Muriel, Alastair; Smith, Jeffrey A.

Working Paper

On educational performance measures

IZA Discussion Papers, No. 5897

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Muriel, Alastair; Smith, Jeffrey A. (2011) : On educational performance measures, IZA Discussion Papers, No. 5897, Institute for the Study of Labor (IZA), Bonn, https://nbn-resolving.de/urn:nbn:de:101:1-201108152046

This Version is available at:
https://hdl.handle.net/10419/51945

DISCUSSION PAPER SERIES

IZA DP No. 5897

On Educational Performance Measures

Alastair Muriel
Jeffrey Smith

August 2011

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

On Educational Performance Measures

Alastair Muriel
Institute for Fiscal Studies

Jeffrey Smith
University of Michigan,
NBER and IZA

Discussion Paper No. 5897
August 2011

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5897
August 2011

ABSTRACT

On Educational Performance Measures*

Quantitative school performance measures (QPMs) are playing an ever larger role in education systems on both sides of the Atlantic. In this paper we outline the rationale for the use of such measures in education, review the literature relating to several important problems associated with their use, and argue that they nonetheless have a positive role to play in improving the educational quality. We delineate several institutional reforms which would help schools to respond "positively" to QPMs, emphasizing the importance of agents' flexibility to change the way they work, and the importance of a sound knowledge base regarding "what works" in raising attainment. We suggest that the present institutional setups in both England and the US too often hold schools accountable for outcomes over which they have little control – but that such problems are far from insurmountable.

JEL Classification: H52, I2, I28

Keywords: school quality, performance measures, education incentives

Corresponding author:

Jeffrey Smith
Department of Economics
University of Michigan
238 Lorch Hall
611 Tappan Street
Ann Arbor, MI 48109-1220
USA
E-mail: econjeff@umich.edu

* We thank participants at the LEMMA / cemmap / ADMIN conference on Measuring School Effectiveness at Oxford, July 5-6, 2010 for helpful comments, Anna Vignoles for comments and Lorraine Dearden for comments, encouragement and extraordinary patience.

1.Introduction

In recent years, quantitative performance measures (QPMs) have come to play an ever-larger role in the educational systems of both England and the United States.England has published schoolperformance tables (popularly known as “league tables”) since the early 1990s, and these tables have grown in both size (as governments added more measures) and in importance, both to parents and to policy discussions, over time. Over roughly the same period, the US has also introduced school-level QPMs at both the state and, more recently, the federal level. The federal efforts in the US, starting with the No Child Left Behind (NCLB) Act in 2001 and continuing more recently with the Obama administration’s Race To The Top initiative (both discussed in more detail below), have partially eroded the traditional federal deference to state and local control of primary and secondary education. In both countries, QPMs have generated both policy controversy and important new research.

The demand for QPMs has its basis in the nature of the education production function and in government provision of educational services. In this paper we lay out the problems that these features of the educational marketplace generate and describe how QPMs might help solve them. We also describe several problems with QPMs as presently implemented. Along the way, we aim to convince the reader that QPMs cannot “do it all,” but that with appropriate institutional changes to better allow education providers to respond to the incentives they embody accompanied by a serious program of additional research to inform those responses, they have much to contribute to educational policy.

The remainder of the paper proceeds as follows: We begin in Section 2 with a quick review of the institutional state of play in England and the US; this background sets the stage for our "big picture" discussion in Section 3 of the various reasons why one might want to introduce performance measures. Section 4 reviews several related literatures to see what support they provide to particular justifications for QPMs. Sections 5 and 6 address potential disadvantages to QPMs in the form of econometric issues and strategic responses, respectively, while Section 7 briefly notes some other issues already well-covered in the literature. Returning to the big picture, Section 8 discusses other institutions that complement QPMs. Section 9 concludes.

2. QPMsin the US and England

In recent years both the US and England have introduced high-profile systems of QPMs in an attempt to improve educational outcomes. In the US, the current national accountability framework owes much of its structure to the No Child Left Behind Act (NCLB) of 2001, which requires all schools receiving federal funding to administer standardized state-wide tests to their students.[1] Schools are required to assess the reading and mathematics performance of students in grades three to eight each year (equivalent to years four to nine in the English education system), with an additional test administered during grades ten to twelve. Schools must show that they are making "adequate yearly progress"(AYP) – though the definition of AYP is left up to the states.Parents of children at schools that fail to meet the AYP target for two years running must be given the option to send their children to a better public school in the same district (if one exists). Failing to meet the target for five years running leads to a school being "restructured" by the district – through closure,transformation into a charter school, or the hiring of a private company to run the school.

[1]See e.g. Peterson and West (2003) for more on NCLB.

Despite these federal requirements, states still retain ultimate control of their accountability systems, because the Act allows them to set their own definition of AYP, and to design (orselect) their own standardized tests.

More recently, the Obama administration unveiled its Race to the Top initiative, encouraging states to compete for federal grants (potentially worth hundreds of millions of dollars) in return for implementing meaningful reforms to their education system. Applications for funding were judged on a range of criteria, with the heaviest-weighted elements being those relating to "improving teacher and principal effectiveness based on performance".[2] Other heavily-weighted criteria aimed to encourage states to adopt *cross*-state (rather than merely state-wide) common standards in their assessments, to implement state-wide longitudinal measurement of pupils' performance, and to foster the creation of charter schools.

While participation in the Race to the Top was purely voluntary, 40 states plus the District of Columbia took part. Despite only 12 states ultimately winning Race to the Top funds, many more signed up to adopt common assessment standards in future,because by doing so they improved their chances of winning.Overall, then, the direction of policy in the UShas been one of increasing harmonization of assessment and accountability regimes across states.

England's move towards test-based accountability began somewhat earlier.The UK education system has always had nationally harmonized assessments at ages 16 and 18 (currently known as GCSEs and A levels, respectively), and in the 1980s many schools began publishing exam

[2]www2.ed.gov/programs/racetothetop/executive-summary.pdf

results in their prospectuses. The system expanded substantially in 1991, with the introduction of standardized achievement tests for all 7 year olds in England's state-funded schools (known as "Key Stage 1 assessments"), and later years saw tests introduced for 11 and 14 year olds (Key Stages 2 and 3, respectively). The results of these tests, as well as GCSE and A level results,are made publicly available, allowing the creation of school performance tables (the "league tables") that enable parents to compare the performance of local schools.

Initially England's school performance tables contained only simple performance metrics, such as the percentage of each school's students attaining five GCSEs at grades A* to C.Recognising the problematic nature of such 'raw' scores, which reflect intake quality as much as school performance (Goldstein and Spiegelhalter, 1996), 1997 saw the introduction of value added measures, showing the average *improvement* of pupils at each school between Key Stages 1 and 2. These were augmented still further in 2007 with the addition of "contextual value added" scores, which adjust raw value added scores to take into account factors outside the schools' control, such as pupils' sex, ethnicity, first language, and family income (as proxied by their eligibility for free school meals).Despite the addition of the value-added based league tables, leading newspapers in England have continued to print only the tables based on outcome levels.[3] High-profile problems with the marking of Key Stage tests in 2008, combined with ongoing opposition to the entire test-based regime by the teacher labor unions, led the government to abolish the tests taken by 14 year olds, replacing them with a system based solely on teacherassessments[4].

[3] See e.g. the Telegraph's GCSE league tables (http://www.telegraph.co.uk/education/leaguetables/8254332/GCSE-league-tables-2010-school-by-school.html)
[4]See e.g. 'Tests scrapped for 14 year olds', BBC News, Oct. 14th 2008 (http://news.bbc.co.uk/1/hi/education/7669254.stm)

3.Performance measures in the big picture

What is the rationale for introducing national QPMs in the market for schooling, when so many markets in our economy(pencils, hamburgers, etc.) function perfectly well withoutthem? The question may sound glib, but answering it requires us to consider the important ways in which the market for primary and secondary schooling differs from other markets. Even before we consider the effects of government involvement, the nature of the product itself raises issues. First, in general, parents do notobserve teacher and school behaviour. This creates a classic principal-agent problem between the parent on the one hand and the teacher and school on the other.[5] In this context, the parent is the "principal" and the school and the teachers are the "agents", retained by the parents to provide educational services to their children. This problem may manifest as the teacher not putting in sufficient effort in either the narrow sense of hours and engagement or in the broader sense of working smart by keeping up with developments in curricula, teaching methods and classroom management. It may also manifest as failure by the school to monitor teachers, to carefully match students to teachers, to hire the optimal teachers and so on.

Second, as with doctors and auto repair shops, the parent consumer often stands at a disadvantage relative to the teacher and the school in terms of knowledge about how best to bring about the desired educational and later life outcomes. Economists call this an information asymmetry; see e.g. Oswald (1986), De Meza and Webb (1987) and Finkelstein and Poterba (2004) for evidence on the nature and effects of such asymmetries in other markets. This

[5] See Prendergast (1999) and Dixit (2002) for classic discussions and Lazear and Gibbs (2009) for a textbook treatment.

information asymmetry makes it difficult for parents to fully evaluate the inputs they do observe, such as the textbooks and assignments their children bring home from school, as well as the broader curricular and management approaches adopted by the school.

Third, the outcome of ultimate interest, namely what sort of adult the child becomes, is not observed until long after much of the educational services have been consumed. Indeed, even many educational outcomes, such as high school completion and university attendance and completion occur years after parents make many of their educational choices. Contrast this with a restaurant meal, where payment does not even occur until after the good has been consumed and its quality revealed.Moreover, unlike restaurant meals, where a bad choice means that the patron is out a few dollars or pounds and where dozens of such meals may be consumed each year, most parents have only one or two children, and make a small number of large (and likely expensive, as they involve residential location) educational choices. This limits both parental learning and risk reduction via diversification.

Worst of all, in terms of making choices about schools and teachers, what the parent really caresabout is not the outcome level, which reflects numerous other factors including the genetic endowment provided by the parent, the home environment, peer influences and so on, but rather a school's long-run valueadded or causal effect on life outcomes. Valueadded refers to the difference that a school makes, relative to some counterfactual. While parents do eventually observe the adult outcomes of their children, they never directly observe the valueadded, even in the longrun. Rather, it must be inferred. This inferential problem is completely general and the

difficult process of finding solutions to it underlies the large econometric literature on program evaluation; see e.g. Heckman, LaLonde and Smith (1999).

The features of the education market just listed suffice to justify some government regulation of providers, such as requiring a teaching certificate for teachers, but do not directly justify either government payment for primary and secondary education or its direct provision. That we have direct government provision via the tax system, coupled with limited (though increasing in both the US and England) amounts of choice within the state system, raises two additional issues. First, an additional principal-agent problem arises between the voters and the government authority that operates the schools. Second, market discipline of schools, already limited due to the issues described above, operates indirectly residential choicesand elections rather than directly through parental choice (as it would in a private market with either vouchers or parental payments). Moreover, state schools pretty much never close, while private ones do.

In sum, the education market differs from other markets in important ways, and would be so even in the absence of state provision of education. The nature of the educational product, the economies of scale that (usually) lead to its production in schools rather than at home, and the specialized knowledge that leads teachers to partially replace parents, combined with government provision, lead to a relative lack of market discipline and thus to concerns about whether or not providers have sufficient incentives to provide value for money.

What can QPMs do to address the lack of incentives for educational agents – schools and teachers – implied by the nature of the educational production function combined with government provision of schooling? They can potentially do several things:

1. They may induce additional effort on the part of administrators and teachers in the form of working harder and working longer;

2. They may induce schools to make better choices regarding recruitment, training and dismissal of teachers and, not unrelated, they may alter the set of individuals who choose teaching careers;

3. They may induce schools and teachers to make better curricular choices, particularly by paying more attention to the research literature;

4. They may provide parents with information that allows them to better bring pressure to bear on under-performing schools, whether via their voting behaviour, their residential location choices, their choices of schools within the government system in contexts that allow such choices, and via their choices to leave the public system altogether for either private schools or home schooling.

4. Performance management, power and knowledge

In order for performance measures to have the positive effects posited in the preceding section, schools and teachers must have two things: the power to make choices that will affect the performance measures and the knowledge to make good choices rather than bad ones.

Forexample, QPMs will not improve the quality of teachers hired if, e.g., principals (or head teachers, in the UK vernacular) must hire from among the qualified teachers the one with the most seniority or if principals have no idea how to distinguish a good teacher from a bad one and so choose the candidate who will provide the best lunchtime conversation.With the above discussion in mind, we now consider the degree to which principals and teachers in the US and England have the power, and the knowledge, to respond positively to QPMs.

4.1 Power

In the US, the degree of flexibility enjoyed by teachers and principals varies significantly between states, between different school districts within each state, and between schools of different types. To a first approximation, the rules governing matters such as teacher recruitment, pay scales, the conditions in which teachers work, and the curriculum they are expected to teach in US public schools are the result of decisions made by (at least) three levels of governance: the school district (often in negotiation with teachers' unions), the state government and (to a lesser extent) the federal government. The balance of power between the school district and the state varies substantially between states, however, with some states allowing school districts to set their own teacher pay and conditions in negotiation with teachers' unions, while other states forbid such collective bargaining, imposing pay scales through state-wide legislation. Neither arrangement, however, leaves significant latitude to school principals themselves, as these systems base teacher pay almost exclusively on years of experience.

In regard to school type, the advent of public charter schools has led to important variation in the degree of flexibility enjoyed by principals in US schools. Charter schools typically escape many

of the collectively bargaining limitations regarding hiring, firing, pay and conditions that bind other public schools (see e.g. Abdulkadirogluet al, 2010), though the particulars vary widely among the states. Principals in such schools enjoy greater freedoms than their peers in regular public schools.

In England, unlike the US, head teachers' flexibility does not vary a great deal from one region to another. Local Authorities (LAs), England's closest analogue to school districts, generally do not have the power to set their own teacher pay and conditions, which are instead agreed by the national government and the teachers' unions.The majority of state-funded schools in England are so-called "community schools," which are subject to a reasonably large degree of LA oversight. These schools receive their funding according to formulae set by their LA, and use central services (such as admissions processes) provided by their LA. They must also adhere to a raft of nationally determined agreements and constraints: teaching within the confines of England's National Curriculum, and adhering to a national system of teacher pay and conditions (agreed each year with the major teachers unions).

Just like the US, however, different types of schools enjoy varying degrees of freedom to deviate from these collective agreements. Successive governments have introduced a range of alternative school types to the English system – from "voluntary aided schools" to "grant maintained schools" to "trust schools" - all enjoying varying degrees of freedom from LA control. The latest (and fastest growing) iteration of this idea is the new class of "academy schools" (a sort of English version of charter schools), introduced in the year 2000,with the authority to diverge from both collective wage agreements and from the National Curriculum. The head teachers in

these school types, just like their charter school counterparts in the US, generally enjoy significantly greater freedoms than their peers in standard community schools.

4.2Choosing teachers

The recent literature suggests that teachers matter a lot to educational outcomes. We might summarize the conclusions of this literature as (i) teachers vary in their effectiveness, and (ii) the magnitude of this variation implies that an excellent teacher can overcome deficits due to poor family background and other factors. Among the more recent studies, Rivkin et al. (2005), using data from Texas, estimate that a one standard deviation improvement in teacher quality (as measured by value added) corresponds to a class size reduction of more than ten students. Aaronson et al. (2007), using data from Chicago public schools, estimate that a one standard deviation improvement in teacher quality raises student mathematics scores by one fifth of the average yearly gains, while Slater et al. (2009), using data on English schools, find even higher variability in teacher effectiveness, with a one standard deviation improvement in teacher quality raising GCSE test scores by at least 25% of a standard deviation.

Given that teachers matter so much to outcomes, the ability to identify, recruit and retain highly effective teachers (and the ability to reform or fire ineffective ones) is one of the key powers wielded by a school principal. In practice, however, two formidable obstacles limit the exercise of this power. The first obstacle is the difficulty in identifying strong teachers to recruit. The literature provides persuasive evidence that teaching effectiveness varies only weakly with observed teacher characteristics, with the exception of experience. For example, Kane et al. (2007) find that teachers' academic backgrounds, including their undergraduate GPA and the

selectivity of their undergraduate institution, do not have predictive power for their value added, but that effectiveness improves with the first few years of experience. Aaronson et al. (2007)and Slater et al. (2009) reach similar conclusions using different data. Rockoff et al (2011) attempt an even more radical approach – administering an in-depth survey to new teachers in New York City, including a range of non-traditional measures such as personality traits, feelings of self-efficacy and cognitive ability – yet find that few of these measures are significantly correlated with subsequent teacher effectiveness. Contrary findings come from Clotfelter et al. (2007), using administrative data from North Carolina,who do find a meaningful effect of both teacher qualifications and teacher test scores on teachingperformance, with the effect being greater for mathematics scores than for reading.Encouragingly, Rockoff and Speroni (2010) find that evaluations of prospective teachers based on long interviews (including mock lessons) also have predictive power for teachers' subsequent effectiveness. Nonetheless, the weight of evidence from this literature suggests that high quality teachers are easier to identify *ex post*than they are to recruit based on their CVs.

At retention time, principals can supplement observed teacher characteristics with their own direct observations of the teacher in making decisions about which teachers to let go. Jacob and Lefgren (2005) provide encouraging evidence on this dimension, comparing school principals' subjective assessments of teachers with the traditional determinants of teacher compensation (teacher's level of education and experience) as well as measures of valueadded.They find that subjective principal assessmentspredict teacher performance significantly better than education or experience, though not as well as valueadded quality measures. In particular, principals appear

to be able to identify the teachers with the very best and worst valueadded, but are less able to distinguish between teachers towards the middle of the distribution[6],

In summary, the literature suggests that hiring teachers presents a difficult problem to principals, given that we presently lack the knowledge to predict teacher effectiveness based on observed characteristics. This highlightsthe value of developing instruments better able to accomplish this important sorting task. On the other hand, some evidence suggests that principals can identify their least effective teachers so that, if they have discretion over retention, they can use that power to raise average teacher quality in their school.

The second obstacle to principals' exercising freedom in recruitment and retention of strong teachers is institutional: in both the US and England, powerful teachers' unions constrain principals' ability to reward strong teachers and to fire weak ones. Teacher salaries in both countries remain determined almost entirely by teachers' years of tenure (rather than, say, their observed effectiveness, or the supply of teachers with particular skills). England experimented with the introduction of 'performance related pay' in the early 2000s, but in practice the system operated as a pay rise for all teachers (Atkinson et al., 2004), rather than a reward for the very best, and the rigid system of automatic pay increases survived.Principals wishing to fire poorly performing staff also face substantial legal and procedural obstacles. While disciplinary procedures exist in both countries for the removal of poorly-performing teachers, they cost so much in terms of time and effort that principals almost never invoke them[7].

[6]Dearden et al (forthcoming) provide evidence that students themselves appear well able to identify stronger and weaker teachers – feedback that principals could no doubt make use of to supplement their own assessments.
[7] For England see e.g. www.guardian.co.uk/education/2010/jul/04/struggling-teachers-woodhead-claims-dismissed, for the US see e.g. www.newsweek.com/2010/03/05/why-we-must-fire-bad-teachers.html

The salary and job security protections enjoyed by teachers make the hiring decision substantially more binding in the school sector than in the private sector. Employers in the private sector may have just as much difficulty identifyinghighly productive workers *ex ante* as school principals – but private sector firms have far greater freedom to adjust remuneration (and fire unproductive staff) accordingly. The inability to reverse hiring decisions in the (public) education sector, or to adjust remuneration in line with performance,means that hiring decisions have higher stakes than in the private sector.. All the more dispiriting, then, that head teachers will only discover whether or not they made the "right" choice when there is little (if anything) they can do about it.

4.3 Choosing how to teach

Do teachers have the knowledge required to become better teachers, if they want to do so in response to a regime of QPMs? Put differently, does the literature provide a settled, evidence-based consensus about how to be a good teacher?Teacher characteristicsmay be only weakly related to teacher effectiveness, but are there nonetheless certain *actions, methods* or *curricula*that consistently improve student outcomes?It seems uncontroversial to state that, at present, no such consensus in this regard exists.

This lack of consensus is certainly not due to a lack of published papers on the topic. Decades of effort and millions of dollars of research funding have spawned a voluminous literature on teachers and teaching – but only a small fraction of this money supported statistically rigorous evaluations of specific, clearly-defined programsand interventions. Indeed, even analyses in the

education literature described as “program evaluations” are often little more than surveys of participants. To take one example among many,consider the 148 evaluations of America’s “Teaching American History”program (which offerspleasant summer courses and workshops to high school history teachers) summarized by Humphrey et al. (2005). The majority of these ”evaluations”rely on teachers’ self-reports of the effectiveness of the program.Only a quarter of them actually analysed the work produced by the students of teachers who took part in the program.Surveying the state of the literature regarding “best practice” for teachers more generally, Humphrey et al. echo Wilson’s (2001) argue that most studies suffer from a sort of circular reasoning: “Too often, the studies start with an assumption that they are examining a teacher with good instructional practice, describe that teaching and the teacher’s knowledge and skills, and then claim that that knowledge and those skills are the characteristics of effective teachers. As a result, it is unclear from the research why certain teacher behaviors lead to good teaching, why certain knowledge and beliefs are associated with good teaching, and the kinds of knowledge that contribute to good teaching.”

For the time being, this dearth of knowledge about “what works” in teaching reflects the abject failure of an entireacademic field, and seems likely to remain an ongoing obstacle to attempts to encourage educators to respond positively to performance management systems.

5. Econometric issues

The literature on QPMs in education highlights two important econometric concerns: imprecision and bias. Imprecision means that the estimates have large standard errors, which result from what economist Art Goldberger jokingly called “micro-numerosity” or, more simply,

small sample sizes. Most classes are not very large. Average class size in the US is around 23 pupils at both primary and secondary levels, while in the UK the average class size is 25 pupils at primary level, 20 pupils at secondary level (OECD, 2010). As noted by Kane and Staiger (2002), such small classes imply very imprecise estimates of performance for individual teachers, whether measured in levels or in valueadded. Aggregating up to the school level helps less in practice than one might expect, because most extant QPMs rely on exams given only to one grade level in a school, implying a school level sample size that is only a small multiple of that for a single classroom. Moreover, most extant QPMs use only a single test score, rather than having students take multiple tests so as to reduce measurement error at the student level.

Wilson and Piebalga's (2008) finding that almost half of England's secondary schools have performance that is statistically indistinguishable from the national average (using contextual value added measures) follows immediately from modest school-level sample sizes. So does the Leckie and Goldstein (2009 and this volume) point that current valueadded estimates in the English system have only modest predictive power for estimates six years in the future. While some of this predictive failure results from changes in underlying school quality over time, most likely results from estimation error.

Turning now to bias, we note that in addition to the basic point that QPMs, like those in the US under NCLB, often consist of outcome levels when we care about valueadded, even valueadded measures may embody substantial bias, depending on their construction and on the mechanisms that sort students into schools and classrooms. Rothstein (2009, 2010) lays out the econometric issues in detail while Kane and Staiger (2008) provide some evidence based on random

assignment of teachers to classrooms in Los Angeles that certain contextual valueadded measures can do a good job of replicating experimental teacher effects in that context.

In sum, small samples and single tests mean big standard errors for school-level QPMs. Averaging performance across years, testing more grades, and testing more than once can help with the sample size problem, but at a cost. The Kane and Staiger (2008) paper provides the key to solving the bias issue: do the research required to determine what you need to condition on in a contextual value-added model in order to get the right answer. Dearden, Miranda et al. (2011), in this volume,provide an example of this approach in action.

6. Strategic responses

Schools and teachers can raise their performance on QPMs by working harder or working smarter or by trying to "game" the measures in ways that raise their measured performance without improving their actual performance. Such strategic responses to QPMs are well documented in many literatures; see e.g. Courty and Marschke (2011) for job training programs, Smith (1995) and Kalman and Friedman (2009) for health care and Wallsten (2000) for research and development subsidies.

Strategic responses take a variety of forms depending on the particular QPMs in use and on the legal and practical limits on school and teacher behaviour. For example, when the QPM consists of mean test score levels, schools may manipulate the set of students eligible to take the test, as documented in Cullen and Reback (2006), or simply encourage weak students to stay home on test days. When the QPM relies on a "threshold" of performance, schools may focus their

energies on *marginal* pupils whose expected performance lies near the threshold, while neglecting both the strongest performing children (who will pass anyway) and the weakest performers (the "lost causes").Neal and Schanzenbach (2010) provide evidence that this occurred in the Chicago Public School system in response to the introduction of NCLB in 2002, while Wilson et al. (2006) show that some(but not all) head teachers in England willingly admitted to engaging in such behaviour in response to England's league tables.

Yet another possible strategic response to QPMs is to change the *mixture* of courses taken by a school's pupils. If some courses make it easier for schools to do well on their performance measures, pupils may be encouraged to take such courses regardless of whether or not they are in the pupils' long term interests. Jin et al. (forthcoming) provide evidence that this has occurred in England's school system in response to league tables,and this forms a key concern of England's recent review of vocational education courses (Wolf, 2011).

In some contexts, schools, particularly non-traditional schools such as charters and academies, may have the leeway to attempt to "cream skim" students who will help them do well on the performance measures. Such selective enrolment has long plagued job training programs in the US. The incentive, and thus presumably the behaviour, arises mainly when using performance measures based on levels, as schools may have an easier time assessing likely outcome levels than assessing likely value-added. See e.g. Epple and Romano (2008) in an education context and Bell and Orr (2002) in a job training context.

Last, but hardly least, the most straightforward strategic response to QPMs consists of simply having teachers cheat on the tests that underlie the performance measures, either by giving out the answers or by filling in or correcting the answers after students turn in their forms. Jacob and Levitt (2003) provide evidence of cheating in the Chicago Public School system. More recently, a widespread cheating scandal emerged in the Atlanta Public Schools (see e.g. Severson, 2011) and Dee, Jacob and McCrary (2011) uncovered manipulation in the New York State "Regents Exams" which certify a higher level of attainment than standard high school graduation.

The literature reveals ubiquitous strategic responses to QPMs. These responses distort school and teacher behaviour, which harms students, and consume real resources. Minimizing such responses requires careful design and, as emphasized by Heinrich and Marschke (2010), will likely represent an on-going process as system designers adjust to the observed strategic responses of educators.

7. Other issues

Partly for reasons of space and partly because they have received a lot of attention elsewhere, we have avoided extended discussions of a few nonetheless important issues, which we briefly touch on here. First, and most important, stand the basic issues around solutions to principal-agent problems. If the government, acting on behalf of parents, seeks valueadded, then it should not reward outcome levels, as it does in the present NCLB system in the US. Here England excels with its contextual value-added measures, though it would do better to drop the levels measures entirely. Heckman et al. (2011) present a clear theoretical discussion of the issues associated with

rewarding levels rather than gains in the context of job training programs; Balou et al. (2004) provide a discussion in an educational context.

Another basic tenet from the standard principal-agent literature holds that if parents(or governments) want the agent to do several things -e.g. increase the student's knowledge of maths, language arts, history and the arts - but only reward progress on a subset of them, schools will focus mainly on the subset and neglect the others.As Wilson et al. (2006) put it in their analysis of the English system, "what gets measured gets done".

Both the US and English systems suffer from this malady to some extent; remedies include additional performance measures or reductions in the rewards and punishments associated with the existing measures.

Second, we have left to the side issues surrounding how parents use QPMs. Parents' understanding of QPMs has implications both for presentation and for understanding the effects presenting them at all. We refer the reader to Hastings and Weinstein (2007) for further discussion and Burgess et al. (2010) for some evidence on what happened when Wales eliminated its school league tables.

Third, we have not had much to say about heterogeneity, which matters here in a couple of senses. For one, schools have many dimensions other than the strictly academic, such as their level of discipline, the religious or ethical training they provide, and their athletic and arts programs. To the extent that parents care about these aspects, QPMs that focus solely on

academic performance may not help them. Schools may also differ in terms of their ability to teach different sorts of students, conditional on the value that they add to their existing students (see Dearden, Micklewright et al., 2011, in this volume, for evidence in this regard). For example, of two schools with the same estimated contextual valueadded, one might do well with students who need remedial help and structure while the other might do well with very bright children. This is the primary and secondary school analogue of the mismatch problem addressed in the higher education literature; see e.g. Dillon and Smith (2011) and, more broadly on heterogeneity in higher education, Smith (2008).

8. Complements

The discussion to this point suggests, in our view, that current performance systems in the US and England ask too much of their QPMs. This section addresses the question of alternatives or, as we would have it, complements to performance management via QPMs.

First, flexibility matters. Principals with no control over personnel decisions, such as hiring, firing and salary determination, or over decisions on student admissions or ejections, or over the school's broad focus or narrow curricular choices, have limited tools indeed with which to respond to the incentives implicit in a system of QPMs. Similarly, teachers with no control over which students end up in their classrooms or over curricular choices face similarly strong limits on what they can do to improve their performance. Relaxing some or all of these restrictions, and thereby providing agents with ways to improve their measured performance other than the sorts of gaming described in Section 6 should lead to better outcomes and fewer scandals.

Second, a knowledge base complements QPMs by providing the means for schools to improve their performance by improving what they do and how they do it. In the US, the Department of Education's "What Works Clearinghouse" has played this role since its establishment in 2002. Like the Cochrane Collaboration, its inspiration from the medical world, it provides rigorous, evidence-based advice for U.S. educators.[8]Its nickname, the "Nothing Works Clearinghouse" testifies to the volume of work that remains to produce a compelling evidence base in education. To the best of our knowledge, no analogous institution exists in England. The UK's largest ever educational research programme, the Teaching and Learning Research Programme, is presently drawing to a close. Sadly, the £40 million spent on over 1,500 "research-informed" resources for educators and policymakers yielded only a small number of statistically rigorous evaluations of specific interventions.

School choice, either within the government system or more broadly, empowers parents to respond to the information provided in QMPs and also, in cases where governments schools face enrolment and funding losses if parents fail to choose them, strengthens whatever other incentives a performance management system provides. Our view of choice as a complement to QPMs goes against the common, and we think misguided, view that choice and performance represent alternative roads to the same destination.Well-designed QPMs empower educational consumers to make better choices.

Fourth, there exist other ways to motivate teachers and principals to work hard and to work smart besides QPMs. One consists of the professionalization of education, as discussed in the literature

[8]http://ies.ed.gov/ncee/wwc/

on educational sociology. The institutions of professionalization, which include but are not limited to undergraduate and graduate programs in education, seek to create and reinforce norms of behaviour that guide teachers and principals to do "the right things" even with no onelooking and even when they receive no direct reward for doing so. See e.g. Abbott (1988) or Lortie (2002) on professionalization. Related to, but distinct from, professionalization, is the notion of intrinsic motivation. Hiring teachers who love to teach and to teach well will help to obtain the desired behaviours in the classroom and the principal's office. See Murray (1988) on intrinsic motivation for teachers and the related discussion, in a job training context, in Heckman, Smith and Taber (1997). Academic readers who doubt the importance of these professional norms and intrinsic motivation should consider why so many professors at research universities devote themselves to excellence in undergraduate teaching, even in the presence of zero (or even negative) financial incentives for doing so.

9. Conclusion

In a broad sense, this paper applies the more general analysis of James Q. Wilson's (1989) excellent tome *Bureaucracy* to the particular case of primary and secondary education. The nature of the educational production function, with its long lags and expert inputs combined with the additional troubles brought about by large-scale government provisionmake designing institutions that make the best use of educational resources difficult indeed. The recent enthusiasm for QPMs and related institutional regimes of reward and punishments as a means for solving these problems flows naturally out of these difficulties, combined with declining costs of data collection and information processing. In our view, the present institutional setups in the US and England too often reward and sanction principals and teachers for outcomes over which they

have little control, whether because the QPMs represent levels rather than value-added, or because they lack the knowledge or the power to effect real change. Not surprisingly, this situation results in frustration, conflict, and a lot of strategic behaviour.

In many cases, the problems with the current setup would yield to institutional change and/or further research, the latter aided by improved data collection. For example, in both the US and England, central governments could require that states or LAs provide researchers with access to student-level data linked to teacher and school data (with appropriate institutional privacy protections), in return for funding. Such data, particularly when combined with information on variation in policies and curricula, would increase the pace of knowledge creation and help to relax the knowledge constraint on schools and teachers that want to do better, particularly if combined with deliberate variation in policy over space and time. Recognition that performance management may complement choice, rather than substitute for it, would also move things forward, as would, particularly in the US, a change in focus from QPMs based on outcome levels to QPMs more like the contextual valueadded measures presently in use in England.

In short, while the educational context admits of no magic bullets, our analysis suggests many straightforward measures that would yield valuable improvements at the margin.

Bibliography

Aaronson, D, Barrow, L and Sander, W (2007), 'Teachers and Student Achievement in the Chicago Public High Schools', *Journal of Labor Economics*, 25(1):95-135

Abbott, A. (1988),*The System of Professions: AnEssay on the Division of Expert Labor*. Universityof Chicago Press, Chicago, IL.

Abdulkadiroglu, A, Angrist, J, Dynarski, S, Kane, T and Pathak, P (2011),'Accountability and Flexibility in Public Schools: Evidence from Boston's Charters And Pilots', *Quarterly Journal of Economics*126: 699-748

Atkinson, A, Burgess, S, Croxson, B, Gregg, P, Propper, C, Slater, H and Wilson, D, (2004), 'Evaluating the Impact of Performance-Related Pay for Teachers in England', Centre for Market and Public Organisation Working Paper no. 04/113

Balou, D, Sanders, W and Wright, P (2004), 'Controlling for Student Background in Value-Added Assessment of Teachers', *Journal of Educational and Behavioural Statistics,* 29(1): 37-65

Bell, S and Orr, L (2002), 'Screening (and Creaming?) Applicants to Job Training Programs: the AFDC Homemaker–Home Health Aide Demonstrations', *Labor Economics* 9(2): 279-301

Burgess, D, Wilson, D and Worth, J (2010) `A Natural Experiment in School Accountability: The Impact of School Performance Information on Pupil Progress and Sorting', CMPO Working Paper no. 10/246.

Clotfelter, C, Ladd, H and Vigdor, J (2007), `How and Why Do Teacher Credentials Matter for Student Achievement?',National Bureau of Economic Research, Working paper no. 12828

Courty, P and Marschke J (2011), 'Measuring Government Performance: An Overview of Dysfunctional Responses', In *The Performance of Performance Standards*, Heckman, J, Heinrich, C, Courty, P, Marschke, G and Smith, J(eds.) Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, pp. 203-230

Cullen, J and Reback, R (2006),`Tinkering Toward Accolades: School Gaming Under a Performance Accountability System', National Bureau of Economic Research working paper no. 12286

De Meza, D and Webb, D (1987), 'Too Much Investment: A Problem of Asymmetric Information', *Quarterly Journal of Economics*102(2): 281-292

Dearden, L, Muriel, A and Zagatti, G (2011), 'Pupil ratings and teacher performance,' Unpublished manuscript, Institute for Fiscal Studies

Dearden, L, Micklewright... REFERENCE IN THIS VOLUME

Dearden, L, Miranda... REFERENCE IN THIS VOLUME

Dee, T, Jacob, B and McCrary, J (2011), 'Manipulation in the Grading of New York's Regents Examinations', Unpublished manuscript available at www.econ.berkeley.edu/~jmccrary/Dee_Jacob_McCrary2011.pdf

Dillon, E and Smith, J (2011), 'The Determinants of Mismatch between Student Ability and College Quality', Unpublshed manuscript, University of Michigan

Dixit, A (2002), 'Incentives and Organizations in the Public Sector: An Interpretive Review',*Journal of Human Resources* 37(4): 696–727.

Epple, D and Romano, R, (2008) 'Educational Vouchers and Cream Skimming', *International Economic Review*, 49(4): 1395-1435

Finkelstein, A and Poterba, J (2004), 'Adverse Selection in Insurance Markets: Policyholder Evidencefrom the U.K. Annuity Market', *Journal of Political Economy* 112(1): 183-208

Goldstein, H. and Spiegelhalter, D. J. (1996),'League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance.'*Journal of the Royal Statistical Society*: Series A, 159: 385-443.

Hastings, J and Weinstein, J (2007),'Information, School Choice, and Academic Achievement:Evidence from Two Experiments', *Quarterly Journal of Economics* 123: 1373-1414

Heckman, J, Heinrich, C and Smith, J(2011), 'A Formal Model of a Performance Incentive System', In *The Performance of Performance Standards*, Heckman, J, Heinrich, C, Courty, P, Marschke, G and Smith, J(eds.) Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, pp. 29-64

Heckman, J, LaLonde, R, and Smith, J(1999), `The Economics and Econometrics of Active Labor Market Programs', in Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics, Volume 3A*. Amsterdam: North-Holland, 1865-2097

Heckman, J, Smith, J and Taber, C (1997), 'What do Bureaucrats do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance into the JTPA Program', In Libecap, G (ed), *Advances in the Study of Entrepreneurship, Innovation and Economic Growth, Volume 7: Reinventing Government and the Problem of Bureaucracy*, Grenwich Conn, JAI Press: 191-218

Heinrich, C, and Marschke, G(2010), 'Incentives and Their Dynamics in Public Sector

Performance Management Systems', *Journal of Policy Analysis and Management* 29(1): 183–208

Humphrey, D, Chang-Ross, C, Donnelly, M, Hersch, L and Skolnik, H (2005), *Evaluation of the Teaching American History Program*, Jessup, MD: U.S. Department of Education

Jacob, B and Lefgren L (2007), 'Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education', *Journal of Labor Economics* 26(1): 101-136

Jacob, B and Levitt, S (2003), 'Rotten Apples: an Investigation of the Prevalence and Predictors of Teacher Cheating', *Quarterly Journal of Economics*, 118(3): 843-877

Jin, W, Muriel, A and Sibieta, L (2011), `Subject and Course Choices in England: Insights from Behavioural Economics', Centre for Understanding Behaviour Change research paper, Department for Education, London

Kane T, and Staiger, D (2002), 'The Promise and Pitfalls of UsingImprecise School AccountabilityMeasures', *Journal of Economic Perspectives* 16(4): 91-114

Kane T, Rockoff, J and Staiger, D (2008),'What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City',*Economics of Education Review*, 27(6): 615-631

Kane, T and Staiger, D (2010), 'Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation', National Bureau of Economic Research Working Paper no.14607.

Kelman, S and Friedman, J (2009) `Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service.' *Journal of Public Administration Research and Theory* 19: 917-946.

Lazear, E and Gibbs, M (2009) *Personnel Economics in Practice*, Wiley & Sons

Leckie, G and Goldstein, H (2009), 'The Limitations of Using School League Tables to Inform School Choice', *Journal of the Royal Statistical Society*, 172(4): 835-851

Leckie, G and Goldstein, H, (2011), THIS VOLUME REFERENCE

Lortie, D (2002), *Schoolteacher: a Sociological Study,* Chicago: University of Chicago Press

Murray, C (1988), *In Pursuit of Happiness and Good Government*, New York: Simon & Schuster

Neal, D and Schanzenbach, D (2010), 'Left Behind by Design: Proficiency Counts and Test-Based Accountability', *Review of Economics and Statistics* 92(2): 263-283

OECD (2010), *Education at a Glance: OECD Indicators*, Organisation for Economic Co-operation and Development, Paris

Oswald, A (1986), 'Unemployment Insurance and Labor Contracts Under Asymmetric Information: Theory and Facts', *American Economic Review*, 76(3): 365-377

Rockoff, J and Speroni, C (2010), 'Subjective and Objective Evaluations of Teacher Effectiveness', *American Economic Review*, 100(2): 261-66

Peterson, P and West, M (2003), *No Child Left Behind? The Politics and Practice of School Accountability*, Washington DC: Brookings Institution Press.

Prendergast, C, (1999), 'The Provision of Incentives in Firms',*Journal of Economic Literature* 37(1): 7–63

Rockoff, J, Jacob, B, Kane, T and Staiger, D (2011), 'Can You Recognize an Effective Teacher When You Recruit One?', *Education Finance and Policy* 6(1): 43-74

Rothstein, J (2009), 'Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables', *Education Finance and Policy* 4(4): 537-571

Rothstein J (2010), `Teacher Quality in Educational Production: Tracking, Decay and Student Achievement,' *Quarterly Journal of Economics* 125(1): 175-214.

Rivkin, S, Hanushek, E and Kain, J (2005), 'Teachers, Schools and Academic Achievement', *Econometrica* 73(2): 417-458

Severson, K (2011), 'Systematic Cheating is Found in Atlanta's School System', *New York Times*, July 5th 2011 (http://www.nytimes.com/2011/07/06/education/06atlanta.html)

Smith, P (1995), 'On the Unintended Consequences of Publishing Performance Data inthe Public Sector', *International Journal of Public Administration*, 18 (2/3): 277-310

Smith, J (2008), 'Heterogeneity and Higher Education', in McPherson, M and Schapiro, M (eds.),*Succeeding in College: What it Means and How to Make it Happen*. New York: College Board. 131-144.

Slater, H, Davies, N and Burgess, S (2009), `Do teachers matter? Measuring the variation in teacher effectiveness in England', Centre for Market and Public Organisation Working Paper no. 09/212

Wallsten, S (2000), 'The Effects of Government-Industry R&D Programs on Private R&D: The case of the Small Business Innovation Research Program,'*Rand Journal of Economics* 31(1): 82-100.

Wilson, D, Croxson, B and Atkinson, A (2006) 'What Gets Measured Gets Done', *Policy Studies*, 27(2): 153–71

Wilson, D and Piebalga, A (2008), `Accurate Performance Measure but Meaningless Ranking Exercise? An Analysis of the English School League Tables', CMPO Working Paper no. 07/176

Wilson, J (1989) *Bureaucracy: What Government Agencies Do and Why They Do It*, New York: Basic Books.

Wilson, S (2001), 'Research on History Teaching' in Richardson, V (ed.), *Handbook of research on teaching* (4th ed., pp. 527-544). Washington, D.C.: American Educational Research Association

Wolf, A (2011), *Review of Vocational Education – the Wolf Report*, Department for Education, London