KATHOLIEKE UNIVERSITEIT LEUVEN

School accountability: (how) can we reward schools and avoid cream-skimming?
by

Erwin OOGHE
Erik SCHOKKAERT

Public Economics

November 2009

DISCUSSION PAPER

School accountability:

(how) can we reward schools and avoid cream-skimming?[*]

Erwin Ooghe and Erik Schokkaert[†]

December 2009

Abstract

Introducing school accountability may create incentives for efficiency. However, if the performance measure used does not correct for pupil characteristics, it will lead to an inequitable treatment of schools and create perverse incentives for cream-skimming. We apply the theory of fair allocation to show how to integrate empirical information about the educational production function in a coherent theoretical framework. The requirements of rewarding performance and correcting for pupil characteristics are incompatible if we want the funding scheme to be applicable for all educational production functions. However, we characterize an attractive subsidy scheme under specific restrictions on the educational production function. This subsidy scheme uses only information which can be controlled easily by the regulator. We show with Flemish data how the proposed funding scheme can be implemented. Correcting for pupil characteristics has a strong impact on the subsidies (and on the underlying performance ranking) of schools.

[†]Erwin Ooghe: Department of Economics, KULeuven, erwin.ooghe@econ.kuleuven.be. Erik Schokkaert: CORE, Université catholique de Louvain and Department of Economics, KULeuven, erik.schokkaert@econ.kuleuven.be.

1 Introduction

In many countries the funding of public schools used to be based largely on inputs. Local schools were subject to strict quality regulation and they had little autonomy in how to organize themselves. Moreover, parents had almost no freedom of choice. Consensus is growing that this archetypical system of public school financing does not create sufficient incentives for efficiency. Giving schools more autonomy may motivate teachers and school administrators and improve performance. Yet, increasing autonomy also means that one needs some standard for evaluating school performance. The UK introduced quasi-markets for education. In principle parent choice then becomes the main mechanism of control and report cards on school performance can give the parents useful information to choose a school for their children. In the USA, the "No Child Left Behind Act of 2001" imposed on the states the requirement to adopt an accountability system based on externally collected pupil test scores. In response US states had to set up a report card-system by which the performance of individual schools can be gauged against the performance of others. Some states went further and introduced in their funding system bonuses (sanctions) for well (poorly) performing schools.

There is a growing amount of empirical evidence suggesting that introducing school accountability indeed improves the performance of the pupils in terms of measured test scores, although it is still unclear whether explicit financial bonuses and penalties are necessary (Wössmann, 2003; Hanushek and Raymond, 2004, 2005; Jacob, 2005; Figlio and Rouse, 2006; West and Peterson, 2006; Burgess et al., 2007; Chiang, 2009). At the same time, it has also become clear that introducing accountability may induce a set of potentially undesirable strategic reactions, such as teaching to the rating, removal of low-achieving students from school, student retainment, even adapting the caloric content of the school lunches at the testing date (Jacob, 2005; Figlio and Winicki, 2005; Burgess et al., 2005; Reback, 2008). High powered incentives work, but their effects may be unexpected and depend on the specific design of the performance measurement scheme.

The design of an adequate performance measurement scheme therefore raises important conceptual issues. First, how to define what is the relevant output? Focusing on cognitive outcomes may lead to a relative neglect of non-cognitive factors. Focusing on one specific domain of knowledge (math) may lead to a relative neglect of other, untested, domains (history). Using a specific test may lead to schools preparing their pupils for this specific test to the detriment of the broader

2

knowledge that one is really aiming at. Using as a criterion the number of students that pass a predefined threshold level may lead to a concentration of efforts on those pupils that are at the margin of passing and to the neglect of very poor or extremely good performers. All this suggests that one should be explicit about the final objectives one wants to reach by introducing accountability and then adapt the measurement instruments to these objectives (Neal, 2008). Second, even if one agrees about the output measure, there still is the issue of how to measure school performance, i.e., the effect of school policies on the chosen output indicator(s). Schools can only be held responsible for those factors that are under their own control. In some cases their decision freedom may be restricted by the regulator. More importantly, the average test scores obtained heavily depend on the characteristics of their pupil population, both directly and indirectly through the peer group effect. Insufficiently correcting for pupil characteristics may lead to a very biased evaluation of school performance (Meyer, 1997; Ladd and Walsh, 2002; Hanushek and Raymond, 2003; Taylor and Nguyen, 2006; Neal, 2008). This biased evaluation may induce an inequitable and inefficient remuneration scheme. Moreover, it will give the schools incentives for cream skimming. Cream skimming is possible as soon as the quality of the pupils is related to characteristics that are observable for the school, such as socio-economic background or previous school results. By attracting students that are easier to educate, schools can improve their measured performance without increasing their real efficiency. For some pupils it might then become difficult to find an adequate school.

In this paper we focus on that second issue. Is it possible to devise a funding scheme that introduces incentives for better performance in terms of test scores without creating incentives for cream-skimming? How to correct observed test scores for those determinants which are not controlled by the schools themselves? To focus on this question, we take it for granted that the definition of the relevant outputs has been settled before. We also neglect the practical implementation issues, that have been discussed in the literature. We assume that the phenomenon of pupils moving from one school to another is taken into account in a satisfactory way. We neglect the fact that the measurement scheme may yield unreliable (in the sense of highly volatile) results for small schools, due to the limited number of observations (Kane and Staiger, 2002). Most importantly, we assume that sufficient data are available to calculate value-added, i.e., the gain in test scores, at the

level of the individual pupils. It is well known that informationally less demanding accountability schemes (based on the level or the difference in *average* test scores, or the gain in average scores of a cohort) will never be sufficient to correct for differences in the individual characteristics of the pupils (Meyer, 1997; Hanushek and Raymond, 2003). Starting from the most favourable informational assumptions, i.e., assuming that individual score gains can be calculated, allows us to focus on the basic conceptual issues.

Our model is formulated as a problem of how to link financial incentives to performance in terms of test scores. Such high-stakes testing is only one possible interpretation of the model, however, and our approach is also relevant in other settings. Consider an educational system in which parents can basically choose any school within the public sector and school funding is based on the number of pupils. This is more or less the system in countries like New Zealand and Belgium. Schools will then try to improve their performance in order to attract more pupils and report cards can improve market transparency. However, in this setting it is also easier for schools to build up a better reputation if they attract stronger pupils and this tendency is reinforced by peer effects. There is therefore a danger of segregation. Even if explicit cream-skimming is legally forbidden, it is not difficult for schools to devise strategies that make themselves more attractive for better students and less attractive for students from weaker socio-economic groups or from ethnic minorities. Report cards designed to make the market more transparent, should then certainly spread information about outcome measures with due correction for pupil characteristics. Or, one could consider moving away from simple lump sum funding per pupil and introducing special financing arrangements for weaker groups of pupils (Del Rey, 2004). The problem of formulating such special financing arrangements is formally equivalent to our problem. Moreover, individualized funding taking into pupil characteristics in a system of free school choice is formally equivalent to the design of differentiated voucher schemes (Epple and Romano, 2008), the only difference being that the voucher goes "directly" to the schools.

We introduce our theoretical framework in section 2. This framework is derived from the social choice literature on fair allocations (Fleurbaey, 2008). We will argue that in general, i.e., for any educational production function, there is a deep conflict between creating incentives for efficiency and avoiding incentives for cream-skimming. However, in a special (not necessarily unrealistic case),

4

the two can be reconciled. We characterize a funding scheme satisfying the two requirements.[1] We then illustrate our approach with Belgian data. In section 3, we take the first step of estimating an educational production function with special attention for individual pupil characteristics. In section 4 we show how to derive and interpret the resulting funding scheme. Section 5 concludes.

2 Reward without cream-skimming

We consider a set of at least two pupils, denoted by I. We assume that for the measurement of school performance and accountability, agreement is reached about the use of a single-valued indicator of output, say $y \in \mathbb{R}$. Output y is a function of (1) factors for which the school is not (held) accountable (compensation factors collected in a set C), e.g., innate intelligence and social background of pupils, and (2) factors for which the school is (held) accountable (responsibility factors collected in a set R), e.g., the number of instruction hours in the different disciplines, the organization of the school and the motivation of its teachers. Both factors together completely explain the performance of a pupil and can be summarized by a vector $x = (c, r) \in \mathbb{D} = \mathbb{R}^{|C|+|R|}$. Typically, the compensation factors are pupil level variables, while the responsibility factors are at the school level; therefore, we will, loosely speaking, refer to c as the pupil type and to r as the school policy. We use $f : \mathbb{D} \to \mathbb{R}$ to denote the function mapping pupil type and school policy into output, thus, $y = f(x) = f(c, r)$.

The assumption that y is a scalar is not very restrictive, as y may be seen as a weighted combination of several output indicators. Moreover, at this abstract level, y does not necessarily refer to the score(s) on a cognitive test. It can be a non-linear transformation of such scores, it can refer to the distance to a threshold level or even to the earnings potential of pupils as a result of educational performance (as advocated by Cawley et al., 1999). Following the literature on value added measures, y can also be interpreted as the individual gain in test scores during a given period. We discuss some of these interpretations at the end of section 2.2, but, for convenience, we call y here a simple test score. Initial test scores will enter our model as one of the pupil level variables in C. Note that in this interpretation the function f is the standard explanatory model of test scores as estimated in the educational literature; see, e.g., Hanushek (2006) for an overview. Such

---

[1] Our analysis in this paper is formally similar to the analysis of risk adjustment and cream-skimming in health insurance in Schokkaert et al. (1998) and Schokkaert and Van de Voorde (2004).

estimation will typically involve unobserved (fixed or random) effects at the level of the pupil and the school as well as idiosyncratic error terms. In the next sections we will discuss how to treat these effects in a practical application. For the theoretical analysis, however, we can consider these unobserved effects to be part of the characterization of pupils. Each specific effect then has to be assigned either to C or to R.

We define a school funding scheme $s : \mathbb{D}^{|I|} \to \mathbb{R}^{|I|}$ as a mapping of the profile $\mathbf{x} = (x_i)_{i \in I}$ into a subsidy vector $s(\mathbf{x}) = (s_i(\mathbf{x}))_{i \in I}$. For later use we can decompose profiles $\mathbf{x}$ as $(\mathbf{c}, \mathbf{r})$. Of course, since $y = f(x)$, simple output-related subsidy schemes are one specific example of $s(\mathbf{x})$. More generally, however, we look for a funding scheme that does reward schools that are performing well, but at the same time corrects for differences in pupil characteristics. What form should $s(\mathbf{x})$ then take? To answer this question, we draw inspiration from the axiomatic social choice literature on fair allocations (Fleurbaey, 2008) and we will formulate two formal axioms, capturing the requirements of rewarding performance while avoiding cream-skimming.

First, we deal with reward for better performance. To remove the ambiguity due to differences in pupil characteristics, we focus on performance comparisons between pupils with the same characteristics. Differences in output between such pupils can only be due to differences in school policy, and a good funding scheme should reward the better performing schools. We formalize this requirement as:

REWARD: For all $\mathbf{x}$ in $\mathbb{D}^{|I|}$, there exists a proportionality factor $\alpha > 0$ such that, for all i, j in I, if $c_i = c_j$, then $s_i(\mathbf{x}) - s_j(\mathbf{x}) = \alpha(y_i - y_j)$.

The "reward" axiom imposes that the subsidy difference between the schools should be proportional to the output difference, if the latter cannot be explained by differences in pupil characteristics. The right-hand side of the reward equation consists of two parts, a parameter α and the output difference $(y_i - y_j)$. The parameter α can be interpreted as a conversion factor transforming output (e.g., test score results) into money. Its value will reflect the importance attached to performance incentives. Lowering α will allow to downplay the monetary consequences of test score results; we will return to the choice of α later in this section. Next, reward requires the subsidy difference to be proportional to the output difference, which seems at first sight rather restrictive. However, recall that output y is not necessarily a raw test score, but could also be a non-linear transformation of

such scores; we come back to this possibility later.

Second, since schools do not control the variables in C, output differences that are only due to differences in pupil characteristics should not be rewarded in the funding scheme. Equivalently, two schools that follow the same policy should be treated in the same way. We formalize this as

NO CREAM-SKIMMING 1 (NCS1): For all $\mathbf{x}$ in $\mathbb{D}^{|I|}$, for all i, j in I, if $r_i = r_j$, then $s_i(\mathbf{x}) = s_j(\mathbf{x})$.

The basic idea of the axiom NCS1 is to link the subsidy scheme only to variables that are controlled by the schools. This is a necessary condition to get an unbiased performance indicator and it can also be seen as an equity requirement. For reasons explained in the introduction we use the term "no cream skimming". It is indeed obvious that incentives for cream skimming are removed in a subsidy scheme satisfying NCS1, since there will be no reward for improving test scores by attracting better pupils without changing policies.

The funding scheme s can be interpreted in different ways. In principle, we could think of a system in which the subsidies have to cover all school expenditures. In most educational systems, this is not very realistic however. It is therefore more relevant to interpret $s(\mathbf{x})$ as a bonus scheme that aims at rewarding better performing schools and comes on top of (a) a budget to cover the fixed costs of the school, independent of the number of pupils, and/or (b) a basic financing scheme consisting of a fixed amount per pupil. Of course, the requirements introduced in this section are relevant in both interpretations. We will return to the budget constraint later in this section.

2.1 An impossibility result

If one wants to create incentives for better performance while at the same time correcting for the effect of pupil characteristics, both the reward and the NCS1-axioms seem eminently sensible. It is therefore very striking that it is impossible to design a reward scheme that combines reward and NCS1 for all possible output functions f. This result is well known (under many variants) in the social choice literature (Fleurbaey, 2008), but to the best of our knowledge has until now remained unnoticed in the literature on school accountability. For our purposes, it is sufficient to illustrate the proof with a simple example.

Figure 1 about here

Consider a continuous pupil type c, say socio-economic status, and two different school policies denoted r_1 and r_2. Figure 1 presents output as a function of socio-economic status for both school policies. Take now two specific levels of socio-economic status, denoted c_1 and c_2 and construct the profile $\mathbf{x} = (x_a, x_b, x_c, x_d) = ((c_1, r_1), (c_1, r_2), (c_2, r_1), (c_2, r_2))$. Applying reward tells us that b should get a higher subsidy than a, i.e., $s_b(\mathbf{x}) > s_a(\mathbf{x})$, since both pupils a and b have the same background, but school policy 2 succeeds in bringing pupil b at a higher output level. For the same reasons $s_c(\mathbf{x}) > s_d(\mathbf{x})$. NCS1 imposes that $s_a(\mathbf{x}) = s_c(\mathbf{x})$ and that $s_b(\mathbf{x}) = s_d(\mathbf{x})$, as the same school policies (r_1 and r_2 respectively) apply to both pupils. All things together we get a cycle.

It is not difficult to grasp the intuition behind this impossibility result. School policy r_1 is apparently more effective for pupils with a higher level of socio-economic status, school policy r_2 is more effective for pupils with a lower socio-economic status. In this situation it is obviously impossible to reward better performance without at the same time giving incentives to attract specific types of students. In some sense, with the educational production technology of Figure 1, segregation (and cream-skimming) lead to a better overall performance. If one wants to reward performance, one will have to violate NCS1 and one will have to tolerate segregation. If one wants to avoid segregation, one will have to give up reward. In the social choice literature intermediate schemes have been formulated that satisfy weakened versions of the axioms (Fleurbaey, 2008). However, there is also another way out of the incompatability. This is to restrict the domain of admissible output functions $f(x)$. As an example, note that the incompatibility disappears in Figure 1 if the lines for r_1 and r_2 are parallel to each other. We will follow the latter route in the next subsection.

2.2 Characterization of a subsidy scheme

In this section we will derive a funding scheme that satisfies reward and NCS1. However, to get a full characterization result, we introduce a multi-profile version of the no cream-skimming condition, stating that changes in the overall pupil population, without any change in the school policies, should not affect the distribution of the subsidies.

NO CREAM-SKIMMING 2 (NCS2): For all $\mathbf{x}, \mathbf{x}'$ in $\mathbb{D}^{|I|}$, if $r_i = r'_i$ for all i in I, then $s_i(\mathbf{x}) = s_i(\mathbf{x}')$ for all i in I.

Note that NCS2 does not imply NCS1. However, imposing NCS2 together with reward does imply NCS1. To see this consider an arbitrary profile $\mathbf{x} = (\mathbf{c}, \mathbf{r})$ in $\mathbb{D}^{|I|}$. Construct a new profile $\mathbf{x}'$ in $\mathbb{D}^{|I|}$ with $\mathbf{x}' = (\mathbf{c}', \mathbf{r}') = ((c, c, \ldots, c), \mathbf{r})$. NCS2 requires that $s_i(\mathbf{x}) = s_i(\mathbf{x}')$ for all i in I. Reward implies that for all i, j in I, $s_i(\mathbf{x}') - s_j(\mathbf{x}') = \alpha[f(c, r_i) - f(c, r_j)]$. Combining the two results yields that for all i, j in I, $s_i(\mathbf{x}) - s_j(\mathbf{x}) = \alpha[f(c, r_i) - f(c, r_j)]$, and thus, for all i, j in I with $r_i = r_j$, $s_i(\mathbf{x}) = s_j(\mathbf{x})$. This is condition NCS1.

Given the impossibility result in the previous subsection, we will therefore have to restrict ourselves to specific output functions if we want a funding scheme to satisfy reward and NCS2. How do this subset of 'compatible' output functions and the resulting subsidy scheme look like? The next proposition gives a definite answer: the output function $f(c, r)$ has to be additively separable between pupil type and school policy variables and the subsidy should be an affine transformation of the output part which is explained by school policy.

PROPOSITION 1. Let $f : \mathbb{D} \to \mathbb{R}$ be a function mapping types $x = (c, r)$ into output $y = f(x)$. A subsidy scheme $s : \mathbb{D}^{|I|} \to \mathbb{R}^{|I|}$ satisfies REWARD and NO CREAM-SKIMMING 2 if and only if there exist

1. functions $g : \mathbb{R}^{|C|} \to \mathbb{R}$ and $h : \mathbb{R}^{|R|} \to \mathbb{R}$, with $f(c, r) = g(c) + h(r)$ for all $x = (c, r)$ in $\mathbb{D}$,

2. functions $a : \mathbb{R}^{|R| \times |I|} \to \mathbb{R}$ and $\alpha : \mathbb{R}^{|R| \times |I|} \to \mathbb{R}$,

such that for all $\mathbf{x}$ in $\mathbb{D}^{|I|}$ and for all i in I, we have

$$s_i(\mathbf{x}) = a(\mathbf{r}) + \alpha(\mathbf{r}) h(r_i), \tag{1}$$

with $\alpha(\mathbf{r}) > 0$.

The proof is given in the Appendix. The funding scheme (1) allows the decision-maker to choose freely the parameters $a(\mathbf{r})$ and $\alpha(\mathbf{r})$. We add two simple requirements:

BUDGET BALANCE: There exists an amount $B \geq 0$ such that, for all $\mathbf{x}$ in $\mathbb{D}^{|I|}$, we have $\sum_{i \in I} s_i(\mathbf{x}) = B$.

NON-NEGATIVITY: For all $\mathbf{x}$ in $\mathbb{D}^{|I|}$, for all i in I, $s_i(\mathbf{x}) \geq 0$.

Imposing a budget constraint for the regulator is certainly a reasonable thing to do. Fixing the budget B defines the first unknown $a(\mathbf{r})$. The relevancy of the non-negativity condition depends on the interpretation given to the funding scheme. If we take $s(\mathbf{x})$ to be the only financing source for the schools, it certainly is highly recommendable that subsidies cannot be negative. The situation is less extreme when $s(\mathbf{x})$ only refers to a bonus scheme (on top of other financing sources). However, while in this case it is in principle possible to have negative "subsidies" (fines) as a kind of sanctions, it is more likely that the regulator would prefer to award only positive bonuses. In fact, non-negativity is not very demanding as it only imposes an upper bound on $\alpha(\mathbf{r})$.

Proposition 2 follows directly from proposition 1, with $a(\mathbf{r})$ defined by the budget constraint and $\alpha(\mathbf{r})$ restricted by non-negativity. Let μ be the mean-operator, i.e., $\mu[z] = \frac{1}{|I|}\sum_{i \in I} z_i$ for an arbitrary vector $z = (z_i)_{i \in I}$.

PROPOSITION 2. Let $f : \mathbb{D} \to \mathbb{R}$ be a function mapping types $x = (c, r)$ into output $y = f(x)$. A subsidy scheme $s : \mathbb{D}^{|I|} \to \mathbb{R}^{|I|}$ satisfies REWARD, NO CREAM-SKIMMING 2, BUDGET BALANCE and NON-NEGATIVITY if and only if there exist

1. functions $g : \mathbb{R}^{|C|} \to \mathbb{R}$ and $h : \mathbb{R}^{|R|} \to \mathbb{R}$, with $f(c, r) = g(c) + h(r)$ for all $x = (c, r)$ in $\mathbb{D}$ and

2. a function $\alpha : \mathbb{R}^{|R| \times |I|} \to \mathbb{R}$,

such that, for all $\mathbf{x}$ in $\mathbb{D}^{|I|}$ and for all i in I, we have

$$s_i(\mathbf{x}) = \frac{B}{|I|} + \alpha(\mathbf{r})\left\{h(r_i) - \mu[h(r_i)]\right\}, \tag{2}$$

with $\alpha(\mathbf{r}) > 0$, and, in case $\mu[h(r_i)] \neq \min_{i \in I} h(r_i)$, also $\alpha(\mathbf{r}) \leq \frac{B/|I|}{\mu[h(r_i)] - \min_{i \in I} h(r_i)}$.

The subsidy that a school obtains for one of its pupils equals a per-capita share of the total budget $B/|I|$ plus a correction factor based on the difference between the output part of that pupil for which the school is responsible and the average 'responsible' output part (averaged over all pupils). Given the additively separable specification of $f(c, r)$, equation (2) can be rewritten as

$$s_i(\mathbf{x}) = \frac{B}{|I|} + \alpha(\mathbf{r})\left\{(y_i - \mu[y_i]) - (g(c_i) - \mu[g(c_i)])\right\}, \tag{3}$$

showing that the subsidy for a given pupil is equal to a fixed lump-sum amount $B/|I|$ plus a fraction (depending on α) of (1) her relative performance (the difference between her individual performance y_i and the average overall performance $\mu[y_i]$) plus (2) a correction for her characteristics, which will be positive (negative) if $g(c_i) < (>)\mu[g(c_i)]$. Of course, in practice the funding will be calculated at the level of the school.

To implement the funding scheme (3), we have to know how the educational production function f and its components g and h look like. A necessary condition is that the function f be additively separable between compensation and responsibility variables. If we interpret y as a test score and $f(c, r)$ as an educational production function, this is an empirical question. We can statistically test whether the additively separable specification is an acceptable approximation of the true data generating process. This approach will be followed in the next sections.

Finally, as mentioned before, it is not necessary to interpret y as a raw test score. First, suppose test scores $t = f'(c, r)$ and output $y = m(t)$, a non-linear transformation of these scores. Then the function $f(c, r) = m(f'(c, r))$ has to be additively separable in c and r, and this can be true even if the original educational production function $f'(c, r)$ is not separable. Of course, this approach is not a panacea, since $m(.)$ has normative implications and cannot be adapted freely. Second, value added can be written as $\Delta y_i = y_i - y_{i,0}$, with $y_{i,0}$ the initial test score of individual i. If the initial test score belongs to the compensation factors in C, we can define $\Delta y_i = f(c_i, r_i) - y_{i,0} = g(c_i) - y_{i,0} + h(r_i) = g'(c_i) + h(r_i)$. The subsidy scheme now becomes

$$s_i(\mathbf{x}) = \frac{B}{|I|} + \alpha(\mathbf{r})\left\{(\Delta y_i - \mu[\Delta y_i]) - (g'(c_i) - \mu[g'(c_i)])\right\}.$$

2.3 From first best to second best?

It is worthwhile reflecting about the normative status of the axioms that have been introduced in the previous sections. In some sense they can be seen as partial objectives of the regulator, but we did not introduce explicitly a fully specified objective function; see, e.g., Fleurbaey and Maniquet (2008). Moreover, the axioms are not formulated in terms of the final outcomes to be achieved, but rather as restrictions on the instruments (the subsidies) that can be used. The propositions then show that these restrictions completely fix the resulting funding scheme (and in the general case are even incompatible). On the other hand, the restrictions reflected in the axioms do capture

normative desiderata. This is certainly true for NCS1. It is also true for reward, however, and here the choice of y offers a degree of freedom to the regulator. We mentioned already that y can be a function of the performance in different domains. Moreover, non-linear transformations of the raw test scores offer scope for introducing elitist or egalitarian considerations.

Another striking feature of our approach is that we did not include an explicit behavioral model of the school; see Barlevy and Neal (2009) who analyze incentive pay schemes in a model with teacher behavior. We did not describe how the allocation of subsidies affects the choices made by the school (nor, for that matter, by the pupils). One can therefore interpret the previous analysis as essentially first-best. In our approach, the function $f(c, r)$ should be seen as a reduced form equation and we implicitly assumed that it does not change if additional bonuses are awarded. This restriction is less severe than it may seem. First, there is a growing amount of empirical evidence that increasing financial resources of schools has at best a very limited effect on performance (Hanushek, 2006; Wössmann, 2003). It is true that giving bonuses and/or sanctions could motivate schools to organize themselves in a more efficient way, and this more efficient organization may improve results. However, that effect is captured in our framework by changes in the chosen vector r, without changing the educational production function $f(c, r)$. Second, the scheme that follows from Proposition 2 is incentive compatible. If the schools manage to improve performance (to increase y) while keeping pupil characteristics constant, they will be rewarded. Note, moreover, that while the formulation in equation (2) seems to offer scope for strategic behaviour (e.g. diminishing class size without any real results) or even creates incentives for open misreporting of efforts, this is more difficult in the reformulated equation (3). Test scores are collected in a standardized way and the c-variables refer to pupil characteristics that cannot be changed by the schools and can be controlled easily by the regulator.

2.4 An alternative interpretation: performance measures

We have interpreted our axioms and results in terms of a funding scheme. This is not the only possible interpretation, however. One could as well argue that $s(\mathbf{x})$ represents only a performance measure. Both the axioms reward and NCS1 remain valid in this measurement interpretation. (Reward could be rebaptized as "performance sensitivity", NCS1 as "correction for pupil charac-

teristics".) The impossibility result also remains relevant. However, the additional requirements of budget balance and non-negativity make much less sense, so that we should probably stick to the result in Proposition 1. The fact that $a(\mathbf{r})$ and $\alpha(\mathbf{r})$ can be freely chosen then indicates that our measurement of school performance is at the interval level, and the choice of parameter values boils down to an arbitrary standardization.

3 Empirical illustration, step one: explaining test scores

We now turn to one specific illustration of the general framework. For this illustration we assume that y is one-dimensional and only includes scores on a mathematics test. Our data are from Flanders, which is the biggest region in Belgium, with a separate educational policy. In this section we focus on the estimation of the function $f(c, r)$. We first describe the data and then turn to the estimation results. We will test for the additive separability of $f(c, r)$. The consequences for the school funding scheme are discussed in the next section.

3.1 The data

The data comes from the SiBO-project, whose aim is to describe and explain differences in the primary school curriculum of Flemish pupils. We look at test score results in mathematics, socio-economic background variables, and classroom data for a cohort of pupils during the first two grades (at the normal age of 6 and 7), corresponding with school years 2003-2004 and 2004-2005 respectively.

At the beginning of the first grade (September-October 2003), and at the end of the first and the second grade (around May-June of 2004 and 2005, respectively), pupils were tested in mathematics. The math tests consist of between 40 and 80 questions (depending on the grade), grouped into different topics. Two small remarks: (1) we do not have test scores at the beginning of grade 2, and (2) the tests contain different questions (different difficulty levels), which explains why test score results decline over time. Figure 2 presents a kernel density estimate of the math scores, rescaled into percentages.

Figure 2 about here

The distributions are reasonably well-behaved, showing no floor and only limited ceiling effects. Besides test scores, we also have socio-economic background variables and class-related data, summarized in Tables 1 and 2 respectively. The former include gender and age of the pupil, and education level and mother tongue of both parents. From the pupil's age, we constructed a dummy variable indicating whether the pupil is behind or ahead of age, and we distinguished the cases where this is due to a decision by the school itself or rather was already a fact at the moment the pupil entered the school. The class-related data include the total experience of the teacher, the class size, the common instruction time for math (in hours per week) and whether two teachers are teaching the class together or not. We also construct a peer-effect variable as the average initial test score (begin grade 1) of all pupils in a certain grade at school.

Tables 1 and 2 about here

Observations on pupils can be missing for two reasons: missing test scores at the end of a grade and/or missing covariates. Table 3 summarizes these reasons per grade, together with the number of pupils involved (n) and the average initial math test score result for the subgroup ($\bar{y}_0$).[2] Note that there is a difference in average initial test scores between the tested and non-tested pupils, while this difference is less clear between pupils with and without missing covariates.[3] We have 6373 pupil-time observations in total, with 2315 pupils appearing in both grades (4630 pupil-time observations), 658 only in grade 1, and 1085 only in grade 2. These pupils are distributed over 121 different schools.[4]

Table 3 about here

---

[2] We use a Heckman selection model to impute initial test scores at the beginning of the first grade for 484 pupils. The estimation is based on the pupils' background characteristics, and—for the selection equation only—we exploit the fact that some schools agreed to participate in the SiBO-project, but not every year, which we consider to be an appropriate instrument.

[3] The value for the non-tested pupils in grade 2 is unreliable due to the very low number of observations with incomplete covariates.

[4] To limit the reduction in total sample size, and, given that missing covariates is not systematically linked with lower (initial) test score results in Table 3, we add an additional classification 'missing' to the covariate dummies; we do not report the corresponding estimates which are, as expected, all insignificant.

3.2 Empirical model and results

Let y_{ijt} be a single-valued test score of pupil i at school j at time t and z_{ijt} a vector of observable regressors. We use a basic linear panel model

$$y_{ijt} = \beta' z_{ijt} + u_i + v_{jt} + w_{ijt}, \tag{4}$$

with β a vector of marginal effects and with the overall error term decomposed into a time-invariant pupil-level effect u_i, a school-grade level effect v_{jt} and an idiosyncratic error term w_{ijt}. Since the mobility of pupils over schools is very limited in the sample, we assume conditional mean independence, i.e., $E\left[u_i|v_{jt}\right] = 0$, to separate the unobserved pupil-level effect u_i from the unobserved school-grade level v_{jt}.

The linear panel specification satisfies the additive separability condition (as defined in propositions 1 and 2), independent of how the right-hand variables will be classified into compensation and responsibility factors. The question remains however whether it is a reasonable specification for the data. To test linearity, we performed a Box-Cox regression on the pooled data; see, e.g., Cameron and Trivedi (2005). To be more precise, three models were tested: a Box-Cox transformation of the dependent variable only, say $y^\theta = \beta' z$, the same Box-Cox transformation for the dependent and the (continuous) covariates ($y^\theta = \beta' z^\lambda$, with $\theta = \lambda$), and a flexible specification ($y^\theta = \beta' z^\lambda$).[5] Table 4 presents, for each model, the 95% confidence interval for θ (and possibly, λ), as well as 'likelihood ratio'-test results (the χ^2-value and the corresponding p-value) for the linear, loglinear or inverse hypothesis.

Table 4 about here

Note that the estimate for θ is close to 1 in all three cases, though statistically rejected in the reported likelihood ratio-tests. Still, it is clear that the linear specification fares much better compared to the log-linear or inverse specification, suggesting that a linear specification is a reasonable approximation. To test for separability, we do have to classify the right-hand variables as either

---

[5] Two additional remarks. First, since the dependent variable and some of the continuous covariates (like initial test score) are equal to zero for some observations, we added a very small amount to it; we obtain similar results by dropping these observations. Second, since separability between (some of) the covariates has not been tested yet, we reestimated these three Box-Cox regression models, while including all interaction effects between the covariates; this inclusion of interactions did not change the conclusion of a 'mild' rejection of additivity.

compensation or responsibility factors. For the intermediate benchmark case—loosely speaking, pupil level variables/effects are compensation factors, while school level effects are responsibility factors (see later, for a more precise description)—, we test separability by testing the hypothesis that all interaction effects between compensation and responsibility factors are equal to zero. The null hypothesis is statistically rejected ($F(104, 656) = 11.77$ with $p = 0.000$). However, since we are interested in predicting output (see later), it is important to note that the mean absolute deviation between the predicted output in both cases (resp. without and with interactions) is small (2.1%) compared to the average standard errors of the predictions (resp. 1.3% and 2.4%).

Tables 5 and 6 about here

Table 5 provides a description of the variables used in the estimations, while Table 6 summarizes the results. As was to be expected, the initial test score plays an important role in all models. Its coefficient is rather robust and smaller than 1, indicating that the added value, i.e., the gain in test scores, is larger for pupils with a lower initial test score.[6] The background variables play a more modest role and their effects depend on whether or not the initial test score is taken up as a covariate. In model (c) without initial test score, boys do better than girls, being ahead of age is not significant while lagging behind is correlated with a lower math performance, having Dutch-speaking and better educated parents improve test scores and these effects are stronger and more significant for mothers compared to fathers. In model (d) with initial test scores as an additional regressor, the estimated coefficients for the background variables change in magnitude and even in sign. We provide two striking examples. First, once we correct for initial test scores, having Dutch-speaking parents gets a negative coefficient. Indeed, pupils with non-Dutch speaking parents have (on average) a worse preparation before starting primary education. Therefore their initial test score underestimates their potential, leading to a catching-up effect in the first grades. Second, the effect of father education is now stronger than that of mother education. One hypothesis could be that mothers have a larger effect on initial test scores (during the pre-primary education period), while fathers have a larger effect on the primary education growth of their children. Comparing model (d) and (e), adding class data does not change the coefficient estimates for the individual-

---

[6]Subtracting the initial test score y_0 from both sides of a regression equation $y_t = \beta y_0 + \cdots + \epsilon$ leads to value added $y_t - y_0$ on the left-hand side and $(\beta - 1) y_0$ on the right-hand side.

specific variables very much. Among the class variables, only instruction time and the peer effect play a significant and positive role.

Before proceeding, recall Table 3. Because of the high number of missing observations due to missing test score results in one of the periods, we did not check and/or correct for selection bias. To check for selection bias we use a variable addition test; see Verbeek and Nijman (1992) and Wooldridge (1995). More precisely, we add two dummies to the covariates indicating whether the pupil is tested at the end of period 1 (respectively period 2) or not. The results indicate that missingness might be informative, but only for the pupils who drop out after grade 1. To check whether selection correction influences our estimation results, we added a selection equation to each period in the spirit of Hausman and Wise (1979), allowing for correlation between the individual level effects in the selection and output equation. However, the corrected estimates do not statistically differ from the uncorrected estimates.[7] Finally, note that Ladd and Walsh (2002) have shown that a failure to correct for measurement error in the initial test scores may lead to a bias against low-performing students. We do not dispose of good instruments to tackle the problem and we have therefore neglected it. In fact, our empirical application is only meant to be an illustration of how the theoretical concepts introduced in the previous section can be implemented.

4   Empirical illustration, step 2: financing schools

The next step is to use the results from estimating equation (4) to calculate the school subsidies following expressions (2) or (3). We have shown in the previous section that the assumption of additive separability is an acceptable approximation in our data. Two further issues remain. First, we must classify the right-hand side variables $(z_{ijt}, u_i, v_{jt}, w_{ijt})$ as either compensation or responsibility factors. Second, (some of) the unobserved components have to be predicted. We discuss these methodological issues in greater detail in the next subsection. We then discuss how the funding scheme would look like for the Flemish schools in our sample.

---

[7]The reported selection correction model assumed random (rather than fixed) school effects, resulting in a so-called multi-level model. An attempt with fixed school effects did not converge, probably due to the high number of school-time dummies in the selection equation.

4.1 Implementation

The partitioning of x in c and r is in the first place a normative exercise. The regulator has to decide about for which factors he wants to hold the schools responsible and for which factors he is willing to compensate. The procedure we have proposed will work for all possible partitionings of x. For illustrative reasons, we will focus on what we consider to be the most relevant benchmark case, in which we split up all *observable* factors z_{ijt} into

1. compensation factors (denoted $z_{c,ijt}$): these include the peer effect, the time effect and all pupil-related variables, except the "ahead and behind age"-dummies when the decision about retainment or skipping a class is taken by the school itself.

2. responsibility factors (denoted $z_{r,ijt}$): these include the school-grade level variables except for the peer effect, as well as the variables ahead and behind age, when the decision is taken by the school.

Note that it does not matter where the constant term is assigned to. With respect to the *unobservables*, we assume schools to be responsible for the unobserved school-grade effect v_{jt}, but not for the pupil effect u_i and the idiosyncratic error term w_{ijt}.

To summarize, we interpret equation (4) as

$$y_{ijt} = \underbrace{\left(\widehat{\beta}'_c z_{c,ijt} + u_i + w_{ijt}\right)}_{g(c_{ijt})} + \underbrace{\left(\widehat{\beta}'_r z_{r,ijt} + v_{jt}\right)}_{h(r_{ijt})} \tag{5}$$

Plugging this expression in the subsidy scheme (2) and using $\mu[v_{jt}] = 0$, we get[8]

$$s_{ijt}(\mathbf{x}) = \frac{B}{|I|} + \alpha \left\{ \widehat{\beta}'_r (z_{r,ijt} - \mu[z_{r,ijt}]) + v_{jt} \right\}. \tag{6}$$

The subsidy formula depends on the observed responsibility factors $z_{r,ijt}$ and on the unobserved school-grade effect v_{jt}. The latter must still be predicted. There are basically two ways to proceed from here (see, e.g., Longford, 1994, for a detailed discussion). One possibility is to use the posterior mean to predict v_{jt}. This estimator is stable in small samples, but it is biased. We opted for using the unbiased OLS estimate, although this may be less stable. Note however that, with our data, the differences between both methods are extremely small: the correlation between the resulting

---

[8] Since we focus on a fixed profile, we replace $\alpha(\mathbf{r})$ by α.

subsidy schemes is 0.999. In addition, the OLS-estimate has a decisive theoretical advantage, in that it allows us to express the total subsidy for a school—the sum of the subsidies for its pupils— as a function of school output and observable compensation factors, i.e., to implement scheme (3). We argued in section 2 why this implementation is less vulnerable for manipulation and strategic gaming.

The OLS estimator equals

$$
\begin{aligned}
\widehat{v}_{jt} &= \mu_{jt}\left[y_{ijt}\right] - \widehat{\beta}'\mu_{jt}\left[z_{ijt}\right] \\
&= \mu_{jt}\left[y_{ijt}\right] - \widehat{\beta}'_c\mu_{jt}\left[z_{c,ijt}\right] - \widehat{\beta}'_r\mu_{jt}\left[z_{r,ijt}\right],
\end{aligned}
$$

with (for an arbitrary vector a) $\mu_{jt}\left[a_{ijt}\right] = \sum_i a_{ijt}d_{ijt}/\sum_i d_{ijt}$ and d_{ijt} is a dummy-indicator, indicating whether pupil i has a test score at school j at time t ($d_{ijt} = 1$) or not ($d_{ijt} = 0$). Plugging this estimate into equation (6), and using the expression (capturing the assumption of additive separability) $\widehat{\beta}'_r\mu\left[z_{r,ijt}\right] = \mu\left[y_{ijt}\right] - \widehat{\beta}'_c\mu\left[z_{c,ijt}\right]$, we get

$$
s_{ijt}\left(\mathbf{x}\right) = \frac{B}{|I|} + \alpha\left\{\widehat{\beta}'_r\left(z_{r,ijt} - \mu_{jt}\left[z_{r,ijt}\right]\right) + \left(\mu_{jt}\left[y_{ijt}\right] - \mu\left[y_{ijt}\right]\right) - \widehat{\beta}'_c\left(\mu_{jt}\left[z_{c,ijt}\right] - \mu\left[z_{c,ijt}\right]\right)\right\}.
$$

This expression is still at the individual level. In practice funding will be at the level of the school. Calculating the average school subsidy for school j, the first term between curly brackets averages out, and we are left with

$$
s_j\left(\mathbf{x}\right) = \mu_j\left[s_{ijt}\left(\mathbf{x}\right)\right] = \frac{B}{|I|} + \alpha\left\{\left(\mu_j\left[y_{ijt}\right] - \mu\left[y_{ijt}\right]\right) - \widehat{\beta}'_c\left(\mu_j\left[z_{c,ijt}\right] - \mu\left[z_{c,ijt}\right]\right)\right\}, \tag{7}
$$

with $\mu_j\left[a_{ijt}\right] = \sum_{it} a_{ijt}d_{ijt}/\sum_{it} d_{ijt}$. This empirical counterpart of equation (3) is the basic expression that we use to calculate school subsidies in the next subsection. The average subsidy a school receives is equal to the per-capita share $B/|I|$ plus a fraction (depending on α) of (1) the relative school performance (the difference between the average school performance $\mu_j\left[y_{ijt}\right]$ and average overall performance $\mu\left[y_{ijt}\right]$) minus (2) the relative school profile (the difference between the predicted average school performance on the basis of the compensation factors $\widehat{\beta}'_c\mu_j\left[z_{c,ijt}\right]$ and the predicted average overall performance $\widehat{\beta}'_c\mu\left[z_{c,ijt}\right]$). Eq. (7) does not include any responsibility variables. Moreover, the unobservable pupil-level variables u_i and w_{ijt} are averaged out at the school level, as a result of our identifying mean independence assumption. To calculate the subsidies, we only need information about average test scores, average observable pupil characteristics at the school level, and the estimates for β_c.

This benchmark case is only one (albeit in our view the most attractive) possibility to implement our theoretical framework. Other normative choices are possible, e.g. one could have doubts about our classification of the variables "ahead and behind age as chosen by the school" as a responsibility variable (is this a real choice?), or of the peer group-effect as a compensation variable (can it not be controlled to some extent by the schools when they decide about the composition of their classes?). To illustrate the implications of different normative choices, we compare our benchmark case with two extreme cases. One is the "traditional" system without school accountability. In our theoretical framework, this means that all variables are in C. As equation (2) shows this leads to identical subsidies for all pupils and hence to a school funding system which only takes into account the number of pupils. This funding scheme does not give any financial incentives to improve performance in terms of test scores. A second extreme case is the simple accountability-approach in which there is no correction for pupil characteristics at all. In our framework, this means that all determinants are in R. Equation (3) then shows that school funding will be based on uncorrected output scores. As argued before, these are a biased indicator of school performance and using them for calculating the subsidies creates incentives for cream-skimming.

4.2   Results

Let us now look at the results for a selection of 58 schools (out of the 121 in our sample) where a sufficient number of pupils have been tested, i.e., more than 30 pupils and more than 80% of the relevant pupil population at the school. As an innocuous normalization, we take $B = |I|$, which means that schools would receive 1 unit per pupil if they were not held accountable at all.

Figure 3 plots the relative school performance $(\mu_j [y_{ijt}] - \mu [y_{ijt}])$ versus the relative school profile $(\widehat{\beta}'_c (\mu_j [z_{c,ijt}] - \mu [z_{c,ijt}]))$ Each dot in the figure corresponds to the position of one school. Using a simple performance measure, i.e., using only information about test scores, the higher a school is, the better it is considered to be. However, if we correct for pupil characteristics, a school will be considered better the further it is above and away from the diagonal line (containing the points with $\mu_j [y_{ijt}] - \mu [y_{ijt}] = \widehat{\beta}'_c (\mu_j [z_{c,ijt}] - \mu [z_{c,ijt}])$). The differences are striking. There are 14 false negatives (FN), i.e., schools which would be considered to be poor performers in a primitive accountability scheme, but become good performers once we introduce a correction for pupil char-

acteristics. Analogously, there are 21 false positives (FP). On the other hand, information about performance is important. Suppose we evaluated schools only in terms of their pupil characteristics. Then all schools to the left of the vertical axis would be considered to be relatively "disadvantaged", while all schools to the right of the vertical axis have a relatively "advantaged" pupil population. Note that, even after correction, some of the former are poor performers, while some of the latter do even better than what could have been expected on the basis of their population. Still, there are more schools in the zones FN and FP than in the zones denoted A and B. This means that schools with a disadvantaged population are performing relatively well, while schools with an advantaged population are doing relatively poorly. This is not difficult to explain in a setting where there is regulation through the imposition of minimal output norms. To reach the minimum performance level, "disadvantaged" schools must strive harder. In that sense, imposing minimal quality norms without sufficiently correcting for pupil characteristics, generates inequity between schools.

Figure 3 about here

Of course, a similar picture emerges when we calculate the school subsidies (per pupil). Results are shown in Figure 4 for $\alpha = 1/30$.[9] "Corrected" subsidies refer to the funding scheme (7) of our benchmark case. The extreme accountability case with funding based on output scores only, boils down to an application of equation (7), but leaving out the correction term $-\widehat{\beta}'_c \left(\mu_j \left[z_{c,ijt} \right] - \mu \left[z_{c,ijt} \right] \right)$. This yields the "uncorrected" subsidies in Figure 4. Finally, funding per pupil without school accountability, would put all the schools in the center (1,1) of the Figure. The false negative (FN) schools are in the bottom-right, receiving less than their per-capita share if subsidies are uncorrected, but more than the per-capita share if corrected; the opposite happens for the false positives in the top left corner.

Figure 4 about here

Additional insights are obtained by looking at schools that are top receivers ($>80\%$-tile), middle receivers (in between 40%-tile and 60%-tile) and low receivers ($<20\%$tile) for both the uncorrected and corrected subsidy scheme. Recall equation (7) and the decomposition of the proportionality part

---

[9]The value $\alpha = 1/30$ guarantees non-negativity of the subsidies. Further decreasing α would only bring the subsidies closer together on both axes.

into relative performance minus relative profile. Table 7 presents the average relative performance (perf.) and the average relative profile (prof.) for these groups in each of the subsidy schemes. The uncorrected (simple accountability) approach allocates large subsidies to the good performers, with a very strong bias in favour of schools with an advantaged profile of pupil characteristics. Once we turn to the corrected (benchmark) case, however, mean performance is about the same for bottom, middle and top receivers. Top receivers are now mainly schools with a disadvantaged population, low receivers are schools with an advantaged pupil population. Note that this does not mean that relative performance does not play a role in the funding scheme, as equation (7) testifies. It reflects that differences in performance are more strongly linked to pupil characteristics than to observable school variables, and that, in the present Flemish system of quality regulation, schools with more disadvantaged pupils (have to) perform better.

Table 7 about here

5 Conclusion

Recent experiences have shown that introducing school accountability may create incentives for efficiency. However, it is necessary to correct for individual pupil characteristics. Otherwise the performance measures are biased, creating perverse incentives for cream-skimming and leading to an inequitable treatment of schools with a disadvantaged pupil population. To calculate these corrections, one needs information on educational production functions. We have shown how this empirical information from the educational literature can be integrated in a coherent normative framework inspired by the growing non-welfarist literature on fair allocations.

We borrowed from this literature the insight that the requirements of rewarding performance and avoiding incentives for cream-skimming are incompatible on the general domain of educational production functions. However, restricting ourselves to educational production functions that are separable in pupil characteristics and in school policy variables, we characterized an attractive funding scheme that satisfies both requirements. This funding scheme uses only easily controllable information on average test scores and average pupil characteristics at the school level. Once one has an estimate of the educational production function, it can be easily implemented, and the separability assumption can be tested on the data.

22

Our application to Flemish schools shows that correcting for individual pupil characteristics leads to a substantial change in the performance measures, and hence in the subsidies allocated to the different schools. Moreover it revealed the interesting insight that a system with quality regulation without compensations for pupil characteristics forces only schools with a disadvantaged pupil population to perform relatively better in order to satisfy the quality norms. Even if the regulator is not willing to introduce financial accountability in the system (which may be the case in many European countries), it should then consider compensations for schools with a socially disadvantaged population. The only coherent way of calculating these compensations is to base them precisely on the additional effort needed to reach the quality norms, i.e., to improve the scores for pupils that are more difficult to educate. The information needed to calculate these compensations is then very similar to the information needed to implement our full funding scheme.

As we wanted to focus on the problem of correcting for pupil characteristics, we neglected a host of strategic and practical issues that have been documented in the empirical literature (such as teaching to the rating, selective retainment, measurement error, instability of the funding for smaller schools). Taking these into account in a satisfactory way would necessitate introducing a model of school (and pupil) behaviour into our theoretical setting. We then have to go beyond the enumeration of requirements that a good funding scheme has to satisfy and we have to specify a complete fair social ordering (Fleurbaey and Maniquet, 2008). A first step in the direction of a more complete specification of social objectives would be to extend our approach to multidimensional outcome measures (including also non-cognitive abilities) and to look at the potential implications of working with non-linear transformations of test scores to capture different egalitarian or elitist intuitions.

References

Barlevy, G., and Neal, D., 2009, *Pay for percentile.* IZA: Discussion Paper 4383.

Burgess, S., C. Propper, H. Slater, and D. Wilson. 2005. *Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools.* CMPO, Bristol: Working Paper 05/128.

Burgess, S., C. Propper, and D. Wilson. 2007. The impact of school choice in England. *Policy Studies* 28, no. 2: 129-43.

Cameron, C., and P. Trivedi. 2005. *Microeconometrics.* Cambridge University Press.

Cawley, J., J. Heckman, and E. Vytlacil. 1999. On policies to reward the value added by educators. *Review of Economics and Statistics* 81, no. 4: 720-727.

Chiang, H. 2009. How accountability pressure on failing schools affects student achievement. *Journal of Public Economics* 93: 1045-57.

Del Rey, E. 2004. Funding schools for greater equity. *Regional Science and Urban Economics* 34: 202-24.

Epple, D., and R. Romano. 2008. Educational vouchers and cream skimming. *International Economic Review* 49, no. 4: 1395-435.

Figlio, D., and C. Rouse. 2006. Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics* 90: 239-55.

Figlio, D., and J. Winicki. 2005. Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics* 89: 381-94.

Fleurbaey, M. 2008. *Fairness, responsibility and welfare.* Oxford: Oxford University Press.

Fleurbaey, M. and F. Maniquet. 2008. Fair social orderings. *Economic Theory* 34: 25-45.

Hanushek, E. 2006. School resources. In: E. Hanushek, and F. Welch (eds.), *Handbook of the Economics of Education, Vol. 2.*, chapter 14. Amsterdam: Elsevier.

Hanushek, E. and M. Raymond. 2003. Lessons about the design of state accountability systems. In: P. Peterson, and M. West (eds.), *No child left behind? The politics and practice of accountability*: 127-51. Washington D.C.: Brookings.

Hanushek, E. and M. Raymond. 2004. The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association* 2, no. 2-3: 406-15.

Hanushek, E. , and M. Raymond. 2005. Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management* 24, no. 2: 297-327.

Hausman, J, and Wise, D. 1979. Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica* 47(2), 455-473.

Jacob, B. 2005. Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics* 89: 761-96.

Kane, T., and D. Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16, no. 4: 91-114.

Little, R., and Rubin, D. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Ladd, H., and R. Walsh. 2002. Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review* 21: 1-17.

Longford, N. 1994. *Random Coefficient Models*. Oxford: Oxford University Press.

Meyer, R. 1997. Value-added indicators of school performance: a primer. *Economics of Education Review* 16, no. 3: 283-301.

Neal, D. 2008. Designing incentive systems for schools . In: M. Springer (ed.), *Performance incentives: their growing impact on American K-12 education*: forthcoming. Washington D.C.: Brookings.

Reback, R. 2008. Teaching to the rating: school accountability and the distribution of student achievement. *Journal of Public Economics* 92: 1394-415.

Schokkaert, E., G. Dhaene, and C. Van de Voorde. 1998. Risk adjustment and the trade-off between efficiency and risk selection: an application of the theory of fair compensation. *Health Economics* 7: 465-80.

Schokkaert, E., and C. Van de Voorde. 2004. Risk selection and the specification of the conventional risk adjustment formula. *Journal of Health Economics* 23: 1237-59.

Taylor, J., and Anh Ngoc Nguyen. 2006. An analysis of the value added by secondary schools in England: is the value added indicator of any value? *Oxford Bulletin of Economics and Statistics* 68, no. 2: 203-24.

Verbeek, M., and Nijman, T. 1992. Testing for selectivity bias in panel data models. *International Economic Review* 33, 681-703.

West, M., and P. Peterson. 2006. The efficacy of choice threats within school accountability systems: results from legislatively induced experiments. *Economic Journal* 116: C46-C62.

Wooldridge, J. 1995. Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics* 68, 115-132.

Wössmann, L. 2003. Schooling resources, educational institutions and student performance: the international evidence. *Oxford Bulletin of Economics and Statistics* 65, no. 2: 117-70.

Proof of proposition 1

It is easy to verify that, if the output function f is additively separable —i.e., there exist functions $g : \mathbb{R}^{|C|} \to \mathbb{R}$ and $h : \mathbb{R}^{|R|} \to \mathbb{R}$ such that $f(c, r) = g(c) + h(r)$ for all $x = (c, r)$ in $\mathbb{D}$—, and if the subsidy scheme can be written (for all $\mathbf{x}$ in $\mathbb{D}^{|I|}$ and for all i in I) as $s_i(\mathbf{x}) = a(\mathbf{r}) + \alpha(\mathbf{r}) h(r_i)$, for some constants $a(\mathbf{r})$ and $\alpha(\mathbf{r}) > 0$, then it satisfies reward and no cream-skimming. We prove the opposite.

Step 1. *If a subsidy scheme satisfies reward and no cream-skimming then the output function has to be additively separable between compensation and responsibility factors.*

Note first that no cream-skimming requires subsidies to be a function of the responsibility profile $\mathbf{r}$ only, thus there exist a list of functions $(\phi_i)_{i \in I}$, such that, for all $\mathbf{x}$ in $\mathbb{D}^{|I|}$ and for all i in I, $s_i(\mathbf{x}) = \phi_i(\mathbf{r})$. Now, consider arbitrary c, c' in $\mathbb{R}^{|C|}$ and r, r' in $\mathbb{R}^{|R|}$ and construct profiles $\mathbf{x} = (x_a, x_b), \mathbf{x}' = (x'_a, x'_b)$ in $\mathbb{D}^2$ such that $(x_a, x_b) = ((c, r), (c, r'))$ and $(x'_a, x'_b) = ((c', r), (c', r'))$. Note that $\mathbf{r} = \mathbf{r}'$. Reward allows the proportionality factor to be profile-dependent, so, if we apply reward to profile $\mathbf{x}$, we get

$$s_a(\mathbf{x}) - s_b(\mathbf{x}) = \widetilde{\alpha}(\mathbf{x})(f(c, r) - f(c, r')),$$

with $\widetilde{\alpha}(\mathbf{x}) > 0$. Using no cream-skimming (and functions ϕ_a and ϕ_b), we must have

$$\phi_a(\mathbf{r}) - \phi_b(\mathbf{r}) = \alpha(\mathbf{r})(f(c, r) - f(c, r')),$$

with $\alpha(\mathbf{r}) > 0$. Similarly, if we apply reward to profile $\mathbf{x}'$ we get

$$\phi_a(\mathbf{r}') - \phi_b(\mathbf{r}') = \alpha(\mathbf{r}')(f(c', r) - f(c', r')),$$

or, given that $\mathbf{r} = \mathbf{r}'$,

$$\phi_a(\mathbf{r}) - \phi_b(\mathbf{r}) = \alpha(\mathbf{r})(f(c', r) - f(c', r')).$$

Combining both results, we must have

$$f(c, r) - f(c, r') = f(c', r) - f(c', r'),$$

which, recall, has to be true for arbitrary c, c' in $\mathbb{R}^{|C|}$ and r, r' in $\mathbb{R}^{|R|}$. Fixing c' in $\mathbb{R}^{|C|}$ and r' in

$\mathbb{R}^{|R|}$, and defining

$$g \quad : \quad \mathbb{R}^{|C|} \to \mathbb{R} : c \mapsto g\left(c\right) := f\left(c, r'\right)$$

$$h \quad : \quad \mathbb{R}^{|R|} \to \mathbb{R} : r \mapsto h\left(r\right) := f\left(c', r\right) - f\left(c', r'\right)$$

we obtain

$$
\begin{aligned}
f\left(c, r\right) \quad &= \quad f\left(c, r'\right) + f\left(c', r\right) - f\left(c', r'\right) \\
&= \quad g\left(c\right) + h\left(r\right),
\end{aligned}
$$

for arbitrary (c, r) in $\mathbb{D}$, as required.

Step 2. *If a subsidy scheme satisfies reward and no cream-skimming, then the subsidy scheme can be written (for all $\mathbf{x}$ in $\mathbb{D}^{|I|}$ and for all i in I) as $s_i\left(\mathbf{x}\right) = a\left(\mathbf{r}\right) + \alpha\left(\mathbf{r}\right) h\left(r_i\right)$, for some constants $a\left(\mathbf{r}\right)$ and $\alpha\left(\mathbf{r}\right) > 0$ and with h defined in step 1.*

Consider an arbitrary profile $\mathbf{x}$ in $\mathbb{D}^{|I|}$. Construct a new profile $\mathbf{x}'$ in $\mathbb{D}^{|I|}$ with $\mathbf{x}' = \left(\mathbf{c}', \mathbf{r}'\right) = \left(\left(c, c, \ldots, c\right), \mathbf{r}\right)$ for some arbitrary c in $\mathbb{R}^{|C|}$. Recall that no cream-skimming requires subsidies to be a function of the responsibility profile $\mathbf{r}$ only, thus there exist a list of functions $\left(\phi_i\right)_{i \in I}$, such that, for all $\mathbf{x}$ in $\mathbb{D}^{|I|}$ and for all i in I, $s_i\left(\mathbf{x}\right) = \phi_i\left(\mathbf{r}\right)$; note that $s_i\left(\mathbf{x}\right) = \phi_i\left(\mathbf{r}\right) = \phi_i\left(\mathbf{r}'\right) = s_i\left(\mathbf{x}'\right)$ for all i in I. Since all pupils in $\mathbf{x}'$ have the same compensation vector c, reward applied to profile $\mathbf{x}'$ tells us that, for all i, j in I we have

$$s_i\left(\mathbf{x}'\right) - s_j\left(\mathbf{x}'\right) = \widetilde{\alpha}\left(\mathbf{x}'\right)\left(f\left(c, r_i'\right) - f\left(c, r_j'\right)\right),$$

with $\widetilde{\alpha}\left(\mathbf{x}'\right) > 0$. Using no cream-skimming, additive separability of f (from step 1) and the fact that $\mathbf{r} = \mathbf{r}'$ we get for all i, j in I

$$\phi_i\left(\mathbf{r}\right) - \phi_j\left(\mathbf{r}\right) = \alpha\left(\mathbf{r}\right)\left(h\left(r_i\right) - h\left(r_j\right)\right),$$

with $\alpha\left(\mathbf{r}\right) > 0$. Finally, defining $k = \arg\min_{i \in I} h\left(r_i\right)$ we get for all i in I that

$$
\begin{aligned}
\phi_i\left(\mathbf{r}\right) \quad &= \quad \phi_k\left(\mathbf{r}\right) + \alpha\left(\mathbf{r}\right)\left(h\left(r_i\right) - h\left(r_k\right)\right) \\
&= \quad \underbrace{\phi_k\left(\mathbf{r}\right) - \alpha\left(\mathbf{r}\right)\min_{i \in I} h\left(r_i\right)}_{a(\mathbf{r})} + \alpha\left(\mathbf{r}\right) h\left(r_i\right).
\end{aligned}
$$

Fixing the minimal subsidy $s_k\left(\mathbf{x}\right) = \phi_k\left(\mathbf{r}\right)$, and defining $a\left(\mathbf{r}\right) = \phi_k\left(\mathbf{r}\right) - \alpha\left(\mathbf{r}\right)\min_{i \in I} h\left(r_i\right)$, we get the desired representation $s_i\left(\mathbf{x}\right) = a\left(\mathbf{r}\right) + \alpha\left(\mathbf{r}\right) h\left(r_i\right)$, with $\alpha\left(\mathbf{r}\right) > 0$, as required.

Figures and tables

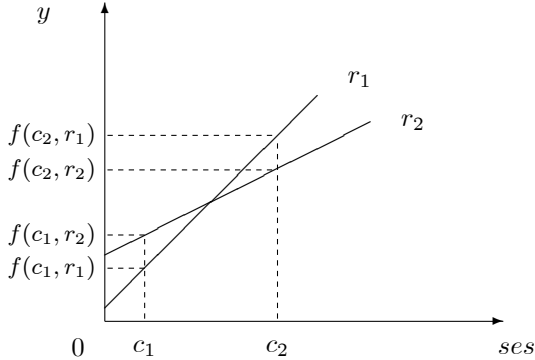**Figure 1:** Impossibility to satisfy reward and no cream-skimming.

Figure 2: Kernel density estimates of math scores at different moments in time.
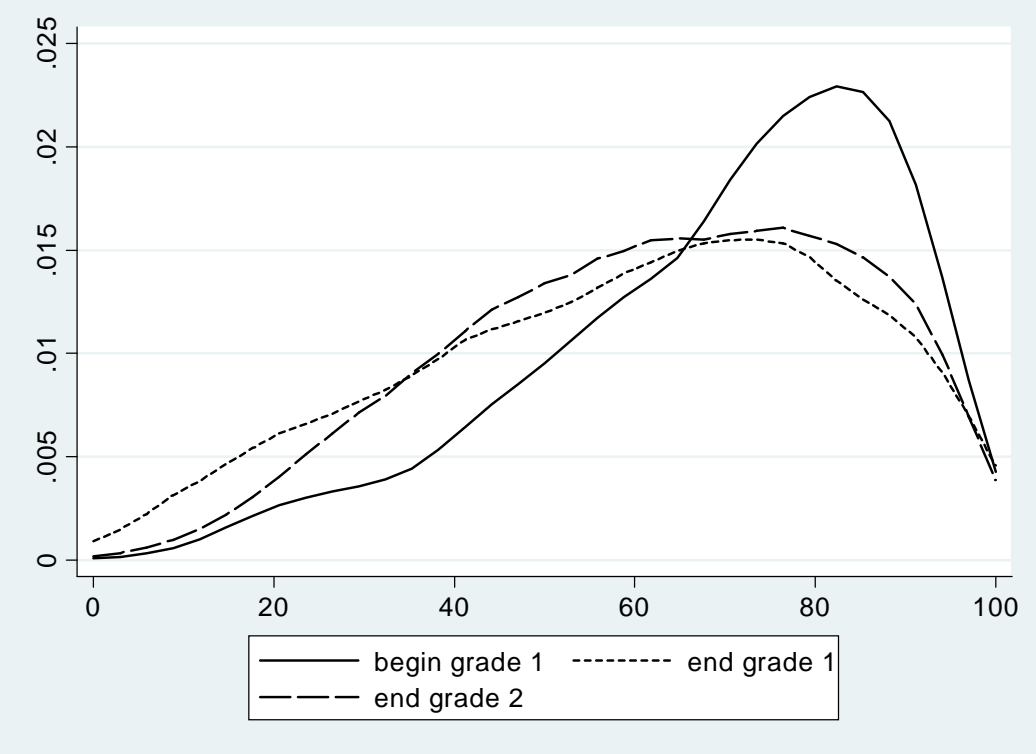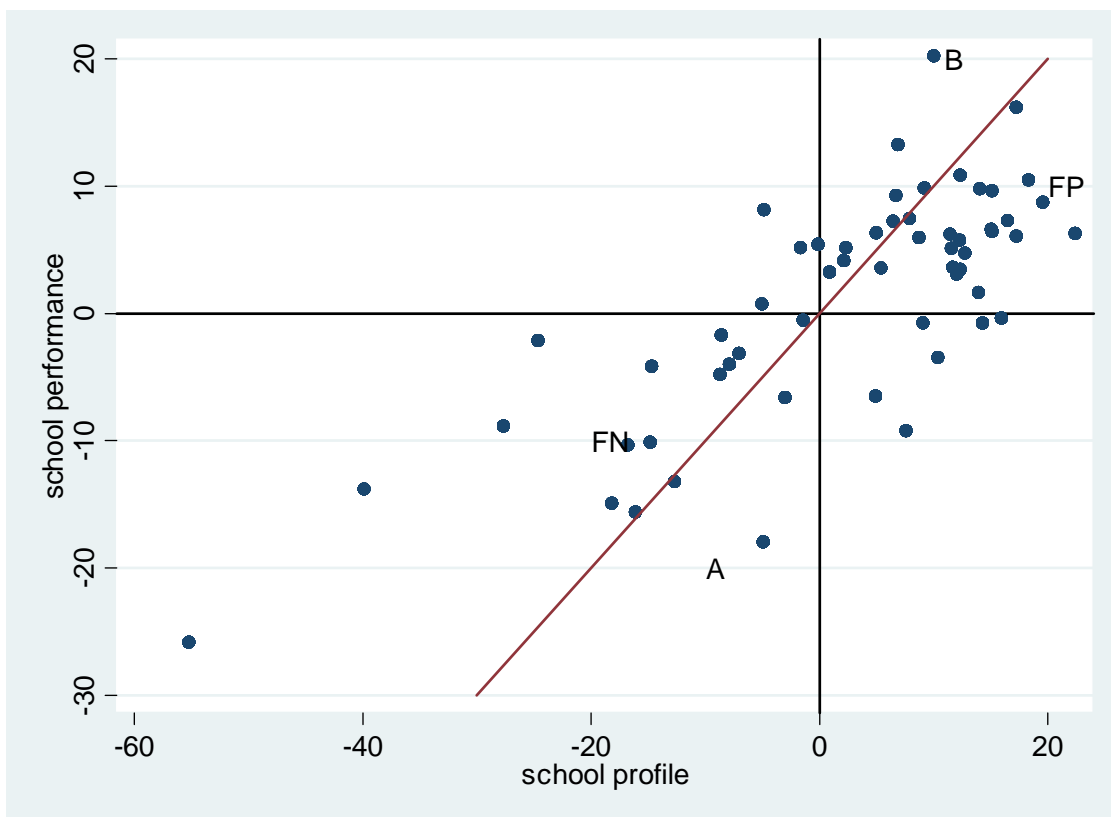
Figure 3: School performance versus school profile.

Figure 4: Uncorrected versus corrected school subsidies.
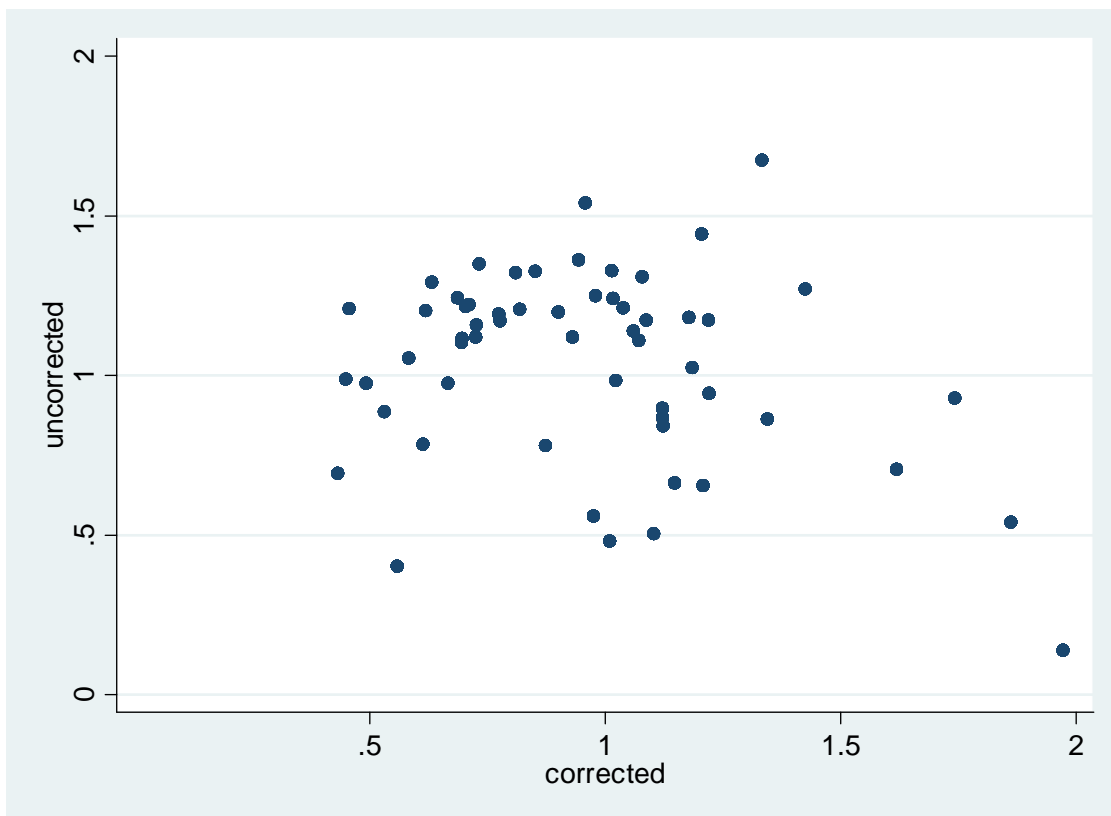
Table 1: Summary statistics for the pupil variables

initial math score	mean	st.dev.	5%-ile	95%-ile
grade 1	69.25	19.62	29.73	94.59
grade 2	70.01	18.19	35.14	94.59

sex	boy	girl		
grade 1	0.50	0.50		
grade 2	0.49	0.51		

language mother	= dutch	≠ dutch		
grade 1	0.91	0.09		
grade 2	0.90	0.10		

language father	= dutch	≠ dutch		
grade 1	0.89	0.11		
grade 2	0.89	0.11		

age	behind	at age	ahead	
grade 1	0.12	0.87	0.01	
grade 2	0.16	0.83	0.01	

mother's highest degree	<secondary	secondary	tertiary (≠univ.)	tertiary (=univ.)
grade 1	0.22	0.37	0.32	0.09
grade 2	0.19	0.37	0.33	0.11

father's highest degree	<secondary	secondary	tertiary (≠univ.)	tertiary (=univ.)
grade 1	0.22	0.40	0.24	0.14
grade 2	0.20	0.40	0.25	0.15

Table 2. Summary statistics for class variables

# of teachers	1	2		
grade 1	0.89	0.11		
grade 2	0.87	0.13		
instruction time	mean	std. dev.	5%-ile	95%-ile
grade 1	6.16	0.86	5	7
grade 2	6.31	0.86	5	7.5
total experience	mean	std. dev.	5%-ile	95%-ile
grade 1	15.15	8.92	3	30
grade 2	17.42	9.52	3	30
class size	mean	std. dev.	5%-ile	95%-ile
grade 1	20.05	3.86	14	26
grade 2	20.12	4.29	14	27
peer effect	mean	std. dev.	5%-ile	95%-ile
grade 1	55.82	2.96	50.32	59.37
grade 2	55.64	3.05	50.46	59.56

Table 3: Selection of pupils over time.

	complete		incomplete	
grade 1	n	$\overline{y}_0$	n	$\overline{y}_0$
tested	2973	69.25	874	67.78
not tested	65	52.08	105	50.86
grade 2	n	$\overline{y}_0$	n	$\overline{y}_0$
tested	3400	70.01	344	70.85
not tested	47	57.49	2	83.79

Table 4: Box-Cox regression results to test for additivity

	θ		$\theta = \lambda$		θ	λ
estimate	1.05		1.14		1.12	1.98
standard error	0.02		0.02		0.02	0.10
H_0	χ^2	p	χ^2	p	χ^2	p
linear: $\theta = 1 \, (= \lambda)$	5.72	0.017	36.07	0.000	138.00	0.000
log-linear: $\theta = 0 \, (= \lambda)$	12775.25	0.000	12889.59	0.000	12991.53	0.000
inverse: $\theta = -1 \, (= \lambda)$	2.9e+05	0.000	2.9e+05	0.000	2.9e+05	0.000

Table 5: Abbreviation and description of the covariates

$time2$	$= 1$	if pupil is currently in second grade, 0 otherwise (thus 0 = in first grade)
$math_0$		initial test score result in mathematics when entering grade 1
$girl$	$= 1$	if girl, 0 otherwise
$ahead\_c$	$= 1$	if pupil is 1 or more years ahead of age not due to school decision, 0 otherwise
$behind\_c$	$= 1$	if pupil is 1 or more years behind age not due to school decision, 0 otherwise
$ahead\_r$	$= 1$	if pupil is 1 or more years ahead of age due to school decision, 0 otherwise
$behind\_r$	$= 1$	if pupil is 1 or more years behind of age due to school decision, 0 otherwise
m_dutch/f_dutch	$= 1$	if mother/father speaks dutch, 0 otherwise
m_edu_sec/f_edu_sec	$= 1$	if mother/father has a secondary education degree, 0 otherwise
m_edu_high/f_edu_high	$= 1$	if mother/father has a tertiary (short type) education degree, 0 otherwise
m_edu_uni/f_edu_uni	$= 1$	if mother/father has a tertiary (long type) education degree, 0 otherwise
duo	$= 1$	if there are two teachers, 0 otherwise
$peer$		average initial test score of all pupils in the same grade at school
it_math		number of hours per week of mathematics instruction in the classroom
$experience$		total number of years experience with teaching
$class\_size$		number of pupils in the classroom

Table 6: Estimation results

math score	model a		model b		model c		model d		model e	
	coeff.	p	coeff.	p	coeff.	p	coeff.	p.	coeff.	p.
constant	60.86	0.000	60.86	0.000	51.88	0.000	66.00	0.000	66.49	0.000
time2	0.14	0.866	0.14	0.866	0.14	0.866	0.14	0.866	-0.21	0.793
$math_0$			0.77	0.000			0.74	0.000	0.74	0.000
girl					-5.21	0.000	-4.91	0.000	-4.92	0.000
ahead_c					4.63	0.210	7.15	0.036	6.92	0.037
behind_c					-4.66	0.000	-4.66	0.000	-4.61	0.000
ahead_r					-1.20	0.782	0.52	0.879	0.63	0.856
behind_r					-10.81	0.000	-4.96	0.003	-4.84	0.005
m_dutch					3.33	0.014	-3.64	0.006	-3.79	0.004
f_dutch					2.46	0.079	-1.31	0.288	-1.31	0.292
m_edu_sec					3.40	0.000	-0.49	0.492	-0.49	0.497
m_edu_high					9.72	0.000	1.60	0.047	1.67	0.041
m_edu_uni					11.58	0.000	3.25	0.016	3.25	0.018
f_edu_sec					1.37	0.138	1.18	0.096	1.19	0.099
f_edu_high					5.20	0.000	3.16	0.000	3.18	0.000
f_edu_uni					6.81	0.000	4.76	0.000	4.74	0.000
duo									-1.85	0.321
peer									1.09	0.007
it_math									2.07	0.035
experience									0.03	0.552
class_size									0.31	0.146
σ_u	19.24		14.32		17.88		13.78		13.81	
σ_v	9.60		6.48		7.94		6.25		11.03	
σ_w	6.78		6.78		6.78		6.78		6.63	

Table 7: Relative performance and relative profile

	low		middle		top	
	perf.	prof.	perf.	prof.	perf.	prof.
uncorrected	-13.85	-17.70	2.59	7.87	10.47	11.72
corrected	-2.16	11.05	1.76	2.24	-2.49	-14.80