://www.eurydice.org/portal/page/portal/Eurydice`.

the detrimental consequence that the strong students from poorer areas are in the worst position to differentiate themselves from the others through grades.

4. Conclusions

This paper studies the informational value of school grades as a signal of underlying competence, in different institutional contexts. We spell out in a simple theoretical model four classes of systems which may produce distortions at the school level (such as when weaker schools grant higher grades at given skill levels). In the benchmark case, with competence standards fixed at system level, school grades reflect competence independently of school type. With different patterns of system behavior (e.g. not failing more than a given percentage of students), grades are usually inflated in weaker schools, uniformly or to a larger extent for weaker students.

The theoretical model is applied to data from the PISA 2003 survey in a sample of 5 countries, namely Australia, Germany, The Netherlands, Italy and the USA. According to our estimates, in Australia and the USA schools heterogeneity does not affect grading practices; in the other countries grades are inflated in weaker schools, uniformly in Germany and The Netherlands, to a larger extent for weaker students in Italy.

Implementing system-wide curriculum-based external exit examinations is of course a sufficient condition for system-wide competence standards. [14] According to our empirical estimates it may not be necessary. In the case of Australia for example, competence standards appear to be fixed at system (country) level, but

[14] Evidence on positive impact of CBEEE on competence is reported in Bishop [4, 5] and Wößmann [21, 22]. Bishop-Wößmann [6] also mention the link with the signalling of academic achievement. Kober et al. [12] warn of the possibility that nationwide standards be too high and raise the drop-out rate.

external exams are held sub-system (state) level. The USA, a decentralized system with elements of system-wide checking, also falls in the non-distorted class (albeit less sharply than Australia). In the other cases, the extent of distortion appears to depend on the variance of school quality and possibly on other characteristics of its distribution.

References

[1] Aczel, J. (1966): *Lectures on functional equations and their applications*, Academic Press

[2] Arrow, Kenneth J. (1973): “Higher education as a filter”, *Journal of Public Economics* **2**, pp. 193-216

[3] Barro, Robert J. (2001): “Human capital and growth”, *American Economic Review* **91** no.2 (May), pp. 12-17

[4] Bishop, John H. (1997): “The Effect of National Standards and Curriculum-Based Examinations on Achievement”, *American Economic Review* **87** no.2, pp. 260-264

[5] Bishop, John H. (2006): “Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn Externalities, Information Problems and Peer Pressure”, in Eric A. Hanushek, Finis Welch (eds.), *Handbook of the Economics of Education*, Vol. 2, Amsterdam, North-Holland, pp. 909-944

[6] Bishop, John H. and Ludger Wößmann (2004): “Institutional Effects in a Simple Model of Educational Production”, *Education Economics* **12** no.1, pp. 17-38

[7] Hanushek, Eric A. and Ludger Wößmann (2007): “The Role of School Improvement in Economic Development”, PEPG 07-01, Harvard University

[8] Hanushek, Eric A. and Ludger Wößmann (2007): “The Role of Educational Quality in Economic Development”, mimeo April

[9] Jacobs, Bas and Frederick van der Ploeg (2006): “Guide to reform of higher education: a European perspective”, *Economic Policy*, pp. 535-592

[10] Jacob, Brian (2007): ”Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments”, NBER Working Paper no.W12817

[11] Kamien, Morton and Nancy Schwartz (1981): *Dynamic Optimization*, Elsevier

[12] Kober, Nancy, Dalia Zabala, Naomi Chudowsky, Victor Chudowdsky, Keith Gayler, and Jennifer McMurrer (2006): " State High School Exit Exams: A Challenging Year", Center on Education Policy, Washington, D.C.

[13] Koretz, Daniel and Mark Berends (2001): "Changes in HIgh School Grading Standards in Mathematics, 1982–1992", *RAND Education*, Prepared for the College Entrance Examination Board

[14] Masters, Geoff, Margaret Forster, Gabrielle Matters and Jim Tognolini (2006): "Australian Certificate of Education: exploring a way forward", Australian Council for Educational Research

[15] Masters, Geoff and Gabrielle Matters (2007): "Year 12 Curriculum Content and Achievement Standards", Australian Council for Educational Research

[16] OECD (2004): "Learning for Tomorrow's World –First results from PISA 2003", OECD, Paris

[17] OECD (2005): PISA 2003 Technical Report, OECD, Paris

[18] Pritchett, Lant (2004): "Towards a New Consensus for Addressing the Global Challenge of the Lack of Education", *Copenhagen Consensus 2004 Challenge Paper*

[19] Spence, Michael (1973): "Job market signalling", *Quarterly Journal of Economics* **87**, pp. 355-74

[20] Stiglitz, Joseph E. (1975): "The theory of screening, education, and the distribution of income", *American Economic Review* **65**, pp. 283-300

[21] Wößmann, Ludger (2005): "The effect heterogeneity of central exams: Evidence from TIMSS, TIMSS-Repeat and PISA", *Education Economics* **13** no.2, pp. 143-169

[22] Wößmann, Ludger (2007): "Fundamental Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries", PEPG 07-02 Harvard University

[23] The World Bank (2004): *World Development Report 2004: Making services work for poor people*

Facoltà di Economia, Università di Palermo

E-mail address: vdardano@unipa.it

Facoltà di Economia, Università di Palermo

E-mail address: modica@unipa.it

Ministero dello Sviluppo Economico, Roma

E-mail address: aline.pennisi@tesoro.it