DOCUMENT DE TRAVAIL DT/2005-05

The Relationship between Educational Expenditures and Outcomes

François LECLERCQ

DIAL • 4, rue d'Enghien • 75010 Paris • Téléphone (33) 01 53 24 14 50 • Fax (33) 01 53 24 14 51
E-mail : dial@dial.prd.fr • Site : www.dial.prd.fr

THE RELATIONSHIP BETWEEN EDUCATIONAL EXPENDITURES AND OUTCOMES[1]

François Leclercq
DIAL, Université Paris 1
UNESCO
f.leclercq@unesco.org

Document de travail DIAL
Avril 2005

ABSTRACT

This paper presents a survey of the large empirical literature in economics that has sought to examine the relationship between educational expenditures and outcomes in both developed and developing countries. The main feature of this literature is the remarkable lack of consensus about the results of standard studies using the 'education production function' conceptual framework, whether at the macro or at the micro level. Experimental evidence that has recently started to accumulate may provide more reliable guidance to policy interventions aimed to improve attainment and achievement. Another strand of literature is emphasizing the incentives structure of the school systems, which affects the way in which available school resources are combined to 'produce' outcomes. However, the ability of economists to adequately model the functioning of schools could be further enhanced by making use of insights from other social sciences, e.g. social psychology and sociology, pertaining to the behavior of teachers and students. Although they remain quite marginal to the field, recent behavioral economics papers may provide a basis for such a renewal of the economics of education.

Key words: Economics of education, education production function, randomized experiments, natural experiments, school resources, school incentives, teachers.

JEL Code: H52, I2.

RESUME

Cet article présente une synthèse critique des travaux consacrés par les économistes de l'éducation à la relation entre dépenses d'éducation et résultats scolaires. Qu'ils portent sur les pays développés ou les pays en développement, et qu'ils utilisent des données agrégées au niveau des pays ou des données individuelles, les travaux utilisant le cadre conceptuel standard – la « fonction de production éducative » – n'ont pas établi de régularité empirique incontestable. L'approche dite « expérimentale » utilisée dans quelques travaux récents pourrait offrir des résultats plus robustes quant à l'impact de politiques éducatives spécifiques sur le nombre d'années d'études et le niveau de connaissances atteints par les élèves. Un nouvel ensemble de travaux s'intéresse désormais aux incitations données aux enseignants par les systèmes scolaires, qui déterminent la façon dont les ressources à la disposition des écoles sont utilisées pour « produire » des résultats scolaires. Cependant, la capacité des économistes à modéliser de façon adéquate le fonctionnement des écoles pourrait être améliorée par la prise en compte de concepts empruntés à d'autres sciences sociales, comme la psychologie sociale ou la sociologie ; bien qu'ils restent rares et soient encore peu cités, quelques articles récents d'économie « comportementale » pourraient conduire à un tel renouveau de l'économie de l'éducation.

Mots clés : Économie de l'éducation, fonction de production éducative, expériences aléatoires, expériences naturelles, ressources des écoles, incitations, professeurs.

[1] This paper was prepared as a background paper for the EFA Global Monitoring Report 2005: *Education for All: The Quality Imperative*, Paris: UNESCO Publishing, 2004.

Contents

List of tables

1. INTRODUCTION

This paper presents a survey of the large empirical literature in economics that has sought to examine the relationship between educational expenditures and outcomes in both developed and developing countries. This introductory section starts by situating the economic approach to the outcomes of school education within the field of the economics of education (1.1), before making general remarks on the state of the existing literature (1.2) and presenting the outline of the paper (1.3).

1.1. The economic approach to the outcomes of school education

The contemporary economic approach to education started developing from the late 1950's onwards with Jacob Mincer's application of human capital theory to the measurement of the economic return to education as the impact that the number of years of schooling an individual received has on her earnings. While this view was challenged, notably through the screening and signaling models proposed in the 1970s, the returns to education literature has kept on innovating conceptually and methodologically, and it is now established that human capital acquired through schooling has a causal impact on earnings that can be measured econometrically (see Card, 1999, for a survey). Recent critical contributions seek to complement this approach by proposing new theoretical derivations of the returns to education as human capital (Heckman, Lochner and Todd, 2003) and considering other dimensions of school education than human capital (Bowles, Gintis and Osborne, 2001), rather than to replace it.

In stark contrast, the literature which, from the late 1960's onwards and especially in the 1980s and 1990s, has sought to measure the impact of school inputs on educational outcomes has failed to reach a consensus on either the conceptual framework that should be used or the results of existing studies. Indeed, the 'educational production function' approach, which dominates this literature, is not based on a proper theoretical model, but on a mere metaphor, an analogy between the functioning of a school and the production process of a firm. Although dozens of papers and hundreds of estimates have accumulated over the last three or four decades, no empirical regularity comparable to estimates of the returns to education has emerged. Indeed, the same special issue of a leading journal recently included two surveys written each by a major contributor to the field and covering the same sets of studies while reaching diametrically opposite conclusions. Although a wealth of precise methodological arguments were exchanged, political considerations obviously underlay this exchange; the editors of the journal prudently avoided taking sides. In the end, after having read such surveys, whether one considers that there is a significant and positive relationship between various school resource variables and cognitive achievement in either the United States or developing countries as a whole is as much a matter of (a priori) *belief* and (informed) *opinion* as it is of (scientific) *knowledge*.

Providing a clear account of the evidence which economists have produced about the relationship between educational expenditures and outcomes thus seems an arduous task, which requires examining the relevance of the conventional economic approach to school education as a *technical* relationship between inputs and outputs as well as methodological issues in implementing the education production function framework.

Indeed, it is easy to see why one may not expect the kind of linear relationship between the resources of schools and the test scores of their students that underlies most of the empirical literature to exist. Educational expenditures are distributed across sectors (primary vs. secondary and higher education, public vs. private schools, centralized vs. decentralized school systems) and over inputs (teacher salaries and training, teaching / learning materials, buildings, management and inspection, etc.). Educational outcomes are extraordinarily diverse, and there is a great deal of arbitrariness in their definition. Because of its early emphasis on the productivity impact of the human capital acquired through schooling, economics has focused on the number of years of schooling and other measures of attainment (the 'quantity of education') and on the achievement of so-called 'cognitive skills', be they basic literacy and numeracy or a more advanced mastery of language, mathematics and science, which are relatively easy to quantify as test scores (the 'quality of education'). Other social sciences, however, have emphasized different dimensions of schooling, e.g. its impact on the psyche and the

socialization of the students, which were key preoccupations when modern school systems were established from the late eighteenth century onwards — economics, which is usually silent about this history, has by and large neglected these dimensions, although some convergence with social psychology, sociology and history may be about to take place[2].

Any relationship between these multiple items of expenditures and outcomes of schooling is defined by the interactions between pupils and teachers, which most existing studies treat as a 'black box'. Economics has no theory modeling the behavior of either pupils or teachers, and has just begun investigating key issues such as pedagogy and the contents of the curriculum. As a result, any estimation of the relationship between education expenditures and outcomes faces the need to take crucial unobservable variables into account: Information about the level and distribution of education expenditures does not provide a complete representation of a school system. As Drèze and Saran (1995), among others, have argued, school systems, besides 'resources', consist of 'incentives' and 'values'. Economics is well equipped to study monetary incentives, and recent developments in the economics of education field have gone away from the estimation of the 'education production function' to the analysis of incentives provided to teachers under various school management types, notably the impact of decentralization and privatization. The economic study of the political and cultural 'values' embedded in school systems, however, is in its infancy.

Keeping this basic conceptual discussion in mind, and noticing that more conventional — though by no means less difficult to handle — specification, data and estimation issues have plagued the field, it is hardly surprising that economics has failed to provide a general answer to a question such as 'Does money matter?' in a way that would be both general and precise enough to guide educational policies across the world[3]. What might arise, though, is a series of empirical regularities which may guide policy provided they are situated in their spatial and temporal context. 'Does money matter?' hardly makes sense, but 'Is further reducing high school class size in the 2000s in the United States likely to be a cost-effective policy to raise test scores compared to changing teachers' employment conditions?' does, provided that one also takes into account the way teachers, pupils and their parents are going to react to these policy changes. As mentioned above, though, whether such regularities have indeed emerged or not is not a settled issue.

1.2. Remarks on the existing literature

This review focuses on the mainstream economics literature. Papers published in the few leading generalist journals usually provide more methodological (and sometimes theoretical) innovations than those published in field journals, e.g. the *Economics of Education Review*, but the latter cover a wider range of issues and are often more precise about their context. The focus on mainstream economics results in a series of biases that have to be kept in mind.

First, the bulk of the developed-country literature has been produced by US-based academics and pertains to the United States. These papers reflect American political debates and as such they focus on 'hot' topics such as class size reduction or private school vouchers. Their relevance to other developed countries is not guaranteed. Second, a large part of the developing-country literature has been produced by or for international organizations and aid donors, most prominently the World Bank. The evaluation of the few specific programs for which good-quality data are available has been the focus of many of them; this leaves little space to envisage public policy from a different angle. In particular, the evaluation literature tends to consider impediments to the universalization of quality elementary education as a set of 'technical' difficulties that can be overcome by providing assistance (based on 'scientific' experimentation) to benevolent governments. The case for evaluating alternative policy proposals before generalizing them is clear enough, but this approach needs to be supplemented

[2] Indeed, it should be kept in mind that the insistence on human capital in current policy debates results from the increase in the returns to human capital generated by recent technical change and the replacement of the pacification of society and the establishment of the State's control of its citizens with the fostering of economic growth as the leading concern of policy-making circles. Nineteenth and early twentieth century educationists used to emphasize the political, social and, indeed, moral, dimensions of education, which twenty-first century economists are beginning to rediscover with partly feigned astonishment.

[3] The phrase has been used in the title of a number of publications, e.g. the book edited by Gary Burtless, *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, Washington, DC: Brookings Institution, 1996.

by a political-economic understanding of the current deficiencies of developing-country school systems[4].

A corollary of the above is that most good-quality studies pertain to a small number of countries, i.e. the United States and developing countries where important World Bank programs have been implemented (e.g. Mexico's Progresa), extensive datasets are available or teams of scholars have been able to run long-term projects (e.g. India or Kenya).

Finally, although the conceptual difficulties mentioned in the previous subsection have of course been noticed by a number of authors, and some recent studies have proposed convincing ways of reducing them, the focus on 'inputs' as the main component of a school system and test scores as its major 'outputs' remains, along with the representation of classroom processes as a black box, which makes it often difficult to articulate the quantitative results of the econom(etr)ic analysis of education with those of the other social sciences.

These pessimistic opening remarks do not imply that the economics literature on the relationship between education expenditures and outcomes has produced no results, but rather that the existing evidence is piecemeal and scattered over very different contexts, making generality and comparability uncertain.

1.3. Outline of the survey

This paper draws on available surveys that cover various parts of the relevant literature and conveniently make the reading of the very numerous older individual studies largely dispensable; it thus emphasizes recent studies that either propose significant methodological improvements or attempt to broaden the focus from inputs and test scores to a more complex and complete representation of school systems.

Section 2 reviews the small cross-country literature, which is largely inconclusive: The few resource variables that have been used do not seem to bear a strong and significant statistical relationship to the few aggregate measures of the quantity and quality of education that are available at that level of aggregation. There is much reason to suspect that specification, data and estimation issues could lead to this result even if expenditures actually had a major impact on outcomes, thus it is necessary to turn to the much more abundant, and indeed, plethoric, micro literature.

Section 3 deals with the 'education production function' literature of the 1980s and 1990s, as surveyed notably by Eric Hanushek, and the controversies raging about his surveys. There is evidence that either total expenditures per pupil or various specific inputs do have an impact on outcomes, but this impact is often small and imprecisely estimated; whether this results from conceptual and methodological deficiencies or indicates that, given current levels of expenditures and management practices, the variation in expenditures really does not explain much of the variance in outcomes in most countries remains an open question.

Section 4 reviews the newer experimental evidence, that directly addresses the impact of selected policies and programs, without necessarily making use of the same theoretical perspective as education production studies. Experiments conducted both in the United States and in developing countries as diverse as Colombia, Kenya and India have found statistically and educationally significant impacts of various types of interventions, but whether these typically context-specific interventions can be scaled up or even applied to other countries is the subject of another debate.

The next three sections consider possible improvements in the economic approach to school quality which have been recently introduced. Section V reviews work that has tried to supplement the production function framework with an analysis of responses by pupils' families to interventions that affect school quality (including the allocation of educational expenditures to families rather than

[4] More generally, the societal implications of the transition to universal elementary education need to be kept in mind, although they go much beyond the current scope of economics.

school, termed ‘demand-side financing’) and has suggested that production function estimates yield insights as to the way input levels are determined.

Section 6 discusses the current shift in emphasis away from ‘resources’ to ‘incentives’ — the policy goal, notably in developing countries, often being to improve measured outcomes without having to increase expenditures. Three related branches may be distinguished in this newer literature: Some papers examine the incentives offered to teachers under different employment conditions, others the impact of the decentralization of educational finance, yet many more have been devoted to the privatization of educational ‘provision’, and notably to the introduction of ‘education vouchers’ in the United States as well as a few developing countries (Colombia, Chile). While most of this literature does not directly address the link between expenditures and outcomes, it is relevant in so far as understanding how inputs are used is necessary to understand why they impact outcomes or not.

Section 7 covers very recent work that tries to renew the conceptual framework used by economists to represent the functioning of schools. This strand of the literature consists of just a few papers, but these, along with the methodological improvements mentioned in sections 3 and 4, may allow economics to go beyond the now sterile debates about the significance of inputs in education production. The emphasis is on behavioral economics papers that have demonstrated the importance of other dimensions of education than cognitive skills and on attempts to explicitly model the behavior of teachers and pupils in the classroom, no longer treating schools as a black box.

Section 8 concludes

2. CROSS-COUNTRY EVIDENCE

The starting point of much of the studies pertaining to the United States is a historical paradox that appears at the macro level: The large increase over the last four or five decades in average real expenditure per student (and other measures of school resources) in primary and secondary schools has not been matched by a comparable increase in average test scores. Hanushek (2003) provides a clear articulation of the argument: ‘The US, operating under a system that is largely decentralized to the 50 separate states, has pursued the conventionally advocated resource policies vigorously. […] Between 1960 and 2000, pupil-teacher ratios fell by almost 40 %. The proportion of teachers with a master’s degree or more over doubled so that a majority of all US teachers today have at least a master’s degree. Finally, median teacher experience […] almost [doubled] since its trough in 1970. […] The question remains, what was obtained for these spending increases? […] The performance [at the age of 17] of students in mathematics and reading is ever so slightly higher in 1999 than 30 years before when spending was dramatically lower. The performance of students in science is significantly lower in 1999 than it was in 1970’ (pp. 67-9)[5].

In fact, this paradox is far from being specific to the US: Using similar test scores, Gundlach, Woessmann and Gmelin (2001) find that ‘the decline of schooling productivity’ has been even sharper in other OECD countries than in the United States. At the same time, popular policy prescriptions relative to the improvement of school quality focus on further increasing expenditure and resource levels.

An econometric assessment of the existence of a positive, strong and significant causal relationship between expenditures and outcomes at the aggregate level may help solve this paradox, and give some indication as to whether these policy prescriptions are likely to be effective. This is the focus of a small number of empirical papers, which constitute a good introduction to the larger economics of education literature, its promises and pitfalls.

[5] Expenditure and input figures are taken from US Department of Education data, while scores are the results of tests administered to random sample of students aged 17 by the National Assessment of Educational Progress (NAEP). It should be noted, however, that Hanushek does not mention whether the data pertain to both public and private schools, and to which grades his expenditure and input figures apply. More generally, a problematic feature of the literature surveyed in section II and III is its lack of information about the specific contexts which the data in use are meant to represent and the corresponding tendency to oversimplify the questions examined.

This section starts with a summary of five recent contributions (2.1). A general discussion of the reliability of the aggregate evidence follows (2.2).

2.1. The aggregate relationship between educational expenditures and test scores

Hanushek and Kimko (2000) and Lee and Barro (2001) may be the most influential pieces in the aggregate literature, while Al Samarrai (2002) provides a survey of these and earlier studies as well as additional evidence. While not based on test score data, Gupta, Verhoeven and Tiongson's (1999) results suggest a different pattern for developing countries than for developed ones. Finally, Hanushek and Luque (2003) is the latest significant contribution. Appendix Table 1 summarizes the results of these studies.

Hanushek and Kimko (2000)'s main purpose is not to identify the determinants of test scores, but to estimate the impact of 'labor force quality' on economic growth. They derive a measure of labor force quality from international mathematics and science test scores, arguing that cognitive skills are a more consistent measure of the kind of human capital envisaged, for example, by endogenous growth models than the years-of-schooling variable used in the earlier empirical literature on human capital and growth. Test scores are bound to be an endogenous determinant of growth, though, as 'growth provides increased resources to a nation, and a portion of these resources may be ploughed back into human-capital investments' (p. 1191). It is as part of their strategy to show that this is in fact not the case, i.e. that their estimated impact of test scores on growth is causal, that Hanushek and Kimko run cross-country regressions of test scores on school resources and other variables. Their main conclusion is the absence of any impact of school resources on test scores: 'if stronger growth leads nations to increase investment in schools, growth could cause increased achievement. But direct estimation of international production functions does not support this as an avenue of effect' (p. 1185).

The data on educational outcomes are drawn from the tests administered by the International Association for the Evaluation of Educational Achievement (IEA) and International Assessment of Educational Progress (IAEP)[6], measuring achievement in mathematics and sciences, which is both more easily comparable across countries and perhaps more closely associated with growth than achievement in other fields, notably language. Tests were administered to pupils belonging to different cohorts in 1965, 1970, 1981 and 1985 (IEA), 1988 and 1991 (IAEP) in 11, 17, 17, 17, 6 and 19 countries, respectively. 39 countries participated at least once, with only the United States and the United Kingdom participating in all six tests.

Test scores for each country-year observation are regressed on a set of variables representing school resources and other country characteristics. Although the authors indicate that a very large number of school characteristics were considered[7], results are displayed for only three of them — the pupil / teacher ratio in primary schools, current public expenditure per student, and total expenditure on education as a fraction of GDP — which are introduced in separate regressions. The only other country characteristics are the quantity of schooling of the adult population and the annual population growth rate. These variables are usually averaged over 5- or 10-year periods around or shortly before the year each test was administered. Information about school resources and other country characteristics is drawn from a variety of sources; data limitations result in the number of observations hovering between 67 and 70, out of 87 country-cohort-test data points.

The results are summarized as follows by Hanushek and Kimko: 'The overall story is that variations in school resources do not have strong effects on test performance. The estimated effects of various measures of resources are either statistically insignificant or, more frequently, statistically significant but with an unexpected sign. This finding holds regardless of the specific measure of school resources — whether pupil-teacher ratios, recurrent expenditure per student, total expenditure per student, or a variety of other measures' (p. 1192).

---

[6] IAEP tests result from an international implementation of US NAEP tests.

[7] Namely, pupil / teacher ratios in primary and in secondary schools, pupil / school ratio, teaching materials and teacher salary in primary schools, percentage of grade repeaters in primary schools, percentage of primary-school cohort reaching the last grade, public recurrent expenditure in primary school relative to GNP, ratio of recurring nominal government expenditure on education to nominal GDP, current expenditure per pupil and ratio of total nominal government expenditure on education to nominal GDP.

Hanushek and Kimko further collapse all the test score information into two alternative aggregate measures of the 'quality of the labor force' over the 1965-91 period — which results in a sample that has either 30 or 31 observations — and regress them on country averages for the same period of the same school resource variables and country characteristics as in the previous regressions, adding a control for the primary-school enrolment rate. They conclude: 'School resources again are not strongly related to quality. The incorrect sign for pupil-teacher ratio appears whether or not there is a dummy variable for the Asian region, a region with traditionally high pupil-teacher ratios and high student performance. The expenditure measures, while positive, are statistically insignificant' (p. 1194)[8].

Lee and Barro's (2001) paper, which was widely circulated and cited several years prior to its publication, has a much more explicit focus on examining the cross-country determinants of test scores. It follows from Barro and Lee's (1996) earlier efforts to provide an international dataset on the quantity of education and the quality of schools at 5-year intervals over the 1960-90 period, originally meant to be used in growth regressions along with the so-called 'Penn World Tables' of per capita GDP. This data set includes both educational attainment by sex for the population aged 15 and above and indicators of the quality of school inputs (teacher / pupil ratio, spending per pupil, teacher salaries, the fraction of students repeating grades or dropping out of school, each available for both primary and secondary schooling but for teacher salaries and the drop out rate, available only for the primary level), and has full data for 105 countries and incomplete data for 21 more. Lee and Barro supplement this information by using the same IEA and IAEP data as Hanushek and Kimko, though the use of reading test scores and the 1993-98 mathematics and science tests raises their sample size from 39 to 58 countries, and 214 country-year observations. Besides the determinants of test scores, they also consider those of school repetition and dropout rates, as additional proxies for the quality of schooling.

Lee and Barro first regress test scores on pupil-teacher ratios, teacher salaries, expenditures per pupil and the length of the school year, controlling for per capita GDP and primary education of adults. Each specification is simultaneously estimated as a system for the different subjects (mathematics, science and reading), age groups (10- and 14-year-olds) and test years, by seemingly-unrelated-regression (SUR) techniques.

In stark contrast with Hanushek and Kimko, Lee and Barro find that school resources have a significant relationship to test scores: 'The pupil-teacher ratio has a negative relation with test scores, confirming that smaller classes are associated with better pupil achievement. The estimated coefficient, – 0.15, $t = 2.44$, implies that a one-standard-deviation decrease in the pupil-teacher ratio (by 12.3 in 1990) raises test scores by 1.8 percentage points. The log of the average salary of primary school teachers has a positive, though less significant, relation with test scores. The estimated coefficient, 1.62, $t = 1.81$, indicates that a one-standard-deviation increase in the log of average salary of primary school teachers (by 0.9 in 1990) is estimated to raise test scores by 1.4 percentage points' (p. 479). The length of the school term and total educational spending per student, however, turn out to be insignificant. The inclusion of an East Asia dummy does not alter these results, suggesting that they are not driven by East Asian countries that score high on international tests — meanwhile, the large, significant coefficient associated with this dummy implies that the East Asian experience is not adequately explained by the variables used in the analysis. Results obtained for age-specific samples (10- and 14-year-olds) are comparable. Results obtained for subject-specific samples however suggest that the determinants of reading differ from those on mathematics and science. The pupil-teacher ratio has a larger negative impact on mathematics and science, teacher salary a larger positive impact on reading, while the length of school days has a strong positive impact on mathematics and science and a large negative one on reading (this is the only case in which individual coefficients are statistically different, as opposed to the vectors of all coefficients).

Finally, results for repetition and drop out rates are also consistent with a significant impact of school resources, as the pupil-teacher ratio has a positive and significant impact on these variables. In two specifications out of four, teacher salaries exert a strong, negative impact. Lee and Barro conclude that

[8] Hanushek and Kimko nevertheless use these regressions to predict the quality of the labor force for countries for which growth rates and other determinants of growth are available, but not test scores, and use the predicted value in their growth regressions.

'school inputs (especially smaller class sizes, but probably also higher teacher salaries and greater school length) enhance educational outcomes' (p. 485).

Al Samarrai's (2002) survey of the papers by Hanushek and Kimko (2000) and Lee and Barro (2001) along with four earlier studies (Colclough with Lewin, 1993; McMahon, 1999; Schultz, 1995; Woessmann, 2000), provides a broader view of the evidence, though no definitive conclusion. Al Samarrai's reading of the evidence is that there is 'no consistent effect of resources on educational outcomes. Studies using internationally comparable test scores tend to show that resources have a significant impact, but the direction of this impact differs across the studies' (p. 3). Indeed, some studies present results similar to those of Hanushek and Kimko, others results similar to those of Lee and Barro.

Al Samarrai's own evidence is based on cross-section regressions that use 1996 UNESCO data on the quantity of education (primary-school gross and net enrolment rates) and proxy measures of its quality (survival rate to grade five and primary school completion rate). Sample size is respectively 90, 79, 69, 33, i.e. not quite larger than in papers based on test score data, but the sample is said to include more developing countries — Al Samarrai somewhat surprisingly provides no list of sample countries. Explanatory variables include public primary education expenditure as a fraction of GNP, primary expenditure per pupil, the pupil / teacher ratio in primary schools and public spending on primary education, as well as controls for various country characteristics (Gini coefficient, per capita GNP, urbanization rate, proportion of Muslim population, regional dummies). OLS results generally show insignificant or 'wrongly' signed results, which are confirmed using alternative specifications and estimation techniques (including an instrumentation of school resources variables using the secondary school pupil teacher ratio, total education spending as a proportion of GNP and the length in years of the primary cycle, which is not entirely convincing). Al Samarrai's discussion of these results emphasizes issues of data quality and variable omission bias — he cannot control for household expenditure on education, the effectiveness of school management or the composition of public expenditure on education.

Gupta, Verhoeven and Tiongson (1999) examine the impact of government spending on education and health care; the specificity of their paper is its use of a cross section of 50 developing and transition countries, while other studies typically include more developed than developing countries. They do not use test scores, however, but focus on enrolment rates and the retention of pupils in school until grade four, which can be interpreted as a proxy for achievement.[9] They find that, total education spending hardly affects retention, but that the proportion of primary and secondary education spending in total spending has a significant and positive impact. Not surprisingly, therefore, it is the share of resources devoted to elementary schools that would affect their functioning; total resources are too rough a measure to be relevant.

Hanushek and Luque (2003) is the most recent significant contribution to this literature. Building on Hanushek and Kimko (2000), they use data drawn from the latest IEA survey, the Third International Math and Science Survey (TIMSS), collected in 1995 in more than 40 countries, for pupils aged 9, 13 and 17 years. For each age group, a representative sample of 150 schools was designed in each country, and the tests were administered to two classes in each school.[10] Hanushek and Luque examine the determinants of achievement focusing on the class-level averages of mathematics test results for pupils aged 9 and 13 years, yielding two samples of 300 observations for each country. For each age group, they run separate country regressions of the test scores on various school and family characteristics, but base their discussion of estimation results on the proportion across countries of estimates that are significant vs. non-significant, negative vs. positive. Thus, although the individual regressions are run at a fairly disaggregated level within each country, their paper belongs in spirit to

[9] Whether at the macro or at the micro level, the literature on school participation (enrolment and attainment variables) is distinct from the literature on educational outcomes. The latter focuses on the functioning of schools as represented by the education production function and is covered in this survey; the former is not, for it focuses on the human capital investments (of which school quality is only one determinant, typically represented by a few poor proxies) made by households and thus belongs to another field of economics, namely, family and population economics.

[10] Hanushek and Luque mention that implementation issues raise doubts as to whether the results for some countries are representative.

the macro strand of the literature (regressions are run for 12 to 17 countries for children aged 9, and 25 to 33 countries for children aged 13).

Their main set of results is based on OLS regressions including actual class size, dummies for teachers having at least a bachelor's degree and teachers with special training, teacher experience, and the total school enrollment as school characteristics, while family characteristics include the proportion of students whose parents have not completed secondary education and the proportion of families owning various goods that both reflect wealth and may have a directly impact on learning (more than 25 books, calculator, computer, study desk, dictionary). As the table summarizing the results indicates (see Appendix Table 2), almost all of the coefficients associated with school characteristics are insignificant, and split almost equally between negative and positive ones. Only 3 and 2 countries for pupils aged 9 and 13, respectively, exhibit the 'expected' negative and significant relationship between achievement and class size; none exhibits a positive and significant relationship between achievement and teacher education or experience.

Hanushek and Luque conclude: 'Across the sampled TIMSS countries, the overall strength of resources in obtaining better student performance appears rather limited […]. Certain countries […] do stand out as having significant effects, and these should be investigated in more detail. Nonetheless, these results defy many generalizations. It simply does not appear to be the case that outcomes related to school resource differences are more positive in the poorer countries or in the countries that begin with lower levels of resources' (p. 497).

2.2. How reliable is the evidence?

Taken as a whole, the papers summarized above do not suggest that a positive, strong and significant relationship between educational expenditures and outcomes can be identified using aggregate data. There is much reason to believe that this is due to data, specification and estimation issues, and that it is would be difficult in any case to assess the 'real' relationship.

Indeed, the macro literature is quite marginal to the field of education production function studies, for three reasons. First, there exist very few data sets providing sufficient information on both outcomes and relevant measures of expenditures or school quality while covering enough countries for econometric analysis to be meaningful. IEA and IAEP data used among others by Hanushek and Kimko (2000), Lee and Barro (2001) and Hanushek and Luque (2003) yield samples of a few dozen countries, that have no reason to be representative, and in which developing countries are underrepresented[11]. The fact that very few countries participated in several tests makes it impossible to apply panel regression techniques that would help solve technical issues such as the impact of unobservable country characteristics, and to adequately analyze the time dimension of the data.

Second, no paper presents a rigorous derivation of a macro equation based on the aggregation of the education production function that constitutes the core of the corresponding microeconomic literature. There is thus no clear guidance as to which variables should be included as determinants of outcomes, how they should enter the model (most papers assume a linear specification), and which countries should be included for the sample to be reasonably homogenous. Brock and Durlauf's (2001) critical appraisal of the empirical growth literature, which emphasizes the issues of 'model uncertainty' and 'heterogeneity uncertainty', is thus highly relevant here, as it concerns small-sample econometrics in general — even arranged in a panel, country-level data cannot have more than a few dozens, or, at the very best, two or three hundreds observations.[12] The available evidence is thus at best descriptive, i.e. it provides correlations rather than causal estimates, or policy-relevant parameters. Even so, it does not yield a consistent description: Some studies do report statistically significant correlations between

[11] For example, the sample of 40 countries for which Hanushek and Luque (2003) present TIMSS mathematics scores for students aged 13 includes 18 Western or Southern European countries, 9 Central or Eastern European transition countries, 4 other 'Western' developed countries, 4 East Asian developed countries and only 5 developing countries (Colombia, Iran, Kuwait, South Africa and Thailand).

[12] Papers examining the aggregate relationship between educational expenditures and outcomes could be envisaged as a subset of the growth literature initiated by Robert Barro's work on conditional convergence in the early 1990's: Barro himself is one of the main contributors, the same data sets and empirical methods are used, and the relationship between educational expenditures and outcomes is sometimes but the first stage of an instrumental variables estimation of the relationship between the quality of education and growth, as in Hanushek and Kimko (2000).

various input measures and test scores, but there are too few of them for these results to constitute an uncontroversial body of evidence[13].

Third, the prospects for improving on existing papers are rather bleak. Test score data are becoming increasingly available for developing countries, but using only country-level aggregates will remain a way to lose most of the information they contain. Meanwhile, expenditures are likely to be endogenous, for unobserved country characteristics could well determine them jointly with outcomes: Even if a consistent relationship were found between expenditures and outcomes, determining how much of it represented causation rather than mere correlation would require the development of a credible identification strategy. Such strategies typically rely either on explicit theoretical modeling or on the existence of some clearly exogenous source of variation in the explanatory variable. For the moment, there is no rigorous derivation of a macro equivalent to the already controversial micro 'education production function' that would suggest good instruments, and no guarantee that such instruments could be found in available datasets. Meanwhile, although the technique has been successfully used *within countries*[14], it seems difficult to imagine how a 'natural experiment' could generate exogenous variations in educational expenditures *across countries*.

In the end, it is unclear whether this literature provides any reliable information as to the determinants of educational outcomes that could be used to guide policy. As Al Samarrai (2002) himself concludes: 'Given the absence of a clear, strong relationship, the use of cross-country averages to guide individual country education policy in resourcing decisions is unlikely to be meaningful. The results suggest, for example, that to use average levels of education spending in countries that have achieved schooling for all as targets for less successful countries is not useful, *and is almost certainly no substitute for detailed country level analysis*' (p. 19; emphasis added). Indeed, one may wonder whether the available information could make up for a proper historical understanding of the development of the school systems of the countries under study, one that would credibly explain the paradox on which this section opened, for example. In developed countries, the broad economic and social change took place over the last decades cannot but have had affected the functioning of schools and thus the correlation between expenditures and outcomes[15]. In developing countries, the issue is even more complex, for there has been a large increase in enrolment rates at the primary-education level, possibly giving rise to quantity-quality trade-offs which cannot be analyzed using cross-sectional data and are best studied *within* countries rather than across them[16].

To conclude this section, two approaches that have been used in the returns to education literature may yield more convincing macro evidence if applied to the relationship between educational expenditures and outcomes. First, one could use data aggregated at a lower level, e.g. cities, districts, regions or states belonging to the same country or set of neighboring countries, reducing sample heterogeneity. Such data, when available, are bound to be more reliable and to have more observations than cross-country panels. Second, a general assessment of the plethoric micro literature might be expected to yield a consistent set of stylized facts. Attempts to estimate macroeconomic versions of the Mincerian human capital earnings function have not been quite successful, and the empirical growth literature has found it difficult to estimate the impact of human capital flows or stocks on economic growth

[13] Meanwhile, some interpretations of the results are far-fetched, e.g. Lee and Barro's interpretation of the strong coefficient attached to their East Asian dummy: 'The significance of the East Asian dummy may reflect the existence of an "Asian value", which is broadly defined by the cultural and religious features unique to the East Asian countries' (p. 481).

[14] See subsection 4.2 below.

[15] Hanushek (2003) mentions two common explanations of this paradox for the United States. First, the characteristics of students would have changed, i.e. family backgrounds would have become less conducive to learning. Second, the spread of 'special education' programs for disabled students would have increased expenditures but not outcomes as these students are not included in NAEP sample. He counter-argues that the trends in family backgrounds are ambivalent (e.g., the proportion of single-parent families has increased, but so has the quantity of parental education) and 'special education' programs represent a minor fraction of the increase in expenditures. His preferred explanation is that increasing expenditures cannot improve test scores if the incentives structure of the school system is not altered, an issue that is examined in sections 6 and 7 below. Whether one agrees with Hanushek or not, it is easy to see that an econometric examination of these competing hypotheses would require much richer information than is available at the cross-country level; at the very least, the information should be disaggregated at the US state level.

[16] Al Samarrai's (2002) result that primary expenditure per pupil is *negatively* related to enrolment rates is bound to reflect such a quantity-quality tradeoff: Countries in which enrolment is high or has quickly increased may happen to spend less per pupil as a result, but this cannot be taken to mean that increasing spending would *cause* a reduction in enrolment rates.

(Krueger and Lindahl, 2001; Pritchett, 2001)[17], yet economists studying the returns to education can draw macroeconomic conclusions from a consistent body of microeconometric evidence. Unfortunately no such consensus exists in the education production function literature; this is the topic addressed in section 3.

3. THE EDUCATION PRODUCTION FUNCTION LITERATURE

The modern economic approach to the determinants of educational outcomes, namely, the education production function literature, started in the United States in the late 1960's and 1970's, in the aftermath of the 'Coleman report' (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld and York, 1966) that had concluded that, in the United States, family background and the composition of peer groups in school had a much larger impact on educational outcomes and economic success than variations in the characteristics of the schools themselves. Another influential report conducted a decade later in the United Kingdom (Rutter, Maughan, Mortimore and Ouston, 1979) 'showed that schools have a modest but significant impact on the learning outcomes of children, though again family circumstances, socio-economic background and individual ability matter far more than schooling' (Vignoles et al., 2000, p. 2). It should be kept in mind, however, that the debate about the relative importance of schools and families in determining cognitive achievement and subsequent social and economic success is not quite recent. Exactly a century ago, French sociologist Paul Lapie drew similar conclusions from his study of life histories of 722 men who had been enrolled between 1872 and 1893 at the primary school of Ay, a small town in Eastern France: 'Schooling sometimes succeeds in breaking the strings of the net by which economic circumstances control our destinies. Its impact is not great, but it is not nothing'[18].

Much of the modern literature inaugurated by the Coleman report has focused on the United States and, typically, on 'class size', i.e. the pupil-teacher ratio measured at various levels of aggregation, but the growing availability and increasing quality of household surveys in developing countries from the 1980s onwards has allowed the publication of a significant development literature. Indeed, it has been widely argued, following Heyneman and Loxley (1983), that schools have a larger impact than families in developing relative to developed countries.

The explicit aim of this literature has been to guide educational policies across the world by identifying which policy-controlled inputs have the largest marginal impact on achievement, e.g. reducing class size vs. providing more textbooks, increasing teacher salaries or improving teacher training. Indeed, in the absence of proper theories of cognitive processes, pupil-teacher interactions or schools as institutions, the general assumption is that increased educational expenditures and higher levels of input provision will lead to 'better' educational outcomes.

This section draws extensively on three papers by Hanushek, Krueger, and Todd and Wolpin, published in a special issue of the *Economic Journal* (113 [485], February 2003) which together form a valuable and up-to-date introduction to the literature and its contradictions — the derogatory remarks made at the beginning of subsection 1.1 notwithstanding. The most readable recent surveys of the literature, however, are Vignoles, Levacic, Walker, Machin and Reynolds (2000), who focus on US and UK studies, and Glewwe (2002), who emphasizes the need for policy-relevant analyses of developing countries. References to important studies not explicitly mentioned in this text can be found in the extensive reference lists of these five papers.

This section proceeds as follows. Some important methodological issues are introduced from the outset (3.1), before the evidence is presented first as it is summarized in the controversial and mutually

[17] Hanushek and Kimko's (2000) results, which suggest a strong causal impact of the quality of education measured through test scores on and growth, are an exception.

[18] My translation from the French. The original text is: '*Ainsi, l'école réussit parfois à rompre les mailles du réseau dans lequel des causes d'ordre économique enferment nos destinées. Son action n'est pas considérable, mais elle n'est pas nulle*'. Paul Lapie, 'Les Effets sociaux de l'école' ['The Social Effects of Schooling'], *La Revue scientifique (Revue rose)*, 41 (2), 1904, quoted in Baudelot and Leclercq (forthcoming).

inconsistent meta-analyses published over the last two decades (3.2), and then through a closer examination of a few influential studies (3.3)[19].

3.1. The econometric analysis of educational production: methodology issues

As the term suggests, education production function studies are based on the idea that school education can be represented as a production process whereby a vector of inputs, comprising pupils' innate ability' and initial achievement level and various characteristics of the schools they attend (usually teachers' education, training, experience and salaries, the pupil / teacher ratio and the availability of school buildings and teaching / learning materials) is transformed into a vector of outputs, usually cognitive achievement, most frequently measured by test scores in language, mathematics and science[20]. The cognitive process itself as well as the interactions between pupils and teachers are considered as a 'black box': Only inputs and outputs are known, while educational processes are not of direct interest and neither pupil nor teacher behavior is modeled. Furthermore, the focus really is on education as a technical process taking place *within the school*, and interactions between schools and families, or the management of the school system, which largely determine the inputs available to each child and teacher, are not modeled either, but for the occasional inclusion of a few family background control variables. The analogy between education and the production process of a firm, however, guides the choice of variables to be included among the determinants of outcomes and enables a consistent interpretation of the results. Specifying and estimating an education production function raises methodological issues that have long been recognized (see Hanushek, 1979, for an early discussion) and are discussed in this subsection.

Vignoles et al. (2000) identify three main deficiencies of the existing literature: The lack of an established theory, the absence of satisfactory solutions to estimation issues, and the poor quality of many of the available datasets[21]. These deficiencies are closely linked to each other; resulting issues may be discussed here.

While achievement is cumulative, and depends not only on contemporaneous inputs, but also on past ones, the latter are typically not observed. A solution, which requires the collection of test score data both at the beginning and at the end of the period over which inputs are observed, consists in estimating 'valued-added' models, in which either the change in test scores rather than their final level is the dependant variable, or initial test scores are included among the explanatory variables. Vignoles et al. (2000) emphasize that 'an individual's initial ability and attainment […] are obviously important as they allow for the fact that some children are intrinsically more able than others, some have already had better schooling than their peers, and some have experienced greater parental inputs before starting at a particular school. […] the inclusion of these variables effectively "levels the playing field" at time of school entry. Clearly the impact of the child's subsequent schooling must be measured separately from these other factors' (p. 5).

A second difficulty is that the explanatory variables, school resources, are likely to be 'endogenous'. Endogeneity — a recurrent issue central to contemporary empirical economics— arises when

---

[19] Estimating education production functions is not the only approach to these issues; Vignoles et al. (2000) mention 'the parallel school effectiveness research field', surveyed in Teddlie and Reynolds (2000; see Unnever, 2001, for a brief review of the book), which is not covered here any more than it is in their paper, as 'it is [the educational production function literature] that has been most concerned with the impact of resources on outcomes' (p. 1) and debates in mainstream economics usually do not extend to school effectiveness research. Yet another approach consists in the parametric ('stochastic frontier regression') or non-parametric ('data envelopment analysis') estimation of 'education production frontier models'. While the estimation of education production functions aims to identify the parameters of a structural relationship of general relevance, education production frontier models aim to identify the efficiency frontier of a specific sample, and the best practices of the schools that are located on this frontier. Only the production function is covered in this review; see Vignoles et al. (2000), section 1.2 pp. 12-16 for a discussion of production frontier models.

[20] As noted by Hanushek (2003), the focus of the returns to education literature remains the quantity of education (grade attainment), while the focus of the education production function literature is the quality of education (cognitive achievement). This contrast is quite surprising, given that school (and family) characteristics determine both attainment and achievement, and that both the quantity and quality of education determine adult outcomes. Note that Glewwe (2002) includes a survey of studies on the returns to the quality of education (i.e. studies which, rather unusually, include test scores or other measures of skills among the determinants of individual earnings).

[21] As remarked by Hanushek (2003) about the US literature, 'analysis seldom has relied on data collected specifically for the study of the educational process. Instead, it has tended to be opportunistic, employing available data to gain insights into school operations' (p. 74). The increasing tendency in empirical microeconomics to rely on purpose-collected data when large standard datasets do not provide enough information should contribute to solve this issue.

unobserved variables explain both the dependent and the explanatory variable, leading to a confusion between the true causal impact of the latter on the former and the mere correlation between them due to their common determination by the unobserved variables. Endogeneity can be overcome if a set of variables can be observed which are uncorrelated with the unobserved variables, and determine the dependent variable only through their impact on the explanatory variable. The use of such 'instrumental variables' (IV) is becoming standard practice in empirical studies, and finds a natural application in the potential endogeneity of school resources as determinants of educational outcomes.

Indeed, two sources of endogeneity that have long been recognized in the policy evaluation literature need to be taken into account, namely, 'selective migration' and 'purposive program placement'[22]. 'Selective migration' is an issue when parents can choose the school their children attend on the basis of its resources, e.g. by moving close to what they perceive to be a 'good school' or by sidestepping legal school mapping by having them choose optional courses that are not available in the 'bad' school of their neighborhood — not to mention enrolling them at a private school. Such behavior is very likely in contexts where parents are very mobile or can freely exert school choice[23]. Schools with more resources end up having pupils with more favorable family backgrounds, which leads to an overestimation of the impact of resources on outcomes. 'Purposive program placement' is an issue when educational policies target extra resources to deprived areas or weaker students. Schools with more resources may then appear to have weaker outcomes, thus leading to an underestimation of the true causal parameter[24]. These two phenomena are not mutually exclusive, for example a school improvement program targeted to a deprived area may well end up benefiting mostly its more affluent residents.

The endogeneity problem may be solved through structural modeling of educational expenditures yielding complex simultaneous equations systems that include both an education production function and expenditure allocation rules, but this is a difficult exercise. Most recent papers rely on a simpler use of 'instrumental variables' (IV) that can be thought of as determining school resources but being uncorrelated with family characteristics and having no direct impact on test scores[25]. IV estimation has two major limitations. First, it is often difficult to identify instrumental variables and credibly argue that they are uncorrelated with unobservable determinants of expenditures and outcomes. Second, 'a further problem with using IV models is that they can only identify the effect of a change in school inputs for a particular sub-set of the pupil population where there has been a variation in school inputs […] If the characteristics of the target population […] are not representative of the total population, results may be biased' (Vignoles et al., 2000, p. 7).

Finally, the lack of a proper theory and the inadequacy of the data available often result in model misspecification. Omission biases are pervasive, for variables are included or excluded according to data availability rather than scientific considerations — very few datasets include rich information on both school and peer, family and community characteristics. A linear functional form is usually assumed without other a priori reasons than convenience, although evidence of non-linearities was found using translog functional forms and quantile regression techniques. Aggregation bias is another pervasive problem, for resource variables, in particular, are often available only at the level of administrative units that include large numbers of schools (even states in many papers using American data) and thus do not reflect the actual resources experienced by pupils. The use of aggregate data

---

22 See Strauss and Thomas (1995) for a discussion in the context of education and health policy interventions in developing countries.

23 Choosing to reside in a neighborhood with a 'good' school for one's children to attend is a common practice in the urban areas of many countries. School-quality induced differentials in housing demand across neighborhoods tend to translate into property prices differentials. Using data covering the housing markets of suburbs of Boston in the mid-1990's, Black (1999) shows that it is possible to estimate parental valuations of school quality in the US by comparing the prices of houses located on different sides of the same streets that constitute the boundaries of school attendance districts, but are otherwise identical once observable characteristics are controlled for. She finds that, on average, parents are 'willing' to pay 2.5 % more to enroll their children in schools whose average test scores are 5 % higher.

24 Such compensatory policies are very common and are often at the core of political debates about education, e.g. France's 'Priority Education Zones' ('*Zones d'Éducation Prioritaire*'). In the United Kingdom, 'the system of educational funding […] includes such a compensatory measure. In particular, the national funding system allocates resources to local education authorities on the basis of pupil numbers in various age bands, weighted by factors to reflect social need […] and also, in some cases, higher cost indices. [Local Education Authorities'] own funding formulae then pass these funds to schools in ways which reflect need in varying degrees. Schools themselves are then free to allocate resources to students, for example by differential teaching group sizes' (p. 5).

25 Such IVs are typically derived from specific features of education policies and follow the paradigm of experimental econometrics, they are thus treated in section IV below.

would increase omission biases and lead researchers to attribute to school resources the impact of omitted variables, e.g. community characteristics at the same level of aggregation.

Given the lack of theoretical guidance to estimation, a careful consideration of available data seems the key to obtaining convincing results. Providing a consistent statistical framework for such an exercise is the aim of Todd and Wolpin's (2003) paper, which 'focuses on the problem of how to specify and estimate a production function for cognitive achievement in a way that is consistent with theoretical notions that child development is a cumulative process depending on the history of inputs applied by families and schools as well as on children's inherited endowments' (p. 5). They argue that commonly used specifications imply highly restrictive assumptions about the underlying 'technology', and derive from a common econometric model a series of estimators adapted each to a different kind of data limitation, discussing identifying assumptions, their plausibility and conditions under which they can be tested as well as data requirements. Three alternative families of estimators are proposed, based on 'contemporaneous', 'value-added' or 'cumulative' specifications, which in this order have increasingly heavy data requirements (from information on contemporaneous inputs and outcomes only to full information on the history of both) but decreasingly restrictive assumptions.

The above discussion of methodological issues should have convinced the reader that the reliability of education production function studies is extremely variable, depending on the specification, the quality of the data and the rigor in statistical and econometric analysis. This is bound to make systematic comparisons between studies rather difficult, as Todd and Wolpin (2003) emphasize: 'Accounting for the variety of estimating equations adopted in the […] literature, it is easy to see how studies, even those based on identical datasets, draw different conclusions' (p. 31). Better data collection and more careful attention to methodology might yield clearer results in the near future; for the time being, Todd and Wolpin's discouraging conclusion needs to be kept in mind while examining the heated controversies that surround attempts to assess the existing 'evidence' using quantitative meta-analytic techniques.

3.2. Meta-analyses of education production function studies

The education production function literature now includes a few hundreds of studies, so that a meta-analysis of the evidence may be expected to provide robust empirical regularities as to which inputs stimulate achievement, and to what extent — provided that one accepts the implicit assumption that there exists a homogenous 'technology' whose parameters would be common to all education systems across the world, conditional on a sufficient number of adequately selected control variables.

The key reference in this literature is a series of surveys published by Eric Hanushek (1986, 1995, 1996 a, b, 1997, 1998, 2002 a, b, 2003), which have been extremely influential and controversial at the same time[26]. According to Hanushek indeed, 'the central conclusion is that commonly used inputs policies — such as lowering class sizes or tightening the requirement for teaching credentials — are almost certainly inferior to altered incentives within the schools' (p. 64). Increasing resources available to schools would not have a major impact on outcomes: 'Quality is […] virtually impossible to dictate through policy. The quest for improved quality has undoubtedly contributed to recent expansions in the resources devoted to schools in the US and other countries. Eager to improve quality and unable to do it directly, government policy typically moves to what is thought to be the next best thing — providing added resources to schools. Broad evidence from the experience in the US and the rest of the world suggests that this is an ineffective way to improve quality' (p. 66).

The technique used by Hanushek to summarize hundreds of published estimates is 'vote counting'. Vote counting consists in exhaustively inventorying studies which satisfy adequate quality requirements, and computing the proportion of estimates of the impact of each input that are positive vs. negative, or statistically significant vs. insignificant. Hanushek (2003) presents two series of results pertaining to the United States on the one hand, and developing countries on the other.

[26] One may say that Hanushek's reviews fulfill the same scientific and journalistic functions in the education production function literature as George Psacharopoulos' recurrent compilation of estimates does in the returns-to-education literature.

The US sample 'includes all published analyses prior to 1995 that include one of the resource measures described below, that have some measure of family inputs in addition to schools, and that provide the sign and statistical significance of the resource relationship with a measurable student outcome. The 89 individual publications that appeared before 1995 and that form the basis for this analysis contain 376 separate production function estimates' (p. 75). Educational outcomes considered are test scores in three quarters of the estimates, and other outcomes such as continuation in school or dropout behavior in the rest of the estimates. Resources considered are real classroom resources (teacher-pupil ratio, teacher education, teacher experience), financial aggregates (teacher salary, expenditure per pupil), other resources (facilities, administration, teacher test scores).

The display of the results (Appendix Table 3) distinguishes between statistically insignificant and significant estimates, and among the latter, between positive and negative ones[27]. The number of estimates by school resource variables varies from 41 to 276. 53 to 86 % of the estimates are insignificant. 5 to 14 % are negative and significant and only 9 to 37 % are significant and positive as expected, although they outnumber negative estimates for all the variables. Hanushek then presents alternative tabulations for the teacher-pupil ratio and expenditure per pupil in which studies are classified according to quality criteria. He first deals with aggregation bias by distinguishing between studies based on single- and multiple-state samples and, among the latter, between studies which use state-level data and studies which use more finely disaggregated data, and thus rely on within-state as well as between-state variance. Estimates based on between-state variance only are likely to be biased due to the omission of state characteristics which are correlated with resource policies that are chiefly determined at the state level; estimates which rely on within-state variation can be more confidently expected to correspond to the impact of resources themselves. It turns out that the proportion of positive and significant estimates is higher in multiple- than in single-state samples and among the former, in samples based on state-level data. This is suggestive of aggregation bias. Hanushek then presents a tabulation restricted to value-added models, which should be more reliable than contemporaneous specifications. Very few of these estimates are significant and positive, but for those pertaining to teacher experience (37 %, up to 41 % for studies using single-state samples).

The developing-country sample (Appendix Table 4) includes 96 estimates pertaining to the following variables: teacher-pupil ratio, teacher education, experience and salary, expenditure per pupil, school facilities. There are 12 to 63 estimates per variable, and the results are much more varied than those of the US sample. First, there are less insignificant estimates, the proportion varying between 26 and 61 %. Second, among significant estimates, a majority are positive for all variables, and the proportion of negative estimates is below 10 % but for the teacher-pupil ratio and teacher salary. Third, an absolute majority of the estimates are significant and positive for teacher education, expenditure per pupil and school facilities. Hanushek acknowledges that 'a minority of the available estimates suggests much confidence that the identified resources positively influence student performance' and 'there is generally somewhat stronger support for these resource policies than that existing in US analyses', but still concludes that 'the evidence is not conclusive that pure resource policies can be expected to have significant effect on student outcomes' (p. 84).

Hanushek's reading of the evidence has been repeatedly challenged, notably by Alan Krueger[28]. Krueger's latest contribution (2003) criticizes Hanushek's methodology, focusing on the 1997 paper. His main point is that counting estimates, as opposed to studies, to compute the proportions of positive, negative and insignificant results gives more weight to studies from which larger numbers of estimates are drawn. Meanwhile, both publication procedures and Hanushek's selection rules result in a negative correlation between the number of estimates drawn from a study and the proportion of them that is positive and significant. Researchers who obtain 'unexpected' negative or insignificant impacts of resources on test scores are more likely to be required by journal referees to provide alternative estimates to establish the robustness of their results than those who obtain 'expected' positive and significant coefficients. Further, Hanushek would have a tendency to over-sample estimates which

[27] The actual regression coefficient may be of the opposite sign, e.g. a negative coefficient associated with the pupil-teacher ratio is considered a positive estimate for it implies that teacher resources are positively associated with achievement.

[28] Indeed, the publication in the same book or journal issue of two contradictory papers by Hanushek and Krueger has become something of a ritual (1998, 2002, 2003). For a very similar discussion in the developing-country context, see Hanushek, 1995 and Kremer, 1995.

corroborate his views: '[…] the selection and classification of estimates in several of the studies is open to question, and could in part account for the curious relationship between the numbers of estimates taken from a study and the study's findings' (p. 44).

Hanushek's (1997) survey covers 59 studies of the impact of class size on test scores in the United States, from which a total of 277 estimates were drawn. As a matter of fact, the number of estimates drawn from each study varies between 1 and 24. 9 studies represent 15.3 % of the studies but 44.4 % of the estimates, while 17 studies providing only one estimate represent 28.8 % of the studies but 6.1 % of the estimates. It is easy to see that, if studies from which small and large numbers of estimates were drawn differ in a systematic way, the results of vote counting will be affected by the choice of the unit of analysis. Krueger provides some evidence that there is indeed a negative correlation between the number of estimates per study and the proportion of positive estimates. He then compares Hanushek's tabulation with three alternative tabulations which give equal weight to all studies ('one study, one vote' rather than 'one estimate, one vote') with rough adjustments for study quality (Appendix Table 5). While Hanushek's tabulation yielded a proportion of 14.8 % (13.4 %) positive (negative) and significant results, Krueger's yield 25.5. % (10.3 %), 34.5 % (6.9 %) and 33.5 % (8.0 %). Positive estimates thus are in a large majority, although the proportion of insignificant estimates remains high.

A difficulty with vote counting is that there is no benchmark against which to assess the results: Above which threshold is the proportion of positive and significant estimates too high to be plausible if there were actually no effects of inputs on outcomes? Krueger provides such a benchmark, by computing the ratio of positive to negative estimates (whether significant or not) and displaying the probability that this or a higher ratio would be obtained in 59 independent random draws (corresponding to the number of sample studies) in which positive and negative results were equally likely. The proportion is 0.500 using Hanushek's tabulation, but 0.059, 0.034 and 0.009 using Krueger's. Krueger concludes that 'studies with positive studies are twice as likely as studies with negative results; the probability of observing at this many studies with positive results by chance is less than one in a hundred' (p. 44).

There is thus no consensus as to the results of existing meta-analyses. Authors who have used more formal techniques than both Hanushek and Krueger agree with the latter that the statistical probability of finding significant proportions of positive estimates if resources had no impact on achievement is extremely low.[29] However, as argued by Hanushek (2003), the use of these methods to compare results based on different specifications, data and estimation techniques is problematic.

Following Vignoles et al. (2000), one may wonder whether meta-analysis is at all a relevant approach to the education production function literature, for three reasons. First, it suffers from publication bias, as published studies are a non-random sample of existing or potential studies. Controlling for this type of sample selection bias unfortunately appears impossible. While Krueger argues that an author reporting unexpectedly insignificant or negative impacts of inputs on outcomes will be required to provide more alternative estimates for his study to be published, a bias in the opposite direction is probably more pervasive: Studies reporting expectedly positive impacts are more likely to be published. Second, meta-analysis requires reasonably large samples, hence a tendency to include the more numerous low-quality studies along with the rare high-quality ones in the sample. However, 'the statistical aggregation of work that is of a low methodological quality is likely to be uninformative' (Vignoles et al., 2000, p. 20). Third, Hanushek's samples, for example, cover a long time period during which the parameters of the education production function have changed.

This raises a deeper conceptual issue than discussed by Vignoles et al. (2000), which is the relevance of mimicking a structural approach to econometric analysis when studying the determinants of educational outcomes in the absence of a proper theoretical framework explaining pupil and teacher behavior. The parameters of the education production function are not clearly defined theoretically,

[29] See Hedges, Laine and Greenwald (1994) and Hanushek's answer (1994), as well as Greenwald, Hedges and Laine (1996). Also see Dewey, Husted and Kenny (2000) for an exploration of the sensitivity of vote-counting to the selectivity of the sample according to specification-related quality criteria.

hence pretending to identify them is somewhat perilous. Given that usual specifications of education production functions exclude many potentially important unobservables (notably family and community characteristics as well as the institutional features of each specific school system representing its 'incentives' and 'values'), they can hardly be interpreted as yielding estimates of a technical relationship that would be of universal relevance. As Todd and Wolpin (2003) argue, 'a leading candidate for explaining why studies reach such different conclusions is that the statistical models used to estimate these relationships are misspecified and fail to account for the major determinants of achievement' (p. 5).

3.3. Selected examples of recent education production function evidence

In the end, it may make more sense to concentrate on a subset of high-quality studies of obvious policy relevance, without failing to mention the specific context in which the impact of a given input on a given outcome has been identified. Building on a limited number of well-defined stylized facts to answer complex, context-specific questions is likely to create less confusion than over-simplifying issues while mixing reliable and unreliable evidence, provided that one avoids drawing hastily general conclusions from the results of just a few studies[30].

Vignoles et al. (2000) provide a detailed review of recent US and UK studies, distinguishing between different inputs and making clear how well each study deals with methodological issues (endogeneity, aggregation bias, functional form, omitted variable bias)[31]. For the US, they find that the evidence on aggregate expenditure is not reliable, so that one needs to consider different inputs separately. Higher-quality studies do find a significant impact of class-size reduction on student achievement, but this impact would be too small to justify the implied increase in expenditure on cost-efficiency grounds. The evidence on teacher characteristics is also rather ambiguous. Teacher experience seems to have a positive impact on achievement, but this is non-linear, as only the first few years of experience are associated with significant, positive and large coefficients. Meanwhile, there is some robust evidence of a positive impact of teacher salaries, but none concerning teacher education. Interestingly, more precise estimates are found when using detailed input measures (e.g. distinguishing between initial and further years of teacher experience, or teacher qualifications by subject areas) rather than less informative summary measures.

UK studies are typical of the literature pertaining to other developed countries than the US: 'The UK literature has relatively few methodologically strong studies. It is also patchy and lacks both depth and breadth of coverage with respect to the different phases of education and datasets used. The research has been restricted by the lack of suitable and accessible data […]' (p. 36). However, Vignoles et al. do mention a few recent education production function studies using school-level data: 'More recent studies utilising student level data from the NCDS [National Child Development Survey], together with school and LEA- [Local Education Authority] level resource input data, have been more successful in controlling for endogeneity and detecting some school resource effects. The larger range of variables has enabled these studies to make progress in reducing omitted variables and endogeneity bias. They have also used more sophisticated and varied model specifications. These studies have produced some evidence of school quality variables impacting positively on non-exam outcomes. However, these studies have not always utilised measures of school output with high construct validity and have had recourse to only limited data for school level resource utilization and for instruments for controlling for the endogeneity of the school resource variables' (p. 49). On the whole, there is little reliable evidence on most inputs, but for some teacher quality variables; interestingly, several studies find a significant impact of the type of secondary school attended on exam results — students enrolled

[30] Unsurprisingly, Hanushek and Krueger disagree on this point as well. 'There is a tendency by researchers and policy makers to take a single study and to generalize broadly from it. By finding an analysis that suggests a significant relationship between a specific resource and student performance, they conclude that, while other resource usage might not be productive, the usage that is identified would be.' (Hanushek, 2003, p. 89). 'Personally, I think one learns more about the effect of class size from understanding the specifications, data, methods and sensitivity of the results in the few best studies than from summarizing the entire literature' (Krueger, 2003; p. 60).

[31] See their paper for references. The US sample includes 10 studies, of which Hanushek, Rivkin and Kain (2000) is summarized in this subsection, and Hoxby (2000 a), Krueger (1999) and Krueger and Whitmore (2002) are discussed in section 4 below, for they are based on experimental methods. Vignoles et al. also include a couple of studies based on cross-section regressions discussed in section 3 above (Barro and Lee, 1996, and Gupta, Verhoeven and Tiongson, 1999) or pertaining to other developed countries, namely Finland and Israel (the latter is the paper by Angrist and Lavy, 1999 discussed in subsection 4.2 below).

in private or 'grammar' schools tend to perform better than those enrolled in 'comprehensive' and especially 'secondary modern' schools. This points to the importance of taking institutional issues into account, a topic that will be addressed in sections 6 and 7 below.

In a recent paper, Rivkin, Hanushek and Kain (2002) are able to solve a number of methodological difficulties by using the unusually rich dataset compiled since 1993 by the Texas Schools Project at the University of Texas at Dallas (UTD). This dataset, which is drawn from the administrative records of the Texas Education Authority, covers all 3,000 public schools in the state, and cohorts of over 200,000 students each. The available information includes student and teacher characteristics along with scores to the tests administered every year through the Texas Assessment of Academic Skills (TAAS) — a limitation is that individual pupils cannot be matched with individual teachers; the match is between pupils and teachers belonging to the same grade within a school. Rivkin, Hanushek and Kain focus on the mathematics scores of 4th, 5th and 6th graders, in a panel spanning over the 1993-97 period. Taking advantage of the panel nature of the data, they introduce fixed effects for students, grades (within each school) and years (within each school), thus controlling for unchanging students or school characteristics or educational policies that are either common across all grades or unique to specific grades, as well as variations in time not captured by the other variables included in the analysis. This reduces biases in the estimates due for example to parental residential choices based on school quality and the use of student characteristics by school managers when placing students into programs and classes.

Rivkin, Hanushek and Kain present two sets of results. First, they attempt to estimate the variation in overall teacher quality, defined as the contribution of each teacher to her pupils' achievement gains during the school year. Interestingly, within each school, there is little correlation between the achievement gains during the same year of pupils belonging to different cohorts and thus enrolled in different grades, who share the same school environment but have different teachers. Meanwhile, there is much more correlation between the achievement gains made in different years by different cohorts of pupils studying with the same teacher. This suggests that differences in teacher quality explain a large part of differences in achievement gains over the school year. Rivkin, Hanushek and Kain formalize this idea. They use a 'difference-in-difference-in-difference' estimator that compares the difference between the achievement gains made by students belonging to the same cohort in a given grade and the achievement gains made by the same students in the next grade with the same difference for students belonging to the next cohort, and relates it to the proportion of new teachers every year in each grade (the latter measures variation in teacher quality). The regression of the triple-differenced test scores on the rate of teacher turnover yields a coefficient from which an estimate of the variance in teacher quality can be derived. Teacher turnover has a significant and positive impact on test scores in the various specifications used, and the preferred estimate implies that 'a one standard deviation increase in average teacher quality for a grade raises average student achievement in the grade by at least 0.11 standard deviations of the total test score distribution' (p. 19).

The second set of results attempts to identify the sources of this variation in overall teacher quality, and in particular to check whether they are related to readily observable teacher characteristics like education and experience. This is done estimating more usual production functions, regressing the gain in achievement during a school year (a value-added specification) on class size, the proportion (for each grade within each school) of teachers with 0 or 1 year of experience, and the proportion of teachers with a master's degree. Significant class size effects are found for grades 4 and 5, though not 6. The proportion of inexperienced teachers has a negative impact on test scores, while the impact of teacher education is insignificant. On the whole, these effects are larger in grades 4 and 5 than in grade 6, which suggests that they fade out as students progress through the curriculum. They are also much smaller than the estimated variation in overall teacher quality would suggest: '[…] the magnitude of the effects of these variables pale in comparison to the total effect of teacher quality. This finding explains much of the contradiction between the perceived role of schools as important instruments in the struggle to raise living standards and reduce social and economic inequality and research that relegates schools to a subsidiary position far below the family: there [are] very important but difficult to measure quality differences among teachers and schools' (p. 3). The education production function approach thus allows Rivkin, Hanushek and Kain to identify differences in teacher quality, but not to explain them.

Turning to developing countries, Glewwe (2002) provides a detailed examination of education production function studies which used data specifically collected for that purpose and are thus among the best available[32]. Harbison and Hanushek (1992) focus on primary schools located in rural areas the Brazilian Nordeste. They regress test scores in reading and mathematics collected in 1981, 1983 and 1985 on school characteristics (a facilities index, a writing materials index, the availability of textbooks, a dummy for multigrade teaching), teacher characteristics (the pupil-teacher ratio, teacher experience, salary, education and training, as well as the teacher's own test scores). Several of these variables (the facilities and writing materials index, the availability of textbooks, teacher salaries and education) have the 'expected' positive and significant impact on test scores, while others (multigrade teaching, the pupil-teacher ratio, teacher experience and training) attract usually insignificant, sometimes even negative coefficients. The effects are not very large. Glewwe and Jacoby (1994) examine middle schools in Ghana. They also regress test scores in reading and mathematics (collected in 1988-89) on a large number of school and teacher variables, and find mostly small and statistically insignificant effects. Exceptions include an indirect effect of teacher experience through increased attainment, and a relatively large impact of repairing leaking classrooms, which helps to reduce the number of school days lost due to rains. Glewwe, Grosh, Jacoby and Lockheed (1995) use data covering Jamaican primary schools that include an unusually broad set of school characteristics. They regress test scores in reading and mathematics collected in 1990 on over 40 school and teacher characteristics, including pedagogical processes and management structure, but still find that most coefficients are statistically insignificant. Exceptions include textbook availability and teacher training within the past three years, as well as routine academic testing of students and textbook use, while the amount of class time devoted to written assignments has a negative impact. These results are suggestive of the importance of taking actual classroom processes into account. Kingdon (1996) uses data collected in 'middle schools' (providing the first years of secondary education) in the city of Lucknow, India, in 1991. She regresses reading and mathematics scores on teacher characteristics (years of general education, years of teacher training, marks received on official teacher examinations, years of teaching experience, salary) and school variables (class size, an index of buildings and equipment, hours per week of academic instruction). She finds that teachers' examination marks and years of general education, as well as buildings and equipment and the number of hours of instruction have a positive impact either on reading or mathematics, or on both. However, class size has a puzzling *positive* impact on reading scores, and the impacts are not very large in general.

Glewwe then provides a critical assessment of these studies. Various attempts are made to include usually unobserved variables, e.g. Glewwe and Jacoby and Kingdon include measures of 'innate ability', and all studies include as many school characteristics as possible. However, omitted-variable biases remain likely, for variables adequately describing pupil, parent or teacher motivation as well as classroom processes are not available. On the whole, while these studies show a better awareness than earlier studies of the problems surrounding the estimation of education production functions, they are not able to overcome all of them.

The main conclusion to be drawn from the education production function literature is… that it is inconclusive. Hundreds of estimates published over the last three decades and pertaining to the US, a few other developed countries and a relatively large set of developing countries have yielded no uncontroversial empirical regularities on which educational policies could be based. A consensus is building on the inadequacy of the datasets and empirical methods used in most existing studies. While an increased availability of richer data such as Rivkin, Hanushek and Kain's (2002) Texan sample coupled with more rigorous analysis along the lines proposed by Todd and Wolpin (2003) might result in the publication of more reliable studies, there is growing uncertainty that it will. Glewwe (2002) thus concludes that 'the econometric problems inherent in conventional estimates of educational production functions are so daunting that it would be unwise to place much confidence in their results. […] Future work that attempts to estimate production functions should eschew conventional estimation methods […]' (p. 465). As far as the results are concerned, there is no consensus either on whether summarizing them through meta-analytical techniques — as opposed to surveying a few key studies — makes sense (and, if this is the case, how the meta-analysis should be conducted) or on

[32] Fuller (1987) and Fuller and Clarke (1994) provide comprehensive surveys of the earlier developing-country literature, and like other authors reach the conclusion that the evidence on the determinants of achievement is unclear and unreliable.

whether regularities do emerge or not. Given this level of uncertainty, it appears that the conclusions drawn by the various authors who have surveyed the field depend on their a priori opinions as much as anything else, and could hardly be used to guide educational policies.

Indeed, attempts to estimate education production function parameters using structural econometric techniques are becoming less frequent than the evaluation of the impact of various policy interventions on educational outcomes, using experimental techniques. Section 4 reviews this newer and more conclusive literature.

4. THE EXPERIMENTAL EVALUATION OF EDUCATIONAL POLICY INTERVENTIONS

The difficulty of estimating educational production functions has led to the development of a new approach to the measurement of the impact of school resources on educational outcomes: The experimental evaluation of policy interventions. This approach does not seek to identify the parameters of a theoretical model of educational production of potentially universal relevance, but to measure the impact of a given change in school resources on educational outcomes in a specific institutional context. The critical feature of this approach is that the exposure of students to the change is determined by a rule that the researcher can identify and that is independent of other determinants of achievement. The paradigm is no longer the empirical test of economic theory but the practice of randomized trials that is standard in biology and medicine. Indeed, it is argued that the randomization of the selection of students participating in a policy intervention removes the endogeneity of the variables representing this intervention. This allows relatively simple measures of the difference between students belonging to the 'treatment group' (who benefited from the intervention) and the 'control group' (who did not) to yield unbiased estimates of the intervention impact (the so-called 'impact of treatment on the treated').

In fact, as Todd and Wolpin (2003) demonstrate, structural estimates of education production function parameters and experimental estimates of policy intervention effects are not directly comparable. The former pertain to a technical relationship between inputs and outputs, and answer questions such as: 'How would an exogenous change in class size, holding all other inputs constant, affect achievement?' (p. 8). The latter measure the total effect of an intervention, including both its direct impact on educational production and any impact that may be mediated through household behavior, for example, answering questions such as: 'What would be the total effect of an exogenous change in class size on achievement, that is, not holding other inputs constant?' (p. 8). This, and the fact that 'structuralists' and 'experimentalists' tend to form separate scientific communities, justifies that a separate treatment be given to this literature.

Randomized experiments belong to the field of policy evaluation as much as to the economics of education per se (see Burtless, 1995, Duflo and Kremer, 2003, and Newman, Rawlings and Gertler, 1994, for extensive discussions of the use of randomized trials as an evaluation tool). Their proponents argue that they provide a solution to the deficiencies that have plagued the education production function literature, as argued by Kremer (1995) in his answer to Hanushek's (1995) survey of the developing-country literature: 'I concur that little is known about what improves school quality and that additional nonrandomized studies are not likely to add to our information. Randomized trials do hold great promise, however. […] Randomized trials revolutionized medicine early on this century, and there is no reason they should not revolutionize education early in the next century' (p. 251). The evaluation of policy interventions through randomized experiments has long been practiced in the United States and is becoming a standard tool in development programs, notably World Bank-funded projects. Such experiments remain less frequent in the field of education, but have already provided credible evidence about the effectiveness and cost-efficiency of alternative policy interventions.

Experiments are not without having their own weaknesses, however. They are usually conducted at a small scale, thus the generality of their results and the possibility to scale up interventions evaluated as successful are not guaranteed (on this, see Duflo, 2003). They may induce selective migration, and practical difficulties often result in the selection of treated students not being fully random.

Furthermore, conducting experiments in education is often difficult on ethical, political or financial grounds, e.g. it can be difficult to justify the exclusion of the treatment group from a potentially beneficial intervention. An increasingly common way to generate experimental data without raising such issues is to randomize the phasing in of programs that are to be introduced gradually in any case.

Experimental estimation techniques may find another application in non-experimental settings where history or institutions generate exogenous variation in educational policies. A case that has been widely exploited in the returns-to-education literature is the changes across US states and during the twentieth century of laws regulating child labor and providing for compulsory schooling, which have induced exogenous variation in the duration of schooling, at least for individuals who acquired approximately as much education as the legally prescribed quantity. Such circumstances are termed 'natural experiments' or 'quasi experiments'.

This section first examines the results of randomized experiments (4.1), before turning to natural experiments (4.2).

4.1. Randomized experiments

The developed-country literature revolves around a single experiment that dates back to the late 1980's, the Tennessee Student-Teacher Ratio Achievement Ratio (STAR) project, which Krueger (1999) presents as 'the only large-scale randomized experiment on class size ever conducted in the United States' (p. 498). Project STAR covered relatively large schools which had at least three classes per grade, and, within each school randomly assigned one class to each of three groups: a first treatment group with small classes of 13 to 17 students, a control group with regular classes of 22 to 25 students, and a second treatment group with regular classes and a full-time teacher's aide. The latter treatment was discontinued after the first year, as it was found to be ineffective, and the classes which had been assigned to it were merged into the control group. The project lasted for four school years, during which students remained in the same group. In the first year, 1985-86, 79 schools participated in the project; the first treatment group comprised 1,900 students; the control and the second treatment group comprised 4,424 students, all enrolled in kindergarten ; a total of 11,600 students participated over the four years of the experiment. Standardized tests were administered to all students at the end of each year: Estimates of the impact of project STAR pertain to differences in test scores, within each school, between pupils belonging to the treatment and control group.

The results of project STAR were mostly published in educational science journals, but they have been re-analyzed by economists, notably Krueger (1999; Krueger and Whitmore, 2001, 2002) and Hanushek (1999 a, b)[33]. Hanushek (2003) thus summarizes the results of educational science studies: '1. Pupils in small classes perform better than those in regular classes or regular classes with aides starting in kindergarten; 2. The kindergarten performance advantage of small classes widens a small amount in first grade but then either remains quantitatively the same (reading) or narrows (mathematics) by third grade; and, 3. Taking each grade separately, the difference in performance between small and regular classes is statistically significant' (p. 87).

In his reassessment of project STAR, Krueger (1999) presents 'experimental estimates of education production function'. Using data for children enrolled in kindergarten, and in the first, second and third grades, he regresses test scores on dummy variables indicating whether the student belongs to the treatment or control groups, controlling for a variety of student and teacher characteristics (the latter include race, experience and education) and including school fixed effects. He thus obtains estimates in which the coefficient of interest (the impact of class size) is identified through the random assignment of students between the treatment and control group — randomization removes the potential endogeneity of class size that would have affected the same regressions if they had been based on non-experimental data. The impact of small class size is significant and positive for all eight specifications presented for each of the four grades: 'the gap in average performance is about 5 percentile points in kindergarten, 8.6 points in first grade, and 5-6 points in second and third grade' (p. 511). Interestingly, teacher characteristics have little impact on test scores, but for a small, positive

[33] See these papers for references to the educational science literature on the STAR project.

impact of experience peaking at twenty years of experience: 'consistent with much of the previous literature, the STAR data suggest that measured teacher characteristics explain relatively little of student achievement on standardized tests' (p. 514).

The rest of the paper examines the robustness of these results to imperfections of randomization (due for example to switches of students between small and regular classes between grades and the attrition of the sample over the years). Krueger also explores the cumulative impact of remaining in a small class over several school years as well as differences in project impact between groups of students (gender, eligibility to free lunches, race, urban vs. rural residence). Krueger concludes that 'adjustments for school effects, attrition, re-randomization after kindergarten, nonrandom transitions and variability in actual class size do not overturn the results of [earlier studies]: students in small classes scored higher on standardized tests than students in regular-size classes' (p. 528). Meanwhile, the impact of attending a small class is not strongly cumulative: Most of the impact is observed after the first year, with the additional impact of subsequent years being still positive but smaller. Finally, the impact tends to be larger on students eligible for free lunches, blacks students and inner-city students, which suggests that children belonging to disadvantaged social backgrounds benefit most from reduced class size.

The data used by Krueger (1999) follow students only until grade 3 — from grade 4 onwards, all pupils attended regular classes — and it is unclear how much impact the achievement gains allowed by class size reduction in their first years of primary schooling will have on their subsequent schooling and adult outcomes. Krueger and Whitmore (2001) investigate the long-run impact of project STAR by estimating differences in test scores through the eighth grade, the probability of taking college entrance exams and scores obtained in these exams (students enrolled in kindergarten in 1985 had graduated from high school by 1998). While they find some evidence of a long-run effect on test scores, their main result is that the probability for the students who had attended a small class in kindergarten to grade 3 to take the college entrance examinations was 43.7 %, as opposed to 40.0 % for those who had not.[34] The impact was much larger for Black students, 40.2 % against 31.7 %.

Krueger and Whitmore (2002) further invest the gap between Black and White students. Their results suggest that reducing class size has a much larger impact on Black students, whose test scores from kindergarten to grade 3 increased by 7-10 percentile points, as opposed to 3-4 percentile points for White students; test scores in further grades, after the end of the experiments, increased by 5 vs. 1.5 points a year. The probability of taking college entrance examinations is here estimated to increase from 31.8 % to 41.3 % for Blacks as a result of the experiment, as opposed to 44.7 % to 46.4 % for Whites. Krueger and Whitmore then extrapolate from these findings from the Tennessee STAR project to estimate the contribution of trends the in pupil-teacher ratio to the closing of the achievement gap between Black and White students at the national level — which is far-fetched. They argue that almost all of the narrowing of the gap in NAEP test scores observed since 1971 can be explained by changes in the PTR, and that systematizing small classes would reduce the racial gap in taking a college entrance examination by 60 %.

Not quite unexpectedly, Hanushek (2003), based on his earlier work (1999 a, b), takes a different view on project STAR. While his own work also shows a significant and positive impact of class size reduction on achievement (on average, reducing class size by 8 students raises test scores by 0.2 standard deviations), he argues, that, contrary to Krueger's (1999) evidence, the limitations to the randomization of the experiment were serious. Further, project STAR would have over-sampled disadvantaged urban and minority schools, leading to an overestimation of the average program impact. He thus concludes: 'The one limited and flawed experiment in Tennessee cannot be taken as providing the definitive evidence needed for policy changes that cost billions of dollars annually. At best it provides evidence about the potential impact of very large changes in class size applied to kindergarten students, and there is direct evidence that these findings do not generalize to other grades and other situations' (p. 89).

[34] They do not find much impact on scores to these tests, however.

An important feature of project STAR, on which both Krueger and Hanushek insist, is that its impact is larger in grade 1 than in grades 2 and 3, while a cumulative impact would have been more plausible on a priori grounds. As argued by Hanushek (2003), 'this pattern of effects is at odds with the normal rhetoric about smaller classes permitting more individualized instruction, allowing improved classroom interactions, cutting down on disruptions, and the like. If these were the important changes, small classes should confer continuing benefits in any grades where they are employed. Instead, the results appear more consistent with socialization or introduction into the behavior of the classroom.' Krueger (1999) makes a similar point, although he disagrees with Hanushek's conclusion that this implies that a generalization of class size reduction policies to other grades would not be effective.

The results of project STAR are important in so far as they provide admittedly robust evidence of class size effects in the US context. It seems that this intervention in early school years has substantial long-run effects, especially in terms of reducing social inequalities in access to higher education. This echoes Carneiro and Heckman's (2003) argument that such inequalities are due to disparities in the accumulation of cognitive skills from the first years of schooling onward, and cannot be explained satisfactorily in terms of family or school characteristics by the time of high school graduation. Developing pre-schools and improving primary schools would thus have larger returns, in the US at least, than targeting cognitive-skill-based policies to adolescents[35].

The developing-country literature originates in two randomized experiments conducted in the late 1970s, in Nicaragua (Jamison, Searle, Galda and Heynemann, 1981) and the Philippines (Heynemann, Jamison and Montenegro, 1984). The Nicaragua experiment tested the impact of providing distance education, in this case radio mathematics instruction, to isolated schools. 48 grade-1 classrooms received radio education, 20 received mathematics workbooks instead, while another 20 constituted the control group. After one year, average test scores were higher by 1 standard deviation in the radio education treatment group than in the control group, and by 0.3 standard deviation in the workbook treatment group than in the control group: Radio education appeared a more effective way of improving achievement than the more conventional workbook provision.

The Philippines experiment focused on textbooks. 26 schools received one textbook per student in mathematics, sciences and Filipino; 26 received only one textbook per two students; the control group comprised 52 more schools. After one year, the test scores of both treatment groups were higher by 0.4 standard deviation than those of the control group, but there was no significant difference between the two treatment groups, suggesting that the provision of textbooks for each child might have been unnecessary. Despite possible biases due to the fact that the experiments tended to raise enrollment, these results were quite promising and indeed, radio education spread over Latin American during the 1980s and 1990s. However, no more randomized experiments were conducted until the mid-1990s — the production function framework dominated research and policy debates for two decades.

Over the last few years, a number of studies have started accumulating. These studies have been produced either as part of the evaluation of World Bank projects or by researchers based at Harvard University and the Massachusetts Institute of Technology and working on Kenya and India, respectively. Glewwe (2002), Kremer (2003) and Duflo and Kremer (2003) provide brief surveys of these papers, on which the following discussion draws heavily.

Several studies have focused on experiments aimed to increase attendance rather than achievement itself, but are still worth mentioning here. Increasing school participation has been the aim of the well-publicized Progresa program conducted by the Mexican government with the World Bank. Amongst other anti-poverty measures, Progresa includes the payment of cash grants to families, conditional on their sending their children to school. The program was phased in gradually, with randomization at the household level, allowing evaluation. Progresa data have been diffused quite widely and a number of evaluations are available, that may be accessed through the World Bank's website. Schultz (2004) finds that Progresa increased enrolment by 3.4 % for all students in grades 1 to 8, and up to 14.8 % for

[35] The benefits of early childhood interventions in the US are the focus of a sizeable literature that includes randomized experiments. See Currie (2001) for a survey.

girls who had completed grade 6. Other evaluations have also found the program to be successful, so that it has been continued and expanded, as well as emulated in other Latin American countries.

Vermeersch (2002) examines the impact of the provision of midday meals in a sample of 25 treatment and 25 control Kenyan pre-schools, and finds that attendance increased by 30 %. The program was also expected to have an impact on test scores, but it happened that the organization of the meals cut into instruction time, so that an increase in test scores (by 0.4 standard deviations) was observed only in schools in which teachers had received substantial training. Kremer, Moulin, Namunyu and Myatt (1997) estimate the impact of the provision by an NGO of free uniforms, textbooks and additional classrooms on 7 treatment and 7 control Kenyan schools which were performing poorly. The intervention reduced dropout rates in treated schools, resulting in an increase in average schooling completed by 15 % after five years. Furthermore, a large number of parents transferred their children from nearby schools to treatment schools, leading to an increase in class size by 50 %, without any impact on test scores. This suggests that parents were willing to trade off class size for direct costs, in a context in which the former did not have a major impact on achievement. A policy implication is that it may be possible to raise class size and devote the funds thus saved to increasing school participation and providing more inputs to schools. Miguel and Kremer (2004) show that a twice-yearly, school-based treatment of Kenyan children with de-worming drugs, which covered 75 schools and was phased in randomly, decreased absenteeism by 25 %, or 7 percentage points, and spilled over to nearby schools through reduced worm transmission to non-treated children. The total increase in schooling was 0.15 years per person treated. Bobonis, Miguel and Sharma (2002) obtain comparable estimates of a de-worming program in India. Kremer (2003) thus concludes: 'Overall, these results suggest that school participation is quite elastic to cost and that school health programs may be one of the most cost-effective ways of increasing school participation'.

Fewer studies in fact have evaluated the impact of providing additional inputs paralleling educational production function studies. Tan, Lane and Lassibille (1999) present the results of a randomized experiment conducted in the Philippines. 30 schools were allocated either to a control group or one of four treatment groups which received either school meals, pedagogical materials for teachers or structured meetings between parents and school officials combined with either of the other two interventions. But for the provision of pedagogical materials, the interventions did not have a significant impact on dropout rates after one year, but they had significant impacts on tests scores, especially school meals combined with meetings between parents and school officials. The results however, were quite imprecisely estimated.

Banerjee and Kremer (2002) address class size reduction: They estimate the impact of appointing a second teacher to one-teacher schools in India, and find that it increased female school participation but had no impact on test scores. Meanwhile, a program that provided teacher training and other inputs to pre-school teachers who had begun teaching after receiving only minimal training was found to have little impact on test scores.

Glewwe, Kremer and Moulin (2002) find that the provision of textbooks to Kenyan schools increased test scores by about 0.2 standard deviations. However, the impact was concentrated among students who had scored in the top one or two quintiles in tests administered before the program was started; there was no impact on the rest 60 % of the pupils. The intervention thus tended to be biased towards children belonging to relatively privileged families. The authors mention that Kenyan textbooks are written in English and reflect a curriculum designed for elite families in Nairobi, that may be more difficult for rural children to understand. Glewwe, Kremer, Moulin and Zitzewitz (2004) further examine the impact of providing flip charts, designed to be more accessible to all children, but find no significant difference between treatment and control schools.

Banerjee, Cole, Duflo and Linden (2003) report the results of a large-scale experiment conducted in rural India, that focuses on the relationship between pupils and teachers. The intervention consists of a 'remedial education' program run by an NGO, Pratham, in primary schools located in slums of Indian cities. Pratham hires young women (called 'balskahi') who hold the equivalent of a high school degree and belong to the same community as the children, and trains them to teach a group of 15 to 20 children for two hours every morning, and another group in the afternoon. Teachers receive two

weeks of initial training, and further training during the year. The courses focus on core competencies such as the basic numeracy and literacy skills which children should have learned in the second and third standards, and are addressed to children who are lagging behind. The expected benefits of the intervention are quite complex. They include the impact of remedial education on children who attend it; the impact of reduced class size for the other children, who continue attending school with the regular teacher during the remedial education sessions; and the fact that the material covered in those sessions may not need to be re-taught during regular classes.

Banerjee, Cole, Duflo and Linden ran their experiment in the cities of Mumbai where the balsakhi program started in 1994, and Vadodara, where it started in 1999. The experiment consisted in randomizing the expansion of the program in both cities. For example, in Vadodara, the program was extended over the years 2000-2002 to the 98 eligible schools which had not yet been covered. In 2000-01, balskahis were sent to half of those schools, selected randomly, to teach grade-3 pupils. In 2001-02, balsakhis were sent to the same schools to teach grade-4 pupils and to the remaining schools to teach grade-3 pupils. Finally, in 2002-03, balsakhis taught grade-3 pupils were grade-4 pupils had been taught the previous year, and conversely. A similar procedure was followed in Mumbai. Tests of cognitive achievement (literacy and numeracy) were administered in all schools before the intervention and at the end of each school year.

This experimental study has several desirable features that make it particularly relevant both to educational policy and to the methodology of randomized experiments. First, the experiment was conducted over a large scale: More than 15,000 students were included, which allows a precise estimation of the intervention impact. Second, the experiment was implemented simultaneously in two cities, with distinct management teams, which to some extent enables the authors to check how context-specific their results may be. Third, the study was conducted over several years, and the use of the time-dimension of the data allows the authors to solve common econometric issues. Fourth, the balsakhi program itself is characterized by its low cost and easy replication, and has indeed already been scaled up and extended to other cities (it is now implemented in as many as 20 cities). Fifth, this study is unique in so far as it explicitly focuses on the quality of education defined as the relationship between teachers and pupils and pedagogical practices: ‘The intervention is motivated by the belief that children often drop out because they fall behind and feel lost in class.’ (p. 3). ‘According to Pratham, the main benefit of the program is to provide individualized, non-threatening attention to children who are lagging behind in the classroom and are not capable of following the standard curriculum. Children may feel more comfortable with women from their own communities than teachers, who are often from different backgrounds. As the balsakhi’s class size is relatively small, she may tailor the curriculum to the children’s specific needs.’ (p. 5).

The authors first provide estimates of the impact of the intervention on all children enrolled in treated schools, and then within those schools they distinguish between the impacts on children who attended the balsakhi sessions and those who did not (who benefited from reduced class sizes and ‘higher average quality of their classmates’ while the weaker students were attending the balsakhi sessions). The results can be summarized as follows. First, there was no impact on children’s attendance, which may respond more to household than to school factors. Second, there was a significant impact on achievement, which is remarkably similar across cities: ‘Test scores of children who benefited from the program improved by 0.12 to 0.16 standard deviations in the first year, and 0.15 to 0.3 standard deviations in the second year. […] Results are even stronger for children in the bottom of the distribution (in the bottom third of the distribution, the program improved test scores by 0.22 standard deviations in the first year, and 0.58 in the second year’ (p. 23). Moreover, the results suggest the direct impact of balsakhi sessions on children who attended them accounts for most of this improvement in test scores, as opposed to reduced class size for the other children; this also implies that balsakhis are more effective than regular teachers. The authors draw the following conclusion: ‘This study demonstrates both the efficacy of the remedial education program, and more generally, the feasibility of dramatically impacting test scores at very low cost. […] The estimates suggest that reducing class size by hiring a balsakhi is at least twice as effective as reducing class size by keeping children with regular teachers’ (p. 4), while balsakhis are paid only a fraction of teachers’ salaries.

Glewwe (2002) and Kremer (2003) reach similar conclusions concerning randomized experiments in education: They allow a much more convincing estimation of the impact of various policy interventions than conventional education production function studies. According to Kremer, 'randomized evaluations can [even] shed light not only on the impact of specific programs, but also on behavioral parameters and questions of more general interest'. A number of methodological difficulties need to be adequately taken into account, however. For example, the experiments should be conducted over a large scale, to yield large samples and correspondingly precise estimates; data should be collected over several years to make it possible to check whether any significant impact fades out or is strengthened as the intervention is continued, or after it has been discontinued; detailed data on schools and pupils should be collected to check that the randomization of the experiment has been correctly implemented. Kremer (2003) and Duflo and Kremer (2003) further insist that randomized experiments are relatively easy to implement, and that local NGOs are often better partners than governments in that regard. They argue they should become a standard tool for institutions involved in development aid: 'Ideally, the World Bank and other development funders would require pilot programs and randomized evaluations before launching large-scale funding of new policies which are prone to evaluation, just as regulators require randomized trials before approving new drugs.' (Kremer, 2003).

4.2. Natural experiments

Studies of natural experiments are based on econometric techniques that are similar in inspiration to those used in studies based on randomized experiments. In both cases, some source of exogenous variation in a determinant of achievement can be identified. Natural experiments in education are usually the result of the application of some policy rule that can be modeled by the researchers, and is independent of the usual functioning of schools. Their obvious disadvantage relative to randomized experiments is that researchers do not have control over them, and the necessary information may not be available: The number of exploitable natural experiments may not be very high. An important advantage, however, is that natural experiments, being based on actual policies, are necessarily policy-relevant and typically conducted over a large scale. Furthermore, as they are not experienced as experiments by the pupils and teachers they affect, they do not generate so-called 'Hawthorne effects' whereby knowing that one is participating in an experiment affects one's behavior independently of the impact of the intervention experimented with itself[36].

The technique was pioneered by Angrist and Lavy (1999)'s study of class size in Israel — a paper that greatly contributed to renew debates on both methodology and the evidence concerning class size in developed countries. Israeli schools are allocated teachers following a rule formulated by the medieval scholar Maimonides, according to which class size should not exceed 40 pupils. As a result, a school with, say, 39 pupils will have one teacher, but a school with 41 pupils will have two. This generates a systematic, nonlinear and nonmonotonic relation between the number of pupils enrolled in a school and the pupil-teacher ratio, that is uncorrelated with unobserved characteristics of pupils that affect learning[37], and class size prescribed by Maimonides's rule can be used as an instrument for actual class size. Angrist and Lavy use data collected in 1991 and 1992 that include scores to a national test administered to Israeli third-, fourth- and fifth-graders and cover Jewish public schools[38]. The sample is quite large, as it comprises 1,000 schools and 2,000 classrooms for each grade — the data are aggregated at the classroom level. As would be the case with data based on a random experiment, the experimental design allows the authors to include class size, the percentage of disadvantaged students

[36] It has been argued that in the Tennessee STAR experiment, teachers, who favored small class sizes independently of their impact on their pupils' achievement, might have striven to teach more efficiently than under usual circumstances, so that the experiment be shown to have a positive impact on achievement and be continued. In such a case, the increase in test scores would be the result of the change in teacher incentives generated by the experiment rather than that of reduced class size. A generalization of small class size would not yield any increase in test scores, for teachers would no longer have the same incentive to teach better. In the case of project STAR, however, Krueger (1999) finds no evidence of Hawthorne effects.

[37] Such a correlation may still exist if the application of the rule differs according to the socio-economic composition of neighborhoods, for example if privileged areas more easily obtain additional teaching positions as the pupil-teacher ratio nears 40. Angrist and Lavy (1999) however provide evidence that the rule is quite strictly adhered to.

[38] 'Except for higher education, schools in Israel are segregated along ethnic (Jewish/Arab) lines. Within the Jewish public school system, there are also separate administrative divisions and curricula for secular and religious schools' (Angrist and Lavy, 1999, p. 538). In an earlier production function study, Lavy (1998) examined the differences in test scores between Arab and Jewish pupils and found that disparities in school resources provided by the Israeli government largely explained the weaker achievement of Arabs.

in the school and total enrollment for the grade as the only explanatory variables — other determinants of achievement should be independent of class size and their inclusion would not alter the coefficient on that variable, which is the parameter of interest. Angrist and Lavy find a significant and negative impact of class size on the reading and mathematics scores of fifth-graders: A decrease in class size by 1 standard deviation (6,5 pupils) results in an increase in reading test scores by 0.2 to 0.5 standard deviations and an increase in mathematics test scores by 0.1 to 0.3 standard deviations. The effects on the achievement of third- and fourth-graders, however, were often insignificant.

Angrist and Lavy's results have been criticized by Hoxby (2000 a), who argues that any identification strategy that relies on policy decisions may still suffer from endogeneity bias — however indirectly, policy decisions are in the end correlated with parental preferences, especially when school policies are decentralized. This skepticism about natural experiments is shared by Rosenzweig and Wolpin (2000), who argue that one should focus on 'natural' natural experiments in which natural events rather than policy decisions provide the exogenous source of variation in the explanatory variable — a common application is the use of rainfall as an instrument for income in rural areas of developing countries, where agricultural output is a major source of income and is crucially affected by the amount and timing of rainfall.

Hoxby (2000 a) examines two such 'natural' natural experiments in the US state of Connecticut. Her first identification strategy relies on the fact that there is random variation in the number of children born in a given month or year that translates into variation in class size, for children are enrolled in grade 1 in the month of September preceding the year in which they will complete the age of 6: If an unusually large number of children were born in November and December of a given year, and an unusually small number in the following January and February, then class size in grade one will be larger than usual six years later and smaller seven years later. The second identification strategy is used for comparison purposes; like Angrist and Lavy's, it relies on the variation in class size generated by the application of maximum class size rules, although Hoxby's use of panel data allows her to correct for some biases that Angrist and Lavy could not get rid of due to the cross-section nature of their own dataset. Hoxby implements her identification strategies on a dataset covering the 649 state elementary schools of Connecticut (these belong to 146 districts, each of which has its own maximum class size rule) over the years 1986-87 to 1997-98. Class size varies between 10 and 30. The results based on the two independent identification strategies are consistent: There is no significant impact of class size on test scores. Indeed, 'the estimates are sufficiently precise that, if a 10 percent reduction in class size improved achievement by just 2 to 4 percent of a standard deviation, I would have found statistically significant effects in math, reading, and writing. I find no evidence that class size reductions are more efficacious in schools that contain high concentrations of low income students or African-American students' (p. 1282).

Krueger's (1999), Angrist and Lavy's (1999) and Hoxby's (2000 a) results complement each other, but it is unclear which conclusions may be drawn from the contradiction between their results. Indeed, they differ in both methodology and context, and although Hoxby provides some convincing evidence that using Angrist and Lavy's methodology with her own data would yield a fallacious impact of class size on test scores, it is not certain that primary schools in Connecticut and Israel can be compared — if only because class size is much larger on average in Israel than in the United States.

In a more recent paper, Angrist and Lavy (2001) have, refreshingly, addressed another topic than class size, namely, in-service teacher training. As is the case for class size, production function evidence on teacher training is mixed, but it tends to focus on initial training, while in-service training could have a stronger impact. Angrist and Lavy use the fact that a small number of Jerusalem schools received funds earmarked for teacher training in 1995, and match these schools with comparable schools in the same area which did not receive these funds[39]. They find that teacher training improved test scores by 0.2 to 0.4 standard deviations in secular schools. The training program focused on pedagogy rather than subject content, and was relatively inexpensive compared to class size reduction. These results

[39] 'Matching' is an econometric technique that does not explicitly rely on the experimental paradigm, but similarly allows a reduction in endogeneity biases. While it has become extremely common in labor economics, its use in the economics of education is still infrequent, as noted by Glewwe (2002).

echo those obtained by Banerjee, Cole, Duflo and Linden (2003) in urban India, as both point to the importance of the relationship between pupils and teachers.

The developing country literature was also renewed by a paper on class size, namely, Case and Deaton's (1999) study of South Africa at the end of the Apartheid regime. Under the Apartheid, South African Blacks had no political representation and therefore no control over government funding of education, which was heavily centralized. Furthermore, they were imposed tight controls on their place of residence, preventing migrations. This implies that the parents of Black children had no influence on the quality of the schools their children attended. Meanwhile, Case and Deaton state that interviews with officials then in charge of educational policies revealed no consistent rule for the allocation of resources to the schools to which Black children were segregated. They thus argue that the wide variations in pupil-teacher ratios actually observed between those schools can be considered random, and exogenous with respect to family characteristics affecting learning. The natural experiment here is somewhat paradoxically the absence of any well-defined school policy due to the institutionalized racism of the Apartheid regime. Using data collected in 1993, Case and Deaton thus examine the impact of variations in the pupil-teacher ratios on years of completed schooling, current enrollment status and test scores of Black children. They find significant impacts on the three variables. For example, decreasing the pupil-teacher ratio from 40 (the mean value for Black schools) to 20 (the mean value for White schools) would increase attainment by 1.5 to 2.5 years and yield the same increase in reading test scores as two years of additional schooling under existing conditions (the impact on mathematics scores was not significant). Case and Yogo (1999) further use this approach with data from the 1996 South African census to estimate the impact of school quality on earnings, and find significant and negative effects of the pupil-teacher ratio on educational attainment, the probability of employment, and the returns to education (reducing the pupil-teacher ratio by 5 students would increase the return to education by 1 %).

There has been some controversy, however, about Case and Deaton's (1999) results. First, they use data aggregated at the district level, arguing that this further reduces any endogeneity bias. Whether this is a good strategy depends on the magnitude of the resulting aggregation bias, which Hanushek (2003) states leads to an overestimation of the impact of school resources on educational outcomes. Second, the identification strategy, which relies on the absence of any clear school policy, is questionable: As argued by Hoxby (2000 a), 'there is a difference between variation that is not obviously biased and variation that has an explicit reason to be random. The systematic links between school inputs and other determinants of student outcomes may be *obscure* without the variation in inputs being *exogenous*. Explicitly articulating a source of exogenous variation is preferable to simply eliminating all sources of bias' (p. 1245).

Owing perhaps to the scarcity of test score data that could be related with information on education policies, there do not seem to be many other studies of natural experiments. Three papers that examine the impact of educational policies on attainment and the returns to education may be mentioned here, although they to not include results pertaining to achievement. Duflo (2001, 2004) estimates the impact of the construction of more than 61,000 primary schools by the Indonesian government between 1973 and 1978. Her identification strategy relies on differences in the intensity of the program across regions and differences in cohort size over the years. She finds that children aged 2 to 6 in 1974 received 0.12 to 0.19 more years of education for each school constructed per 1,000 children in their region of birth, and that the returns to this education ranged from 6.8 to 10.6 percent (per year of education). The influx of young educated workers however reduced the wages of older cohorts, probably because physical capital did not adjust quickly enough to the increase in human capital.

Chin (2002) examines the impact of Operation Blackboard launched by the Indian government in 1987, that aimed to provide additional teaching materials and school buildings, as well as a second teacher to single-teacher schools. She focuses on the latter component, using the facts that only children enrolled in primary schools after 1987 were affected and that the number of new teachers enrolled varied across states according to the proportion of single-teacher schools. Operation Blackboard is widely considered a failure in India. Indeed, while as many as 150,000 new teachers were hired, Chin estimates that only 25 % to 50 % of them were actually appointed to single-teacher schools, and average class size did not decrease. There was some redistribution of teachers from larger

to smaller schools, however, leading to increases in primary school completion rates, especially for girls and poorer children.

To conclude this section, the use of experimental techniques, whether pertaining to randomized, 'policy' natural or 'natural' natural experiments, and related matching techniques, provides credible strategies to overcome the endogeneity of school characteristics that has plagued the education production function literature. Randomized experiments in particular have already provided convincing evidence on some policy interventions, e.g. de-worming or the appointment of more pupil-friendly teachers. Implementing randomized trials or analyzing policies that generated natural experiments also requires paying attention to the context in which the experiment takes place, and the policy-relevance of these studies is often clearer than that of education production function studies. Furthermore, these studies do not rely on theoretical modeling, so that a wider range of interventions have already been considered than in the education production function literature; empirical modeling is made simpler by the identification of an exogenous source of variation in school resources, which is ultimately a matter of policy analysis, not of statistics or econometrics. The dissemination of the results beyond academic circles is thus bound to be easier. Nevertheless, Kremer (2003) and Duflo and Kremer (2003) argue that there is a risk of publication bias, whereby only the results of successful interventions would be published, while potentially as much could be learned from failures as from 'success stories'.

As things stand, a limitation to the policy-relevance of this literature is that estimates relying on different methods are not directly comparable. First, unless one considers that experimental estimates do recover some universal structural parameters, there is no reason to expect the results of an experiment conducted in Kenya or India to be relevant to other countries. Second, as argued by Todd and Wolpin (2003), not only do production function parameters differ from estimates based on randomized experiments, which also include household responses to the intervention considered, but as well estimates based on randomized and natural experiments may not be comparable: The former represent the 'impact of treatment of the treated' while the latter represent 'local average treatment effects', and the two may differ if the population affected by a natural experiment differs from the one which would have been selected into a randomized experiment. This justifies the organization of sections 3 and 4 along methodological rather than thematic lines, and explains that they give a rather disorganized picture of policy interventions that could substantially improve achievement. But for class size — on which even experimentalists have reached no consensus — there is no substantial body of evidence for specific inputs or policy interventions.

Sections 2 to 4 have reviewed macro, micro and experimental studies which constitute the core of the literature on the relationship between educational expenditures and outcomes. Sections 5 to 7 are devoted to recent developments which may be laying the foundations for a renewal of the economics of education by questioning the relevance of the education production function conceptual framework. Section V retains a structural approach to empirical analysis, and examines extensions that have been proposed to this framework to take into account behavioral equations which determine the input levels which enter this technical relationship. Section 6 considers policies which do not rely solely on the allocation of educational expenditures across inputs to improve educational outcomes, but instead focus on the incentives faced by teachers. Section 7 further argues that the very analogy between schooling and production is fallacious. The educational process itself should not be treated as a black box, and the diversity of both schooling outcomes and their determinants should be taken into account.

5. EXTENSIONS TO THE EDUCATION PRODUCTION FUNCTION FRAMEWORK

The education production function is a technical relationship between inputs and outputs; as such, it is not a behavioral equation, and empirical studies such as those discussed in section 3 do not model the way inputs are determined. These inputs include teacher and school characteristics, which depend on educational policies that are led at different levels of aggregation (from principals to ministers) and family inputs, which depend on household behavior. Now responses from school management authorities and families to pupils' characteristics or policy interventions can affect the *observed* relationship between educational expenditures and outcomes, as opposed to the underlying technical

relationship. School management authorities may choose to allocate more inputs to pupils whose social background is less favorable; parents may reduce their own contribution to their children's education as public expenditures increase. In both cases, the observed and the technical relationship between public expenditures and outcomes will differ.

There is thus a case for supplementing education production function analysis with an explicit modeling of the behavior of parents, teachers and school administrators. This section covers such extensions to the standard conceptual framework. It starts by considering household responses to educational policy (5.1) before reviewing the evidence on 'demand-side financing', which explicitly relies on such responses to improve the functioning of schools (5.2); it then suggests that inferences about the rules governing the allocation of education expenditures may be in fact be drawn from the frequent insignificance of education production function coefficients (5.3).

5.1. Household responses to educational policy

Households are likely to respond to policy interventions which affect school inputs. In the education production function framework, a key question is whether an increase in school inputs leads to an increase in inputs provided by parents (the two sources of inputs would be complements) or to a decrease (the two sources would be substitutes). If one considers the impact of providing through schools material inputs that may alternatively be provided by parents, substitutability is most likely, unless the inputs exhibit increasing returns to scale. For example, the provision of free textbooks or uniforms is bound to reduce parental expenditure on those items rather than to increase it; this is indeed the goal of such policy interventions.

The contribution of families to schooling is however much broader than the mere provision of material inputs; a more general and policy-relevant question such as the relative roles of families and schools in determining educational outcomes is much more difficult to answer in general terms. Can increased levels of school inputs (or more generally, higher quality schooling) compensate for the lower levels of family inputs enjoyed by children from educationally or socially deprived backgrounds? This debate has been raging in the USA since the findings of Coleman and his co-authors were published in the late 1960s; it has equivalents in many other countries, whether framed in terms of economics or quantitative sociology or not.

In an even broader perspective, education at school and education in the family (or, more generally, in the social institutions which take care of children) could be understood as two different processes, which may or may not be compatible. The expansion of formal schooling tends to crowd out the informal transmission of cultures which are not represented in the classes that dominate the school system, e.g. regional languages, oral traditions of minority groups, etc — in which case educational policy results in the substitution of school for family education. At the same time, especially when at least basic schooling has already been made universal, there may be more complementarities between the kind of education transmitted by parents and by teachers.

For the time being, besides comparisons between coefficients on school and family variables in education production function regressions, economists seem to have mostly addressed the narrower question of the relationship between school and household *expenditures*. In a recent paper Das, Dercon, Habyarimana and Krishnan (2004), supplement the education production function with a model of household behavior, in which parents maximize the utility they draw from their consumption and the achievement of their children with respect to their expenditure on consumption and on educational inputs. As a result, if the inputs supplied through schools and bought by parents are technical substitutes, increases in supply-side expenditure will lead to decreases in parental expenditure, resulting in less than commensurate or even nil increases in pupil achievement. Das et al. argue that such optimal household responses may play a part in explaining the frequent insignificance of education production function parameters which do not explicitly model household behavior. They develop a two-period framework that allows them to distinguish between anticipated changes in school expenditures, which households can compensate for, and unanticipated changes, which

households cannot compensate for and whose impact can then be used to recover production function parameters[40].

They test their model using data collected in 2002-03 in Zambia, a country with relatively high enrolment rates and almost exclusively public schooling. They compare the impacts on achievement of large government cash grants uniformly distributed to schools in 2000, irrespective of enrollment — which households could anticipate — and other sources of school funding (from District Education Offices or external donors) that are less widely available, irregular and extremely variable from one school to another — it is thus difficult for households to anticipate them[41]. They find that school cash grants and household expenditures are indeed substitutes, with an elasticity of – 0.35 to – 0.52. Meanwhile, these cash grants have no impact on cognitive achievement while the other, unanticipated, sources of funding do (at the mean, unanticipated funds increase English test scores by 0.05 standard deviations and mathematics scores by 0.25 s.d.). The results apply to both mathematics and language (English) test scores, and are robust to alternative specifications.

This paper is an interesting illustration of the relevance of taking household behavior into account. A limitation is that the kind of substitutability between school and family inputs envisaged is not very informative. First, the authors do not make quite clear to which inputs either cash grants or parental expenditures are allocated, while it is those inputs, and not grants or expenditures themselves, that enter educational production. Substitutability between public and household expenditure on material inputs used by pupils is most likely, but substitutability between, say, teacher salaries and training and household expenditure, while plausible (e.g. the appointment of more competent teacher may reduce the need for private tuition), would certainly require more detailed analysis. Second, it seems neither surprising nor undesirable as such that parents in a poor country going through a long-run economic crisis should reduce their expenditure on education when public funds are made available. Indeed, reducing household expenditure on education to a minimum is an explicit goal of free education policies, which do not imply that other types of parental inputs such as home teaching, should be reduced. While the amount of substitution between supply-side and demand-side expenditure is important where schooling is not free of direct costs, estimating it does not reveal much about the learning process itself, and the relative contribution of parents and teachers to this process.

A few papers mention this broader question. In the context of rural India, Behrman, Foster, Rosenzweig and Vashishtha (1999) seek to disentangle the various channels through which maternal education affects child schooling, and find evidence that the better ability of educated mothers to complement schooling with home teaching is a major channel. While not comparing the magnitude of this impact with that of school quality, this paper is a rare example of an investigation into home teaching as a complement to schooling, as opposed to a (more common) mere consideration of household characteristics as education production function inputs.

In the US context, Hanushek, Rivkin and Kain (2000) remark that in their Texan sample, attending a class taught by a more efficient teacher can substantially offset the disadvantage of a deprived social background[42]. Bonesrønning (2004) estimates the interactions between class size and 'parental effort' (measured by proxy variables such as the frequency with which pupils ask their parents for assistance with homework, how often parents control homework, etc.). Using an IV technique inspired from Angrist and Lavy's (1999) use of Maimonides rule, he finds that increases in class size tend to reduce parental effort. In this case, a specific parental input (help with / control of homework) and a specific school input (class size) are found to be complements. However, serious estimation issues arise and Bonesrønning himself warns in his abstract that 'the evidence is not conclusive'.

[40] To be precise, if school and household expenditures are substitutes, anticipated increases in school expenditure in the next period result in increases in household expenditure in the current period and decreases in the next period, while unanticipated increases in school expenditure will have no impact on household expenditure. As a result, unanticipated increases should have a larger impact on cognitive achievement.

[41] A concern here is the potential endogeneity of these unanticipated funds, which could benefit schools that are already located in areas with characteristics more amenable to schooling — in which case their impact on achievement could be fallacious. Das et al. however provide some evidence that unanticipated funds are not related to observable pupil and school characteristics.

[42] See subsection 3.3 above for a discussion of that paper.

To conclude this subsection, if one retains an education production function framework, incorporating household responses into empirical modeling may be a useful addition. Increases in school inputs with which some parental inputs are substitutes may not result in achievement gains as high as expected, once household behavior is taken into account; on the contrary, complementarities may result in higher gains. This may contribute to the difficulty in estimating production functions, and may shed light as well on differences that have been observed between estimates of the results of the same randomized experiments obtained using experimental and non-experimental techniques (see Duflo, 2003, and Duflo and Kremer, 2003). Indeed, experimental estimates include household responses; non-experimental estimates do not (Todd and Wolpin, 2003; see introduction to section 4 above). A corollary is that policies might be designed that would explicitly rely on households as actors of the schooling system.

5.2. Demand-side financing as a substitute for the provision of school inputs

Policies allocating educational expenditures to families might offer an alternative to standard policies focusing on the resources of schools. Such policies may help parents to afford buying inputs such as textbooks (or paying tuition fees when they exist); they may also, especially among poor households in developing countries, reduce the time children spend working. Such 'demand-side' interventions have been promoted notably by the World Bank as a way to overcome the ineffectiveness of supply-side interventions, and evaluation reports from the Bank constitute most of the literature about them. Enrolment subsidies have also been advocated as an effective means of simultaneously reducing child labor and increasing enrolment[43].

Patrinos (2002) provides a useful review of these interventions[44]. His extensive list includes 32 World Bank programs (in Bangladesh, Colombia, Guatemala, Indonesia, Jamaica, Kenya, Mexico, among other countries), 20 non-World Bank programs in developing countries, and 17 programs in developed countries. Five types of interventions can be distinguished: 'Stipends' (cash payments to families covering schooling costs), 'targeted vouchers' (cash payments conditional on enrolment, allowing the choice of public or private schools), 'targeted bursaries' (cash payments to schools, municipalities or provinces that are reserved for specific purposes), 'student loans' for higher education and 'community grants / community financing', i.e. funds given to a community linked to attendance at a community institution. The list of expected benefits clearly shows that these programs are now seen as an alternative to providing more resources to schools and reforming their management, as it includes higher quantity and quality of education (higher enrollment, attendance and completion rates, higher achievement), efficiency gains and cost-effectiveness.

Random selection of program participants coupled with careful data collection would provide an ideal means of evaluating demand-side interventions, yet it has not been implemented in most cases, so that the literature focuses on the few cases where it has been, such as Mexico's Progresa program[45]. Nevertheless, Patrinos' account of the results clearly suggests that demand-side interventions have a great potential to increase participation, despite some glaring failures. For example, Ravallion and Wodon (2000) find that Bangladesh's Food-for-Education program, launched in 1994, covering 13 % of total enrolment in 1995-96, entitling children of poor families to 15 kg of wheat each month conditional on enrolment in primary school and 85 % attendance, increased attendance by 24 % of the maximum feasible days of schooling, and reduced child work though by a much lesser extent. An issue with this program was that once enrolled, children were still confronted to the inefficiency of the primary school system.

---

[43] There has been a renewal of the economics literature on child work, much of which is not of direct relevance here, as it does not address the relationship between child work and education, but rather focuses on the determinants of child wage labor, an extreme and *relatively* infrequent form of child work — debates revolve around whether child wage labor is caused by subsistence poverty or labor and capital market imperfections, for example, and whether there is a risk that effectively banning it would merely plunge working children into deeper poverty. Several recent papers, however, have addressed the amount of substitution that exists between work and schooling, both of which may be combined. Most of them focus on participation rather than achievement; Heady (2000), however, studies the impact of work on the achievement of Ghanaian children who participate in both school and work, and finds that it is strongly negative. Policies reducing child work, whether of not-enrolled or of enrolled children, could play an important role in increasing enrolment and achievement — improving children's welfare outside school may be another way of improving school outcomes.

[44] Rawlings and Rubio (2003) provide a more detailed review of programs implemented in Latin America, that covers both the health and education sectors and focuses on methodology issues in evaluating program impacts. See Patrinos' and their paper for further reference.

[45] See Schultz (2004), mentioned in section 4.1 above.

Indeed, the evidence on the impact of demand-side financing initiatives on achievement (rather than school participation) is rather mixed. Mexico's Progresa includes the provision of additional resources to schools, but this does not seem to have improved their functioning; the Brazilian program Bolsa Escola, that largely replicated Progresa, also did not lead to an increase in school quality. The pass rate at the secondary school certificate exam of participants in Bangladesh's Female Secondary School Assistance Project was similar to the national average. Patrinos concludes that 'in conditional cash transfer programs, there must be careful consideration of the quality of the education system if one wants the short-run stipend intervention to have a long-run impact on education outcomes' (p. 12). In the end, it comes as no surprise that programs designed to increase school participation by reducing its costs or making it otherwise attractive to parents have an impact of the quantity rather than the quality of schooling — the main objective of these programs is to reduce the negative impact of low family income on enrolment.

5.3. Insights into the political economy of educational expenditures

Optimal responses by education authorities to pupil characteristics can affect education production function estimates, just as household responses to policies do. For example, as explained by Vignoles et al. (2000): '[…] assume that a school is given a fixed budget. Assume also that the school knows that the same level of resource inputs has a very different effect on a child's attainment, depending on the socio-economic background and prior attainment of the child. The school will therefore allocate their fixed amount of resources among their students, taking this fact into account. In other words, the school will systematically allocate resources to each child, such that the learning output of the whole school is maximised. […] This criticism suggests that, contrary to Hanushek's conclusion […] that school resources do not impact greatly on pupil outcomes because schools are inefficient, resources do not appear to impact on outcomes because schools are efficiently optimizing their use of scarce resources' (p. 5)[46]. However, structural modeling of the allocation of resources within and between schools hardly seems to have been undertaken, and the assumption that schools are run to maximize pupil achievement is contentious — it underlies the production function approach but is not quite consistent with empirical results.

Indeed, as argued by Pritchett and Filmer (1999) in a widely cited paper, the actual rules which govern the allocation of educational expenditures across inputs are only secondarily concerned with achievement, if at all. Pritchett and Filmer show that available production function estimates yield insights into that *political* economy. They first remark that the marginal impact of each input depends on the rate of utilization at which it is assessed, because of diminishing returns. Systematic overprovision of an input (e.g. several copies per pupil of the same textbook, or teacher salaries uniformly higher than the corresponding market wage rate) will result in that input having no significant marginal impact, even if it is actually crucial in educational production. Now if school resources were allocated so as to maximize achievement, the marginal products per unit of expenditure of different inputs should be equalized.

Based on the estimates compiled by Fuller and Clarke (1994) and on selected individual studies, Pritchett and Filmer show that this is not the case. Indeed, there is a systematic tendency for inputs from which teachers can derive direct utility, independently of their impact on achievement, to be associated with insignificant or small coefficients. For example, Harbison and Hanushek (1992) find that, in Brazil, material inputs had a much higher impact on achievement than alternative teacher education strategies, which were themselves superior to teacher salaries. Pritchett and Filmer comment that '[…] studies that use non-experimental data systematically evaluate the marginal product of inputs at expenditure allocations that are not education output maximizing because, at the spending allocations typically observed which are the result of choice, the marginal product of inputs is lower for those inputs more highly valued by teachers and hence they are less likely to be found statistically significant' (p. 229). Indeed, '[…] the evidence is grossly inconsistent with the assumption that

[46] A limitation of Vignoles et al.'s argument is that it concerns variations in per pupil expenditure *within* schools, while production function studies typically rely on variations *between* schools or even higher levels of aggregation. What would deserve consideration, therefore, is not so much the use of available resources by school principals, but the rules that govern the allocation of resources across schools (e.g. see subsection 3.1. for the consequences of *purposive program placement* on the estimation of production functions).

resources are allocated to maximize education output (however defined)' (p. 224), so that alternative theories are required to explain why there is such overspending on teacher-related inputs.

A first hypothesis is that schools produce both achievement and 'citizenship' (defined as 'acceptable patterns of social interaction, and cultural norms and beliefs, and political ideals', p. 225), so that observed input levels are not consistent with maximization of achievement alone, but still the result of output-maximizing behavior: Inputs that are more effective in producing citizenship have higher rates of utilization and consequently lower marginal products than if achievement alone was maximized. An alternative hypothesis is that expenditures are allocated so as to optimize both educational production and teacher utility, because of the influence of teachers in determination of educational policy. Now teachers derive utility both from their students' test scores and directly from spending on inputs. An implication is that inputs that enter teacher utility directly will have higher rates of utilization and lower marginal products than would be the case under pure achievement maximization. The distortion thus created depends on the weight assigned to teacher utility versus achievement in policy making as well as on teachers' own relative valuation of achievement and of the inputs that directly enter the utility function.

Pritchett and Filmer argue that the second hypothesis is more directly consistent with the evidence than the first: 'Given the evidence, we think the correct model that describes the actual allocation of educational expenditures for most, if not all, developing countries must include the fact that educators have enormous influence over the allocation of spending and that spending is biased towards those educational inputs that also directly increase the welfare of teachers. Crudely put, teachers lobby (and form unions, and strike, and write) and books and desks do not' (p. 235). Pritchett and Filmer then propose three models of the political economy of school expenditure. The first relies on the principal-agent framework: Parents and school managers do not know the actual production function parameters, which teachers misrepresent to further their interests. Teachers overemphasize the impact of teacher-related inputs, and this results in excessive spending on them. The second is based on a standard collective-action argument: In the competition for the allocation of educational expenditure, teachers, and teacher unions in particular, have more political power than parents. The third argues that policy makers, e.g. ministers of Education use input allocation as patronage, to strengthen their political base.

A consequence is that resource-based policies not accompanied by changes in the incentives structure of the school system may not result in major improvements in outcomes, as the impact of educational policies depends on both total expenditures and their allocation over various inputs. Increases in total expenditures with constant allocation rules may even be less effective than changes in allocation rules for a constant budget. Pritchett and Filmer's analysis of the political economy of educational expenditure matches a growing interest for the incentives faced by teachers as opposed to the resources they can avail of; this recent research is examined in section 6.

6. THE RELATIONSHIP BETWEEN INCENTIVES AND EDUCATIONAL OUTCOMES

School resources are but a component of any school system, and obviously much depends on the way they are utilized. Given the evidence that available resources are very poorly utilized, especially in developing countries, altering the incentives that the system gives to teachers and its other actors appears to be a necessary complement to 'input-based policies'.

Indeed, analyzing teacher behavior and school management has become the focus of a new literature, that largely gives up estimating production functions but instead examines the impact on achievement of various institutional arrangements which differ in the incentives they give notably to teachers. Hanushek (most clearly in his 1995, 2002 b and 2003 papers) has been one of the promoters of this shift in emphasis from resources to incentives: 'In short, the findings [of education production function studies] do *not* indicate that schools and teachers are all the same. Large differences exist, even though these differences are not captured by the simple measures commonly employed. Neither, it appears, are they captured by more detailed measures of classroom organization or pedagogical approach. This leads me to conclude that the educational process is very complicated and that we do

not understand it very well. We cannot describe what makes a good or bad teacher or a good or bad school. Nor are we likely to be able to describe the educational process very well in the near future. My view is that we should learn to live with this fact: living with it implies finding policies that acknowledge and work within this fundamental ignorance' (1995, p. 236; emphasis in the original).

Once one has *renounced* understanding teaching itself, educational policy as guided by economic analysis should focus on the design of adequate incentives: 'The alternative incentive structures include a variety of conceptual approaches to providing rewards for improved student performance and range from merit pay for teachers to privatization and vouchers. Performance incentives recognize that there might be varying approaches by teachers and schools that are productive. Thus, they avoid the centralized "command and control" perspective of much current policy. At the same time, they recognize that simply decentralizing decision making is unlikely to work effectively unless there exist clear objectives and unless there is direct accountability' (2002 b, pp. 2089-90). It should be noted that Hanushek's understanding of these policies is by no means neutral politically, as what he advocates is effectively a transformation of schools into competitive firms whose only output would be cognitive achievement — an evolution that government involvement in education has hitherto prevented from taking place.

Reforms seeking to transform the incentives structure of the school system have already been implemented in a very large number of developed and developing countries, frequently under the influence of the World Bank: Changes in public-school teachers' employment conditions, administrative or political decentralization of school management, development of the private sector, etc. Much of the economics literature about these reforms pertains to the United States and is not directly relevant to developing-country contexts, e.g. the debates on school finance. Hanushek (2002 b, sections 5 to 7) provides a survey of the evidence for the US. Studies about developing countries have started accumulating over the last few years, largely as evaluations of World Bank projects. In his survey published in 2002, Glewwe mentions only a handful of them, but many more papers are now available. Indeed, while survey papers on the relationship between school *resources* and pupil achievement abound, there does not seem to exist a comprehensive survey on the relationship between *incentives* and achievement.

Given the range of issues involved, a detailed discussion of changes in the incentives structure of the school system is way beyond the scope of this review, and beside its subject. This section has a less ambitious aim, which is to mention a few key themes in this newer literature that will likely replace the education production function research program as the core of the economics of education. The discussion examines teacher incentives (6.1), decentralization (6.3) and privatization (6.4).

6.1. Teacher incentives

Pritchett and Filmer's (1999) analysis is especially relevant to the teacher variables typically included in production function studies, such as education, training and experience. Observable teacher characteristics describe the potential for effective teaching, but whether this actually translates into reality depends on the actual behavior of teachers in the classroom. For example, Hanushek, Rivkin et Kain's (2000) results (see section 3.3 above) suggest that teacher quality, while highly variable, is hardly related to these variables. In an intellectually neo-classical and politically right-wing framework, 'merit pay' appears to be a privileged way to motivate teachers, by rewarding those teachers whose contribution to their pupils achievement gains over the school year is greatest.

Hanushek (2002) briefly reviews the evidence for the United States, and concludes that attempts to introduce merit pay have not been successful. Indeed, they have yielded uniform pay structures, instead of the expected performance-wise differentiated structure, and seem to have resulted in teachers increasing teaching time rather than teaching quality. Hanushek, however, asserts that merit pay may have a greater impact through its effect on the selection of new professionals into teaching than through its effect on teachers already positioned, and that studies also taking the former into account may reach different conclusions: 'This assessment assumes that the most significant issue is whether or not teacher are trying hard to do well. If they are not, merit pay may induce more and better efforts, leading to improved student performance. An alternative view is that the most significant

aspect of merit pay proposals revolves around changes in the stock of teachers, or the selection issue. Merit pay schemes might provide incentives for better teachers to stay and for poorer teachers to leave and thus may have little to do with variations in effort' (p. 2105).

Lavy (2003) examines the introduction of financial incentives for teachers in Israeli schools, which seems to have been more successful than its US counterparts: 'The incentive program [under study] is a rank-order tournament among teachers of English, Hebrew, and mathematics. Teachers were rewarded with cash bonuses for improvements in their students' performance on high-school matriculation exams. […] teachers' monetary performance incentives a have significant effect on students' achievements in English and math. No spillover effect on untreated subject is evident and the general equilibrium impact of the program is positive as well. The program is also more cost-effective than alternative forms of intervention such as extra instruction time and is as effective as cash bonuses for students' (abstract).

There is some evidence about the relationship between teacher performance and salaries in developing countries, as well as about the introduction of merit-pay schemes. As is the case for the US, a distinction arises between the relationship of teacher pay with the selection of professionals into teaching and its relationship with the motivation of teachers once appointed.

Kingdon (1996 b) first estimates education production function for secondary schools of urban North India, and then investigates whether those characteristics of teachers that are found to have an impact on achievement are determinants of salaries. She concludes that existing remuneration schemes are not structured so as to motivate teachers towards improving their pupils' achievement: '[…] much of the educational data and debate in India has been on measures that, according to our data, are dubious indicators of school quality. Moreover, there is evidence of inefficient incentive structures for teachers, with teacher characteristics that produce improved student achievement commanding only weakly higher pay, while other teacher traits that have few discernible learning benefits for the pupils having strong salary pay-offs for the teachers' (abstract). In a paper of similar inspiration, Kingdon and Teal (2003) estimate education production functions for government and private schools of the same area and find that, given student characteristics and school resources, private schools obtain better academic results. Contrary to government schools, private schools relate teacher pay to student achievement, and this would explain their efficiency.

In a very different context, Vegas, Pritchett and Experton (1999), examine the structure of teacher pay in Argentina. They show that its uniformity results in variations in real teacher salaries across provinces, making teaching a more or less attractive occupation for highly qualified individuals. As a result, a national pay raise or a national policy on teacher compensation would not be the best way to attract and retain qualified teachers. Vegas, Pritchett and Experton also question the uniformity in the structure of rewarding seniority in teaching.

Changes in teacher pay structure, however, are not the only incentive than can affect teacher behavior. Lopez-Acevedo (2004) provides some evidence as to the impact of a comprehensive program launched in Mexico in 1992, Carrera Magisterial, designed to raise the quality of education. The program consists in providing additional teacher training as well as improving working and salary conditions. Lopez-Acevedo finds that teacher participation in the program has a positive impact on pupil achievement, especially when the program is targeted toward increasing teachers' practical experience and developing content-specific knowledge, and when the degree of supervision by school principals is high. These results echo those of Angrist and Lavy (2001) mentioned in subsection 4.2 above, and suggest a complex relationship between school resources (in this case, teacher training or, more generally, 'investment in teachers'), monetary incentives and teacher motivation, that goes beyond the simplistic rhetoric of merit pay. Hanushek, in various papers, acknowledges that designing adequate teacher incentives is difficult, but his renouncement to understanding actual classroom practice makes the task even more arduous.

Merit pay can also have perverse side effects, as found by Glewwe, Ilias and Kremer (2002), in an experimental study of Kenya. Prizes were to be attributed by parent-run school committees to teachers whose pupils would have had low dropout rates and would have performed well on exams. The impact

on achievement was limited as teachers responded to the incentive by manipulating exam results rather than teaching better. Establishing a direct relationship between remuneration and a simplistic and easy to manipulate outcome such as test scores is bound to have such perverse consequences, especially in contexts where schools are deprived of resources, other factors than salaries de-motivate teachers, and achievement levels officially required to pass test scores are unrealistically high — as is the case in many developing countries.

6.2. Decentralization

Decentralization has often been associated with changes in teacher recruitment, employment and remuneration conditions. A difficulty with this concept is that it has been used to designate a range of very different policies, as highlighted by Pritchett and Filmer (1999): 'In implicit recognition of the incentive problems inherent in centrally controlled public provision of schooling there have been a number of moves to increase local control over schools. These range from "decentralization" type reforms which shift the government authority with control over the provision of schooling; from federal to state (provincial) or state (provincial) to municipal, to "localization" or "school autonomy" initiatives to move more control over schools from whatever levels to the schools themselves' (p. 235). Indeed, while economics has long addressed what Pritchett and Filmer specifically call 'decentralization' (e.g. the public economics literature on fiscal federalism), the new focus of interest is rather 'localization', e.g. the local recruitment of teachers and various forms of 'community participation'[47].

While there is a US literature on issues related to decentralization, e.g. school finance reform (see Hanushek, 2002 b), most of the recent evidence comes from developing countries. There is some evidence that *local funding* of schools permits greater accountability in resource allocation and use. Jimenez and Paqueo (1996) thus find that, in the Philippines, local financing (contributions from the local school board, municipal government, parent-teacher associations, etc.) dramatically increases achievement per peso, mostly by reducing expenditure on personnel. A 1 % increase in the share of financing coming from local sources would be associated with a decline in total costs of 0,135 %, or about the cost of providing for a place for one more student. James, King and Suryadi (1996) similarly find a positive relationship between the share of locally-raised funds and achievement in both public and private schools in Indonesia.

The economics literature on *school autonomy* is much more recent; Glewwe (2002) mentions only two papers. Jimenez and Sawada (1999) estimate the impact of El Salvador's EDUCO program, under which schools are run by parent committees that can purchase school equipment and recruit and dismiss teachers. Achievement regressions for a sample of pupils attending EDUCO and other public schools show that EDUCO school pupils outperform others in terms of attendance (by 3 to 4 days in the past four weeks) and reading skills (by 1.3 standard deviations). A potential limitation is that the implementation of the program was not randomized, so that the authors have to rely on standard selection correction techniques, using debatable identification assumptions. The results nevertheless suggest that decentralized management succeeded in increasing the accountability of EDUCO schools to the local community. King and Ozler (2000) study 'Autonomous' schools in Nicaragua, which are managed by 'Directive councils' comprising the school principal, teachers, parents and students) and are competent to select textbooks, set school fees and recruit and dismiss the school principal — all these competencies belong to central authorities in regular government schools. Using a sample of about 3,000 primary and secondary school students enrolled in a total of 153 autonomous and 89 regular schools, King and Ozler harly find evidence that autonomous schools reach higher achievement levels (the only significant impact, at the 10 % level, is on mathematics scores in primary schools), and face methodological difficulties in controlling for the potential endogeneity of the autonomy variable. Glewwe (2002) concludes that 'overall, the results of both studies of autonomy are intriguing and intuitively plausible, but more and better research is needed before making policy recommendations' (p. 464).

[47] See Bardhan (2002) for a general survey of these new forms of decentralization.

Like merit pay, decentralization can have perverse side-effects. In an important paper, Kremer, Moulin and Namunyu (2003) analyze the situation of Kenya, providing 'a cautionary tale': 'Kenya's education system blends substantial centralization with elements of local control and school choice. […] the system creates incentives for local communities to build too many small schools; to spend too much on teachers relative to non-teacher inputs; and to set school fees that exceed those preferred by the median voter and prevent many children from attending school. Moreover, the system renders the incentive effects of school choice counterproductive by undermining the tendency for pupils to switch into the schools with the best headmasters. A randomized evaluation of a program operated by a non-profit organization suggests that budget-neutral reductions in the cost of attending school and increases in nonteacher inputs, financed by increases in class size, would greatly reduce dropout rates without reducing test scores. Moreover, evidence based on transfers into and out of program schools suggests that the population would prefer such a reallocation of expenditures' (abstract).

Decentralization per se does not necessarily transform the incentives structure of the school system in a way that leads to improved achievement: The question is to what extent it affects classroom practice.

6.3. Privatization

The privatization of education is a natural proposition for neo-classical economists. Indeed, neo-classical theory provides little rationale for governments to run most schools directly, as is the case in most countries in the world. Externalities and other market failures would justify subsidies to education, but not direct provision. According to Hanushek (2002 b): 'Because of the heavy involvement of the public sector in the actual provision of schooling, understanding the efficiency of production becomes an important issue. With competitive, private provision, little attention is given to economic or technical efficiency. Barring obvious market imperfections, there is general *faith* that market forces will push firms toward efficient use of resources. Even with market imperfections, there is generally little attention given to issues of technical efficiency, because firms are *presumed* to produce the highest possible levels of output given the chosen inputs. […] But, the involvement of government in production, frequently in near-monopoly situations, alters the focus considerably. The possibility of inefficient production becomes a much more serious concern.' (p. 2068; emphasis added). There is much faith and presumption indeed in this vision of education, but privatization has become a major *de facto* trend in many developed or developing countries[48] as well as the focus of heated controversies in academic and policy-making circles.

In the United States, most of the debate has revolved around the benefits of competition between schools located in the same area, and of the possibility for parents to choose the school in which they enroll their children. In a market economy, indeed, school choice is a form of *consumer power* that could play an essential role in shaping a more efficient supply of education. However, according to Hanushek (2002 b), the evidence is mixed. Several papers have shown that there is a positive relationship between school efficiency and competition, as measured by the density of neighboring school districts in urban areas (Hoxby, 2000 b, Hanushek and Rivkin, 2003). However, the investigation of the relative efficiency of public and private schools has proved difficult and controversial. The usual approach consists in estimating production functions for each sector separately, while controlling for the selection of students into either. Most of this research has focused on Catholic schools, which represent a high but quickly decreasing proportion of the relatively small private school system of the United States (about 11 % of all pupils attend private primary or secondary schools). Catholic schools have often been shown to achieve higher graduation rates, after controlling for selection effects and school resources, but the results for test scores are unclear. Estimated coefficients differs widely across studies in magnitude, and they are often insignificant (Evans and Schwab, 1995; Neal, 1997). Interestingly, public and private schools located in wealthy suburban neighborhoods tend to be similarly efficient, while private schools outperform their public counterparts in poorer urban areas — which suggests that, besides organizational differences, some contextual, sociological factors may be at play.

[48] For example, in India, most of the remarkable increase in elementary school enrollment recorded during the 1990s has been the result of the spread of private schooling (see Kingdon, 1996 c, 2002).

Much of the recent literature however revolves around the introduction of a specific type of 'school choice', namely the possibility for parents to enroll their children in any public or private schools, for which purpose they receive an education voucher covering tuition and other fees — schools are financed through the vouchers and thus they should be incited to maintain high quality to attract and retain students. The proposition dates back to the writings of Milton Friedman (1962), but till date has rarely been implemented — although it was part of the Republican candidate program for the 2000 presidential election. Theoretical models have been developed in which a school system based on voucher-financed private schools turn out to be optimal compared to public provision of schooling, but the risk of segregation of students according to family background (race, class, income level, etc.) has been a major concern, especially for the opponents to the proposal.

The empirical evidence comes mostly from a program implemented in Milwaukee, Wisconsin, which is not representative of what vouchers could be if they were made into a general policy, as stressed by Hanushek (2002 b): 'This publicly supported voucher, while not an experiment but instead an on-going program, had an evaluation plan set up at the outset […] Subsequently, a number of studies, taking different approaches and reaching somewhat different conclusions, have looked a the same impact of vouchers in Milwaukee' (p. 2112). While parents who had chosen voucher schools appeared to be satisfied with their choice, 'the analyses of student achievement show no real gains in vouchers schools during the first years, but by the fourth year of operation voucher schools are doing as well as or better than the Milwaukee public schools, depending upon the precise performance measure. The findings on achievement also depend upon the precise comparison groups' (p. 2112). In fact, although the literature on voucher schemes is already sizeable (e.g. Neal, 2002; Hoxby, 2003; Ladd, 2002), 'the existing public school choice plans have not received thorough analysis in terms of student outcomes' (p. 2112).

Glewwe (2002) briefly surveys the evidence about privatization in developing countries. In this literature as well, many studies have attempted to compare the efficiency of public and private schools. Experiments randomly assigning pupils to either sector would provide an ideal estimation framework, but it seems that no such experiment has been conducted as yet. Most studies in fact estimate production functions on observational data and either pool samples of public- and private-school pupils and introduce a private-school indicator variable, or run separate regressions. The latter approach is preferable, as it allows for the 'technologies' used in the two sectors to be actually different (besides differences in efficiency per se). The major issue is controlling for selection, as it has proved difficult to identify variables which affect school choice but not achievement.

Cox and Jimenez (1991) compare public and private secondary schools in Colombia and Tanzania, using family background variables as determinants of selection, which is an issue, for they are bound to also affect achievement. Nevertheless, private schools are found to be more efficient than public ones. Kingdon (1996 a) compares government, private 'aided' and private 'unaided' secondary schools in urban North India. She uses a sample of 928 students enrolled in 30 schools including rare data on per-pupil costs for each school. Her results imply that the ratio of per-pupil costs on predicted test scores for an average student is half as high in private unaided schools as it is in government and private aided schools — test score production is less costly in private unaided schools. In this study as well, however, the choice of variables used to control for selection is problematic.

Vouchers have started being implemented in developing countries as well, and have attracted their fair share of controversy (see West, 1997, and Carnoy, 1997, for a useful introductory debate). The first major voucher scheme was introduced in Chile in the early 1980s by the military dictatorship. The proportion of pupils enrolled in primary and secondary private schools rose from 22 % in 1981 to 33 % in 1990. In a recent paper, Hsieh and Urquiola (2003) use panel data on about 150 municipalities to assess the program. They 'find no evidence that choice improved average educational outcomes as measured by test scores, repetition rates and years of schooling.' Meanwhile, the program 'led to increased sorting, as the "best" public school students left for the private sector.'(abstract). These results confirm those of earlier studies mentioned by Glewwe (2002).

Angrist, Bettinger, Bloom, King and Kremer (2001) use a natural experiment to assess Colombia's voucher program. From 1992 to 1997, 125,000 vouchers were offered to the residents of poor urban

neighborhoods, allowing them to attend private secondary schools offered. There was excess demand for those vouchers, which were finally allocated by lottery — generating randomness in school choice that the authors use to create an instrumental variable. Using data on 1,600 lottery participants, half of whom won, Angrist et al. find that lottery winners completed more grades of schooling due to reduced grade repetition, and that, among a sub-sample of 283 students, lottery winners scored between 0.13 and 0.20 s.d. higher in mathematics, reading and writing tests, although only the impact on reading was significant, and at the 10 % level — probably owing to small sample size.

To conclude this section, it is worth stating that while both sufficient resources and adequate incentives are necessary for reaching high school quality, they may not be sufficient, for they directly address neither teaching itself nor the politics of the relationships between teachers, pupils, parents, school managers, education policy-makers and society at large. Likening the school system to a market for cognitive skills that could be governed by perfect competition is but a specific perspective on the nature of teaching and the politics that surrounds it. Interestingly enough, the World Bank's 2004 *World Development Report*, officially acknowledges this, in a text contributed by Pritchett (2003): 'That public provision has often failed to create universally available and effective schooling does not imply that the solution is a radically different approach (complete decentralization, total control by parent groups, generalized choice) or a narrow focus on proximate determinants (more textbooks, more teacher training). […] Classroom practice is what matters. If the underlying causes of failure are not addressed, all these approaches can fail' (p. 113). This is in stark contrast with the Hanushek approach that renounces understanding classroom practice. Now actually understanding it requires realistic modeling of the behavior of pupils and teachers, and of the interactions between them. The next section turns to recent papers that may provide a basis for such modeling.

7. TOWARDS A BEHAVIORAL ECONOMICS APPROACH TO SCHOOLING

The production function for cognitive achievement is at the core of the economics of education, as the main tool that has been used to assess the relationship between school expenditure and outcomes. Dissatisfaction with the methodology and results of macro and micro studies has led to the development of an alternative approach: The experimental evaluation of policy interventions. Yet 'experimentalists' are as concerned with efficiency in the production of cognitive achievement as 'structuralists', and one of their arguments in favor of randomized trials is that they could allow to recover structural parameters (e.g. Kremer, 2003). Even the shift from resources to incentives as the main policy tool to improve cognitive achievement does not imply a major change in the representation of the school system by economists. Studies of teacher incentives, decentralization and privatization are often based on comparisons between production function estimates, e.g. for centralized versus decentralized schools or public versus private schools. The lack of consensus mentioned in section I applies to the results of existing studies and possible methodological improvements, but the research papers discussed in sections 2 through 6 in general accept the analogy between education and production, even if implicitly and/or reluctantly.

7.1. Questioning the analogy between education and production

Questioning this analogy is the purpose of this section. Education production function studies envisage schooling as the production of cognitive achievement by teachers, using a variety of inputs, including pupils. This framework greatly simplifies the design of educational policy, which is reduced to a technical issue, i.e. designing adequate financial incentives for teachers and providing optimal input levels given the estimated marginal product and the price of each input. This view underlies Glewwe's (2002) survey: 'The question addressed is: What school policies are most cost-effective in producing students with particular cognitive skills, such as literacy and numeracy?' (p. 437). '*Education production functions […] contain most of the information that a ministry of education wants to know*. These functions are technological relationships that show how much students learn when placed in certain types of schools with certain types of teachers (conditional on student and household characteristics). Education planners can use this information to assess the impact of each school and teacher characteristic on learning. Combined with cost data on these characteristics, they can "design" schools to maximize learning per dollar spent' (p. 450; emphasis added). This framework has been

widely used empirically with practically no specific theoretical modeling being undertaken, yet serious doubts can be raised about its relevance. One may argue that the fallacy of the production function metaphor — rather than deficient empirical methods — is the main reason why economic research on the determinants of educational outcomes has been so unsuccessful, especially when compared to the parallel literature on the returns to education.

First, educational outcomes include much more than cognitive achievement as measured through standardized tests, and schools may not be managed with achievement maximization as their only or even their main purpose. According to Hoxby (2000 a), 'on the economic front, class size is a primary example of the education production function fallacy. It is conventional to estimate the relationship between educational inputs (like class size) and outputs (achievement) and to call the relationship an "education production function". This nomenclature suggests that inputs translate systematically into achievement, as they do in the production function of profit-maximizing firms. The analogy is a false one, however, because firms' production functions are not just a result of their ability to turn inputs into outputs. A firm's production function is the result of maximizing an objective (profits), given a production possibilities set. It is not obvious that schools have stringent achievement maximization objectives imposed on them. […] class size reduction can fulfill a variety of objectives, not all of which are related to achievement' (p. 1240).

Second, teachers are modeled as employees, *not* as teachers. Indeed, teacher behavior as envisaged in the literature on incentives includes only responses to productivity incentives given by educational authorities; the quality of teaching is assumed not to depend on interactions with pupils. Yet pupils respond to teacher behavior, and conversely, and it is plausible that such interactions have more impact on pupils' cognitive achievement than both resources and monetary incentives — which can explain, for example, why Rivkin, Hanushek, and Kain (2002) find that teacher quality is extremely variable but difficult to explain.

Third, the production function framework comprehensively neglects pupil behavior, which is simply not modeled in any of the papers discussed hitherto[49]. Now cognitive achievement, like any other component of human capital, is *embodied* in pupils. Pupil behavior is key to the learning process, whichever resources teachers avail of or incentives they face; it is also amenable to policy interventions, which can make considerable differences whether at an aggregate (e.g. nation-wide changes in pedagogy) or individual level (e.g. a teacher may not have the same attitude towards all his pupils).

Using the education production function conceptual framework to guide educational policy thus amounts to making stringent assumptions, which are most unlikely to hold empirically. One may argue that these assumptions are benign, i.e. they merely simplify the analysis without biasing its results. My contention is that this is not the case. A firm may be thought of as comprising employers and employees; a school system, besides school management authorities and teachers, comprises pupils, who are just not comparable to, say, a car being assembled in a factory. If a car chassis could refuse to get assembled, or be reshaped through interactions with other car chassis being assembled in the same or another factory, or deliberately hit factory workers, or leave the production line in the afternoon to return on the next morning partly assembled according to different plans than the car designers', no production manager — or economist studying the production process — would fail to take it into account. Yet these are just a few examples of pupil behavior, which economists of education do *not* take into account. Similar arguments could apply to the neglect of teacher behavior and of the other outcomes of school education than cognitive achievement.

To what extent can economists improve their modeling of schools? Learning is not an economic process, and it is unlikely that pupil behavior or interactions between teachers and pupils are essentially economic (but for teachers' responses to financial incentives). Economists may thus want to leave these topics for other social scientists to investigate, e.g. cognitive scientists, educational scientists, social psychologists or sociologists, and provide enough experimental evidence on the impact of policy interventions to guide educational policy. In this case, they could still compare the

[49] It is discussed, though informally, in some experimental studies, notably Banerjee, Cole, Duflo and Linden (2003).

cost-efficiency of the alternative interventions in producing cognitive achievement, but they could not quite say through which channels these interventions have an impact. A school would still be considered a black box; factors which make an experiment replicable or not would not be spelt out if they influenced pupil behavior or pupil-teacher interactions rather than available resources or teacher incentives.

Reaching more general conclusions about the organization of the school system and the design of alternative policies requires taking other considerations into account, which can be done only if economists make use of insights gained from the other social sciences. This amounts to shifting to 'behavioral' economics, i.e. recognizing that education, before being a series of outcomes, is a process which takes place in social institutions that have a history, and basing its representation through models on a psychological understanding of pupil and teacher behavior (the case for behavioral economics is eloquently put by Bowles and Gintis, 2000). The growing criticism of the production function approach over the last few years has yet to lead to the development of such a consistent research program; not all of the critics are behavioral economists. Yet, taken as a whole, the papers introduced below may provide some basis for a more realistic conceptual framework and renewed empirical analysis.

The rest of this section develops these arguments. It presents evidence that the other outcomes of school education than cognitive achievement matter (7.2), and introduces recent attempts to model teacher and pupil behavior (7.3).

7.2. Taking the diversity of educational outcomes into account

This paper started with a comparison between the education production function and returns to education literatures (1), stating that the former had been much more successful in establishing a substantial body of evidence. It may be useful to build again here on a comparison between the two literatures in terms of the outcomes of school education they have considered. Education production function studies have focused on cognitive achievement as measured by test scores in language, mathematics, and often science — indeed the quality of the education a person has received is *defined* as the level of her test scores. Authors attempting to measure the returns to education have long focused on the quantity of education (the number of years of schooling), but the way its impact on earnings should be interpreted (causal impact of human capital, screening or signaling effects, correlation between education and earnings both determined by unobserved innate ability, rewards given to educational credentials, etc.) has been the focus of longstanding debates, and many recent papers have attempted to measure the returns to cognitive skills per se.

Now both the quantity and the quality of education appear to influence earnings, which could imply that cognitive skills are not the only outcome of school education that is valued on the labor market. Meanwhile, authors who have investigated the impact of school characteristics on earnings (rather than test scores) have typically found it easier to obtain significant coefficients of the expected sign: School resources may fail to influence test scores but still have a positive impact on earnings as they enhance the acquisition of other, non-cognitive skills that also enter wage formation. This offers a broad perspective, as non-cognitive skills may include any kind of psychological or behavioral trait that is rewarded on the labor market. Naturally, the traits valued by governments and those valued by the market may not be the same, nor may they be the same as those valued by pupils or their parents; not all of them may be valuable on ethical grounds, e.g. submission to authority.

There is growing evidence that 'non-cognitive skills' or 'psychological and behavioral traits', as they are alternatively called by different authors, explain a large part of the variance in earnings (see Bowles, Gintis and Osborne, 2001 b, for a general survey). In particular, there has been a debate in the United States about the returns to the GED, an examination open to high school dropouts which certifies that they have the same cognitive skills as high school graduates. Heckman and various co-authors (Cameron and Heckman, 1993, Heckman and Rubinstein, 2001; on the GED, see also Tyler, Murnane and Willett, 2000, and Clark and Jaeger, 2002) have shown that the earnings of GED holders were more similar to those of other high school dropouts, than to those of high school graduates. The GED acts as a 'mixed signal' of both strong cognitive skills and weak non-cognitive skills, and hence

is less rewarded than a proper high school degree: 'GED's are "wiseguys", who lack the abilities to think ahead, to persist in tasks, or to adapt to their environments' (Heckman and Rubinstein, 2001, p. 146)[50]. Carneiro and Heckman (2003) conclude that designing policies like the GED which take *only* cognitive skills into account is 'a folly'.

The recognition that psychological and behavioral traits acquired through schooling matter has major implications for the analysis and design of educational policies. If schools produced only cognitive skills, i.e. a component of human capital, there would hardly be a justification for government provision of education; not only would educational policy be reduced to equalizing the marginal products of various inputs, but this equalization would be arrived at more efficiently by private firms than governments. As Kremer and Sarychev (2000) put it, 'the preponderance of public education is a *mystery*, since there are fairly strong a priori reasons to believe that the quality of education would be higher under a voucher system, and the limited available empirical evidence supports this view […]. Standard rationales for public support of education — positive externalities from education and credit constraints that prevent human capital investment — do not explain why the state operates schools, rather than simply financing schools through vouchers' (pp. 3-4; emphasis added). Glewwe (2000) similarly regrets that '[…] many governments favor public schools for a variety of "noneconomic" reasons (examples are perceived equity benefits and political objectives such as promoting a curriculum that gives students a national, as opposed to an ethnic or regional, identity) and thus *policy advisors have little choice but to accept this constraint and focus on ways to improve public schools*.' (p. 437; emphasis added). The same view pervades Hanushek's writings: 'With few exceptions, little policy attention is given to any underlying consideration of the scope and form of governmental intervention. […] The summary from considering the role of government is that the arguments for the currently large intervention — one quite generally including both financing and provision of services — remain not well analyzed. Thus, the remainder of this essay addresses a more limited issue: *How well does government do at what it is trying to do*.' (2002 b, pp. 2067-8; emphasis added).

Taking other outcomes of education than cognitive skills into account is thus necessary to reconcile economic theory with the fact that most schools in most countries in the world are government-operated, a point forcefully made in a series of papers by Pritchett (2002 a, b, c), who focuses on the inculcation of 'beliefs' as a component of school education, in addition to the teaching of cognitive skills. Governments naturally have a direct interest in controlling belief formation through school education, and that control could hardly be exerted through contracts with private firms — contrary to the production of cognitive skills. This explains why all governments with a strong ideological orientation (including democratic ones) or otherwise authoritarian governments have developed public education systems, and may be satisfied with them covering only privileged sections of society or being completely inefficient in imparting cognitive skills, as long as they influence belief formation in the circles which constitute their political base.

Pritchett's model is but a recognition that the history of the expansion of school education and of its control by governments predates contemporary mainstream economics and that school systems were created for reasons which hardly included the production of human capital envisaged as such. Educational policies around the world are still determined according to political criteria which have little to do with any 'scientific' evidence that may be obtained through the estimation of education production functions, a point most clearly made by Pritchett and Filmer (1999): 'Even though, as in all fields that involve human behavior, there are enormous areas of ignorance, the fundamental problem is typically not teacher or technocrat ignorance that could be resolved by more research' (p. 234). 'This implies that perhaps the main role of the estimation of educational production functions is not to inform an idealized output maximizing decision maker of the "true" technical production function and "optimal" composition of inputs, but rather to provide the information necessary to encourage the deeper educational reforms that change the very structure of decision making power' (p. 236).

For the time being, there does not seem to be much research in economics on the formation of non-cognitive skills through school education. An exception is Kremer and Sarychev (2000), who develop a theoretical model in which privatization through education vouchers may not be desirable even

[50] For a longer discussion of this research, see Baudelot and Leclercq (forthcoming).

though private schools are more efficient at producing achievement: '[…] democratic societies will prefer that education be publicly provided, rather than simply publicly financed, since under a voucher system, parents may send their children to schools teaching ideologies that are similar to their own. Over a series of generations, this leads to a more ideologically polarized society. There may be two steady states: one in which an ideologically homogeneous society votes for a public education system, thus maintaining ideological homogeneity, and another in which a more ideologically diverse society votes for school choice, thus maintaining ideological diversity' (abstract).

Proponents of vouchers could counter-argue that the possibility for parents to choose the ideology underlying their children's education is a fundamental right. In any case, a balance should be achieved between providing all citizens some common cultural ground and preserving cultural diversity within society. In Kremer and Sarychev's model, a totalitarian regime would also 'prefer that education be publicly provided'. Democratic societies, however, are able to debate about the 'ideology' promoted by the school system. For example, debates about pedagogy, the notion of authority or the role of religious institutions in education have been concerned as much with non-cognitive as with cognitive skills.

Economists need to take the acquisition of psychological and behavioral traits through schooling into account if they are to provide better guidance to educational policy. This will require building on insights gained from the other social sciences. Indeed, research on non-cognitive skills has built on the landmark book by Bowles and Gintis, *Schooling in Capitalist America* (1976; also see Bowles and Gintis, 2001, and Bowles, Gintis and Osborne, 2001a) that belonged to the then burgeoning field of 'radical economics', was close to the preoccupations of historians of education in the same decade and has had much influence on social psychologists[51]. Adopting a multidisciplinary approach seems necessary as well to arrive at a better understanding of teacher and pupil behaviour.

7.3. Understanding pupil and teacher behavior

Economists have modeled the behavior of teachers as employees who may respond to financial incentives designed to increase pupil achievement. The evidence on the implementation of such incentives programs however is mixed, and it is quite likely that non-economic aspects of teacher behavior, and in particular their interactions with pupils, interfere with them. For example, papers by Hanushek and Rivkin (2003), Hanushek, Kain and Rivkin (1999), Hanushek, Kain and Rivkin (2004), and Jepsen and Rivkin (2002) on the labor market of US teachers find that the relationship between teacher salaries and performance is weak and that student characteristics have a much stronger impact on teachers' decisions to change schools than salary differentials.

There does not seem to have been any in-depth research by economists on the behaviour of teachers *as teachers*, nor much use of results from the other social sciences in conjunction with economic studies of the impact of resources on achievement. One exception is a paper by Goldhaber and Brewer (1997) on which Vignoles et al. (2000) thus comment: 'Their paper is important because it evaluated the effect of both observed and unobserved teacher and school characteristics on students' 10^{th} grade mathematics scores. They found that teacher behaviors and techniques may be more important than simple resource measures. Specifically, teachers who felt well prepared, who had control over lesson content, who spent less time maintaining order, and who used oral questions frequently and emphasized problem solving had a positive effect on pupils. The causality and endogeneity of some of these behavioral variables is questionable, but these results do suggest that researchers might need to focus more on qualitative aspects of teachers and schooling inputs' (p. 32). Specifically, this paper suggests that it is possible, even using standard quantitative techniques, to reverse Hanushek's conclusion that teaching should in the end be treated as a black box.

There has been even less research on student behavior. Students are systematically represented as passive recipients of knowledge, and are assumed to play no active part in learning. For example,

[51] *Schooling in Capitalist America* insisted on the 'correspondence' between the socialization at school, characterized by hierarchy, and at the workplace, and argued that the American school system had been the locus of conflicts of interests that had resulted in it being reshaped according to the needs of employers.

Mocan, Scafidi and Tekin's (2002) comparison of the behavior of adolescents enrolled in US Catholic and public schools assumes that the stricter discipline maintained in Catholic schools should automatically reduce the incidence of 'bad behavior', without even considering the possibility that adolescents may *respond* to discipline by even 'worse' behavior. Lazear's (2001) widely cited attempt to derive a theoretical model of the relationship between class size and achievement, considers that each student has an *exogenous* propensity to cause disruption and does not attempt to understand what determines this propensity — schools adapt class size so as to minimize disruption, the only alternative being, here as well, stricter 'discipline'.

In the end, what is needed is a new economic theory of students, teachers and schools, and a recent paper by Akerlof and Kranton (2002), based on results from the sociology of education, might provide an avenue for such research. Akerlof and Kranton's explicit aim is to improve on the ability of economics to explain the apparent lack of a relationship between school expenditure and outcomes: 'While economists have begun to examine many substantive areas of the sociology of education, our reading of the literature indicates that two related themes have thus far eluded economic analysis. First is a sociological view of *the student as the primary decision maker*. Second is the conception of the school as a social institution' (p. 1172; emphasis added). The lack of consideration of these two themes explains the failure of economists to provide reliable evidence about the functioning of schools: 'Without a model that mirrors this sociology, economic analysis produces only partial answers to key questions. For example, what is the impact of resources on academic attainment? What are the important elements of school reform? Economists can answer *whether* additional resources enhance schooling, but not *when* these resources will be effective, or *why*' (p. 1168; emphasis in the original).

Akerlof and Kranton then question the emphasis given to outcomes in the economics literature, and build on the idea that processes are as important and that achievement does not necessarily play a central part in students' experience of schooling. They '[…] develop a theory where a student's primary motivation is his or her identity and the quality of a school depends on how studies fit in a school's social setting' (p. 1167).

The specificity of their approach is not so much the consideration of the multiplicity of education outcomes as the recognition that students react to the school environment in which they live, and that this behavior can be understood in sociological terms, that it also has a history: 'schools not only impart skills; they impart the characteristics and behavior of ideal students. […] students' identities delineate whether a student accepts or rejects the school itself. […] Why would students reject their school? A large part of the sociology of schooling points to systematic social differences between students and schools as a reason. Histories of urban education, for example, emphasize the clash between the Americanizing schools of the early twentieth century and their immigrant students. Students were not just taught reading, writing, and arithmetic. In school, they were also corrected in details of comportment, including what to wear and how to speak. The teachers believed they were teaching the students what they needed to be economically successful. The skills were socially neutral — just logic and reason. But that is not how it seemed to all students. […] even the most caring teachers can unknowingly offend their students and convey that they are inferior' (p. 1181).

Obviously the contemporary mainstream economic vision of schools as firms producing cognitive skills is not any more neutral: '[economists'] models assume that schools maximize students' marketable skills. But skills are only one of the goals of some schools and of the parents who choose them. […] The goals and curricula of public schools are the product of elected school boards; the nature of these schools, their ideals, may then derive from the political economy of a community. If schools' goals include promoting certain ideologies, school choice may be neither skill-increasing nor ideologically neutral' (p. 1199). Designing policies aimed to increase skills thus requires also considering the other dimensions of schooling *as students and teachers experience it*, which cannot be done if schooling is considered a technical process in which subjective experience does not matter.

Akerlof and Kranton apply their conceptual framework to the history of US high schools from the 1950s to the 1990s, of which they give a very different picture than that of the decline in productivity emphasized by mainstream economists: 'A sequence of three models will capture the historical

development of U.S. schools in the twentieth century. The first model captures historians' depiction of the initial period, in the early twentieth century: schools have a single ideal, which students may accept or reject. The second model conforms to the sociological picture of contemporary public education: schools adjust to their diverse student bodies by allowing students to choose among different ideals within the same school. The third model captures the nature of reform programs: schools reduce the social difference between students and their schools; students then identify with these schools, which promote high academic achievement' (p. 1182).

Now the apparent lack of school expenditures and educational outcomes can be easily explained, and it is clear that both resource-based and incentive-based policies may not improve the situation, unless student and teacher behavior is addressed: '[…] the typical U.S. high school today is "Shopping Mall High". In the language of our framework, such schools fail to promote a particular social category. Instead, principals and teachers preach tolerance, and students are allowed wide choice in classes and curricula. Educators say that such laxity produces academic mediocrity' (p. 1169). To the contrary, '[…] private school administrators and teachers spend considerable resources to delineate prescriptions for student behavior and ensure that students identify with the school and its ideals. Successful experiments in public school reform […] have similar strategies. We see these schools, in our economic terminology, as *investing* in students' self-images and relationship with the school. The schools reduce the initial social differences among the students and create a community, with an ideal of academic excellence' (p. 1170; emphasis in the original).

In the end, the use of school resources may be made more efficient by recognizing that not all resources should be devoted to the production of cognitive skills: 'We interpret the creation of a school community as a choice about the allocation of resources within a school. […] Suppose a school has resources […] which it can divide into two uses, those directly devoted to the teaching of skills […] with the remainder devoted to creating community and reducing […] the social differences between the students and the school (p. 1188).

To conclude this section, it is worth emphasizing that Akerlof and Kranton's paper lays the ground for empirical research that could overcome many of the deficiencies of the existing literature, e.g. its inability to explain the 'productivity decline' in North American and West European schools over the last few decades or its lack of success in providing credible guidance for educational policy in developing countries. The education production function approach, which informs most of the literature, is intrinsically normative, as it reckons that efficiency in the production of cognitive skills is the only relevant criterion for educational policy. The existing literature thus fails to *describe* the way schools around the world actually function, and this is why it is so unsuccessful in *prescribing* how they should function, besides *beliefs* in the (in)effectiveness of, say, class size reduction or privatization.

Economists studying the determinants of educational outcomes venture far from their core field, and need to alter their models considerably if they want them to bear a sufficient relationship with reality for their conclusions to be policy-relevant. My opinion is that no less a change is required here than a shift in paradigm, from 'Walrasian' to 'behavioral' economics, as advocated by Bowles and Gintis (2000). Reluctance to operate such a shift probably explains why, even though it was published in a leading journal and one of its authors received the Nobel prize, Akerlof and Kranton's paper is hardly mentioned in the rest of the literature. Whether or not one is willing to include psychological or sociological considerations into economic modeling, there should be more debates on the adequacy of the mainstream conceptual framework.

8. CONCLUSION

This paper has reviewed the extensive literature in economics that has analyzed the relationship between educational expenditures and outcomes. The empirical results of that literature may be summarized as follows.

No uncontroversial stylized fact pertaining to the relative impacts of inputs like school buildings and equipment, teaching/learning materials or teacher positions, salaries and training has been identified. The attempt to estimate structural parameters of the 'production' of cognitive achievement has, by and large, failed, in stark contrast with the parallel research program that has investigated the returns to the quantity and quality of education in terms of individual earnings. Data, specification and estimation issues certainly explain this failure to a certain extent, but it seems equally plausible that there is no linear relationship between the inputs and outcomes of school education, either because school are 'inefficient', or because other factors not taken into account blur the relationship.

A few examples of successful policy interventions have been identified using experimental techniques, e.g. class size reduction in some cases (but not others) or remedial education by more child-friendly teachers. The econometric analysis of randomized or natural experiments can thus identify interventions that are cost-efficient, but its conclusions are bound to be context-specific and not enough studies have been conducted as yet for any regularity to emerge. Besides, educational policy design also takes other considerations into account than 'maximizing learning per dollar spent'; both the relevance of experiments pertaining to the determinants of cognitive achievement and the need for putting them into perspective with other dimensions of the schooling process and the political economy of education should be emphasized.

Similar conclusions apply to the literature on incentives and outcomes; it is doubtful as yet that changes in teacher incentives may become a much more powerful and simpler tool to improve cognitive achievement than increases in educational inputs have been.

The reasons why economics in the end has produced only limited evidence as to the way schools around the world function probably stems from the deficiencies of the conceptual framework it has used; the main conclusions of this review thus pertain more to economics as such than they do to educational policy. The empirical analysis of the relationship between educational expenditures and outcomes relies on the estimation of an equation such as:

$$y = f(X, \boldsymbol{W}, \boldsymbol{Z}) + u$$

where y, the explained variable, is an educational outcome, and X, the explanatory variable, is educational expenditures or a specific input; $\boldsymbol{W}$ and $\boldsymbol{Z}$ are vectors control variables including school and family characteristics, respectively, and u is a statistical residual that represents, among other things, unobserved determinants of the outcome under consideration.

While it is always possible to estimate such an equation and consider that its coefficients have descriptive value, i.e. they represent the *correlation* between each right-hand side variable and y, the interpretation of the coefficient on X as measuring its *causal impact* on y requires the adoption of some theoretical framework and empirical identification strategy. The most common approach has been to draw an analogy between education and the production process of a firm. Inputs have often been assumed to be exogenous determinants of outcomes. This approach is highly restrictive. Only test scores in language, mathematics and/or science have been considered among outcomes of schooling; inputs have been considered the only significant component of a school system; the endogeneity of the latter has often not been addressed in a satisfactory way, and a linear functional form has been assumed for f. Data quality has also often been an issue, especially in macro studies. More importantly, no theoretical modeling has been undertaken that would explicitly represent cognitive processes, student behavior, teacher behavior and student-teacher interactions.

Alternative approaches have been developed to deal with these issues. Experimentalists have been able to provide convincing answers to the endogeneity problem; they have also considerably reduced specification issues by making the use of control variables dispensable. A few papers propose additional structural modeling that takes the determination of school or family inputs into account. Researchers focusing on incentives have argued that they matter as well as resources and have greater policy relevance. Specific data collection has also become more frequent, whatever the estimation technique in use.

However, the exclusive focus on cognitive achievement as well as the lack of a proper theory of schools remain. To most economists, schools are black boxes that produce test scores. To other social scientists, they are social institutions which provide students with cognitive skills among other intended or de facto outcomes, in a way that can be explained if one considers student and teacher behavior. Economics thus may benefit more from supplementing the education production function production framework using insights gained from the other social sciences than from seeking methodological improvements *within* it. The economics of education seems a natural field of application for behavioral economics, which might provide more convincing empirical results than the education production function literature.

APPENDIX

Table 1 : Summary of the cross-country evidence

Study	Type of data; year; number of observations	Educational outcome	Educational expenditures	Results
Al Samarrai (2002)	Cross-country; 1996; 33 to 90	Primary gross and net enrolment ratios Survival rate to primary grade five Primary school completion rate	Public primary education spending (% GNP) Primary expenditure per pupil Primary pupil-teacher ratio	Primary expenditure per pupil has a positive and significant impact on survival rate to grade five only (10 %), and a negative and significant impact on the enrolment ratios (1 % and 5 %, respectively). All other coefficients are insignificant.
Hanushek and Kimko (2000)	Cross-country panel; 1965, 1970, 1988, 1991; 67 to 70	IEA and IAEP mathematics and science tests	Pupil-teacher ratio Current education spending per pupil Total expenditure on education (% GDP)	Insignificant Negative impact, 1 % Negative impact, 5 %
Woessmann (2000)	Cross-country; 1995; 39	TIMSS mathematics and science scores	Class size	Positive impact, 1 %
Lee and Barro (1997)	Cross-country panel: 1964, 1970, 1975, 1980, 1982, 1984, 1985, 1990 (depending on the outcome considered) 214 to 346	Test scores Primary school repetition rates Primary school drop-out rates	Pupil-teacher ratio Average teacher salary Current education spending per pupil	Pupil-teacher ratio has a negative and significant impact on all three variables (5 %, 1 % and 5 %). Average teacher salary has a positive and significant impact on test scores (10 %). All other coefficients insignificant.
McMahon (1999)	Cross-country; early 1990s; 44 to 50	Primary male and female gross enrolment ratios Male and female fifth grade completion rate	Public recurrent expenditure on primary education (% GNP) Public recurrent expenditure per primary student (% GNP per capita) Public recurrent expenditure per primary student (level)	Total expenditure has a positive and significant impact (1 %). Per capita expenditure has a negative and significant impact (1 %). Positive impact, 1 %

Study	Type of data; year; number of observations	Educational outcome	Educational expenditures	Results
Gupta, Verhoeven and Tiongson (1999)	Cross-country: 1993-94 23 to 42	Gross primary and secondary enrolment rates Persistence through grade four	Primary and secondary education spending (% total education spending) Education spending (% GDP)	Expenditure on primary and secondary education has a positive and significant impact on all three variables (1 to 5 %, 5 to 10 %, 5 to 10 %) Total education spending has a significant and positive impact on secondary enrolment only (5 %); insignificant for the other two variables.
Schultz (1995)	Country panel; 1965-1980; 60 to 191	Primary gross enrolment ratio	Public teacher compensation as (% GNP per working age adult)	Negative impact, 1 %
Colclough with Lewin (1993)	Cross-country: 1986; 82	Primary gross enrolment ratio	Public recurrent expenditure on primary (% GNP) Public recurrent expenditure per primary student (% GNP per capita)	Insignificant Negative impact, 1 to 5 %

Source: Table based on Al Samarrai's (2002) Table 2.1, with the addition of results from Gupta et al (2000) and Al Samarrai (2002). The detailed sources are as follows: Hanushek and Kimko (2000) results taken from Table 3, Woessmann (2000) from Table 1, Lee and Barro (1997) from Table 3, McMahon (1999) from p. 164 and p. 166, Schultz (1995) from Tables 2 and 3, Colclough with Lewin (1993) from Table 2.6 a. Gupta, Verhoeven and Tiongson (1999) Table 1, Al Samarrai (2002) taken from Tables 4.1 and 4.2. Lee and Barro (1997) was published as Lee and Barro (2001), which is referred to in section 2 above.

Table 2 : Results of Hanushek and Luque (2003)

Distribution of estimated production function parameters across countries and age groups, by sign and statistical significance (10 percent level)
Dependent variable: classroom average TIMSS mathematics score

	Age 9 population					Age 13 population				
	Negative		Positive		Number of countries	Negative		Positive		Number of countries
	Significant	Not significant	Significant	Not significant		Significant	Not significant	Significant	Not significant	
Class size	**3**	11	2	1	17	**2**	8	6	17	33
Teacher with at least a bachelor's degree	0	3	12	**0**	15	2	11	12	**2**	32
Teacher with special training	0	7	4	**1**	12	0	12	11	**2**	25
Teacher experience	0	7	6	**4**	17	3	9	17	**4**	33
School enrollment	0	9	5	3	17	2	9	15	6	32

Note: Bold indicates the number of statistically significant results with the expected sign of the effect. Because these estimates rely on actual class size, its expected sign is negative while the estimates for teacher education and experience have an expected positive sign. No clear expectation exists for school enrollment.

Source: Hanushek and Luque (2003), Table 2.

Table 3 : Hanushek's (2003) tabulation of US production function estimates

Percentage distribution of estimated effect of key resources on student performance, based on 376 production function estimates

Resources	Number of estimates	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
Real classroom resources	276	14	14	72
Teacher education	170	9	5	86
Teacher experience	206	29	5	66
Financial aggregates				
Teacher salary	118	20	7	73
Expenditure per pupil	163	27	7	66
Other				
Facilities	91	9	5	86
Administration	75	12	5	83
Teacher test scores	41	37	10	53

Source: Hanushek (2003), Table 3.

Table 4 : Hanushek's (2003) tabulation of developing-country production function estimates

Percentage distribution of estimated expenditure parameter coefficients from 96 educational production function estimates

Inputs	Number of estimates	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
Teacher / Pupil Ratio	30	27	27	46
Teacher education	63	56	3	41
Teacher experience	46	35	4	61
Teacher salary	13	31	15	54
Expenditure / pupil	12	50	0	50
Facilities	34	65	9	26

Source: Hanushek (2003), Table 6.

Table 5 : Krueger's (2003) reanalysis of Hanushek's tabulation of US estimates

Reanalysis of Hanushek's (1997) literature summary of class size studies

Result	Hanushek's weights (1)	Studies equally weighted (2)	Studies weighted by journal impact factor (3)	Regression-adjusted weights (4)
Positive & stat. sig. (%)	14.8	25.5	34.5	33.5
Positive & stat. insig. (%)	26.7	27.1	21.2	27.3
Negative & stat. sig. (%)	13.4	10.3	6.9	8.0
Negative & stat. insig. (%)	25.3	23.1	25.4	21.5
Unknown sign & stat. insig. (%)	19.9	14.0	12.0	9.6
Ratio positive to negative	1.07	1.57	1.72	2.06
p-value*	0.500	0.059	0.034	0.009

Notes: […] Column (1) is from Hanushek (1997, Table 3), and weights studies by the number of estimates that Hanushek extracted from them. Columns (2), (3) and (4) are author's tabulations based on data from Hanushek (1997). Column (2) weights each estimated by the inverse of the number of estimates taken from that study, thus weighting each study equally. Column (3) calculates a weighted average of the data in column (2), using the 'journal impact factor' as weights; articles that are not published in a journal are assigned the lowest journal impact factor. Column (4) […] adjusts for sample selection […] Table is based on 59 studies.

* p-value corresponds to the proportion of times the observed ratio, or a higher ratio, of positive to negative results would be obtained in 59 independent random draws in which positive and negative results were equally likely.

Source: Krueger (2003), Table 2.

BIBLIOGRAPHY

Akerlof G. A. and Kranton R. E. (2002), 'Identity and Schooling: Some Lessons for the Economics of Education', *Journal of Economic Literature*, 40, December, pp. 1167-201.

Al Samarrai S. (2002), 'Achieving Education for All: How Much Does Money Matter?', Working Paper No. 175, IDS, Brighton, December.

Angrist J. D. and Lavy V. (1999), 'Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement', *Quarterly Journal of Economics*, 114 (2), May, pp. 533-75.

Angrist J. D. and Lavy V. (2001), 'Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools', *Journal of Labor Economics*, 19 (2), April, pp. 343-69.

Angrist J., Bettinger E., Bloom E., King E. and Kremer M. (2002), 'Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment', *American Economic Review*, 92 (5), December, pp. 1535-58.

Banerjee A., Kremer M., Lanjouw J. and Lanjouw P. (2002), 'Teacher-Student Ratios and School Performance in Udaipur, India: A Prospective Evaluation', working paper, Harvard University.

Banerjee A., Jacob S. and Kremer M. (2002), 'Promoting School Participation in Rural Rajasthan: Results from Some Prospective Trials', working paper, MIT.

Banerjee A., Cole S., Duflo E. and Linden L. (2003), 'Remedying Education: Evidence from Tzo Randomized Experiments in India', working paper, MIT, September.

Bardhan P. (2002), 'Decentralization of Governance and Development', *Journal of Economic Perspectives*, 16 (4), Fall, pp. 185-205.

Barro R. J. and Wha Lee J. (1996), 'International Measures of Schooling Years and Schooling Quality', *American Economic Review*, 86 (2), May, pp 218-23.

Baudelot C., Leclercq F., Châtard A., Gobille B. and Satchkova E. (forthcoming), *Les Effets de l'éducation*, La Documentation française, Paris.

Behrman J. R., Foster A. D., Rosenzweig M. R., and Vashishtha P. (1999), 'Women's Schooling, Home Teaching, and Economic Growth', *Journal of Political Economy*, 107 (4), August, p. 682-714.

Black S. E. (1999), 'Do Better Schools Matter? Parental Valuation of Elementary Education', *Quarterly Journal of Economics*, 114 (2), May, pp. 577-99.

Bobonis G., Miguel E. and Sharma C. (2002), 'Iron Supplementation and Early Childhood Development: A Randomized Evaluation in India', working paper, University of California, Berkeley.

Bonesrønning H. (2004), 'The Determinants of Parental Effort in Education Production: Do Parents Respond to Changes in Class Size?', *Economics of Education Review*, 23 (1), February, pp. 1-9.

Bowles S. and Gintis H. (2000), 'Walrasian Economics in Retrospect', *Quarterly Journal of Economics*, 115 (4), November, pp. 1411-39.

Bowles S. and Gintis H. (1976), *Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life*, New York: Basic Books.

Bowles S. and Gintis H. (2001), 'Schooling in Capitalist America Revisited', working paper, University of Massachussetts and Santa Fe Institute, January.

Bowles S., Gintis H. and Osborne M. (2001 a), 'Incentive-Enhancing Preferences: Personality, Behavior, and Earnings', *American Economic Review*, 91 (2), May, pp. 155-8.

Bowles S., Gintis H., and Osborne M. (2001 b), 'The Determinants of Earnings: A Behavioral Approach', *Journal of Economic Literature*, 39 (4), December, pp. 1137-76.

Brock W. A. and Durlauf S. N. (2001), 'What Have we Learned from a Decade of Empirical Research on Growth? Growth Empirics and Reality', *World Bank Economic Review*, 15 (2), pp. 229-272.

Burtless G. (1995), 'The Case for Randomized Field Trials in Economic and Policy Research', *Journal of Economic Perspectives*, 9 (2), Spring, pp. 63-84.

Cameron S. V. and Heckman J. J. (1993), 'The Non-Equivalence of High School Equivalents', *Journal of Labor Economics*, 11 (1), pp. 1-47.

Card D. (1999), 'The Causal Effect of Education on Earnings', ch. 30 *in* O. C. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, 3A, Amsterdam: North Holland, pp. 1801-61.

Card D. and Krueger A. B. (1992), 'Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States', *Journal of Political Economy*, 100 (1), February, pp. 1-40.

Carneiro P. and Heckman J. J. (2003), 'Human Capital Policy', Working Paper No. 9495, NBER, February.

Carnoy M. (1997), 'Is Privatization through Education Vouchers Really the Answer? A Comment on West', *World Bank Research Observer*, 12 (1), February, pp. 105-16.

Case A. and Deaton A. (1999), 'School Inputs and Educational Outcomes in South Africa', *Quarterly Journal of Economics*, 114 (3), August, pp. 1047-84.

Case A. and Yogo M. (1999), 'Does School Quality Matter? Returns to Education and the Characteristics of Schools in South Africa', Working Paper No 7399, NBER, October.

Chin A. (2002), 'The Returns to School Quality When School Quality if Very Low: Evidence from Operation Blackboard in India', working paper, University of Houston, February.

Clark M. A. and Jaeger D. A. (2002), 'Natives, the Foreign-Born and High School Equivalents: New Evidence on the Returns to the GED', Working Paper No. 462, Industrial Relations Section, Princeton University, April.

Colclough C. and Lewin K. (1993), *Educating All the Children: Strategies for Primary Schooling in the South*, Oxford, Clarendon Press.

Coleman J. S., Campbell E.Q., Hobson C.J., McPartland J., Mood A. M., Weinfeld F.D. and York R.L. (1966), *Equality of Educational Opportunity*, Washington, D.C., U.S. Government Printing Office.

Cox D. and Jimenez E. (1991), 'The Relative Effectiveness of Private and Public Schools: Evidence from Two Developing Countries', *Journal of Development Economics*, 34 (1-2), November, pp. 99-121.

Currie J. (2001), 'Early Childhood Education Programs', *Journal of Economic Perspectives*, 15 (2), Spring, pp. 213-38.

Das J., Dercon S., Habyarimana J. and Krishnan P. (2004), 'When Can School Inputs Improve Test Scores?', Policy Research Working Paper No. 3217, World Bank, February.

Dewey J., Husted T.A. and Kenny L.W. (2000), 'The Ineffectiveness of School Inputs: A Product of Misspecification?', *Economics of Education Review*, 19 (1), February, pp. 27-45 .

Drèze J. and Saran M. (1995), 'Primary Education and Economic Development in China and India: Overview and Two Case Studies' *in* K. Basu, P. K. Pattanaik and K. Suzumura (eds.), *Choice, Welfare, and Development: A Festschrift in Honour of Amartya K. Sen*, Oxford and New York, Oxford University Press, Clarendon Press, pp. 182-241.

Duflo E. (2001), 'Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment', *American Economic Review*, 91 (4), September, pp. 795-813.

Duflo E. (2003), 'Scaling Up and Evaluation' *in Annual Bank Conference in Development Economics Proceedings*, Washington, D.C., World Bank.

Duflo E. (2004), 'The Medium Run Effects of Educational Expansion: Evidence from a Large School Construction Program in Indonesia', *Journal of Development Economics*, 74 (1), June, pp. 163-98.

Duflo E. and Kremer M (2003), 'Use of Randomization in the Evaluation of Development Effectiveness', paper prepared for the World Bank Operations Evaluation Department Conference on Evaluation and Development Effectiveness in Washington, D.C., July 15-16.

Evans W. N. and Schwab R. M. (1995), 'Finishing High School and Starting College: Do Catholic Schools Make a Difference?', *Quarterly Journal of Economics*, 110 (4), November, pp. 941-974.

Friedman M. (1962), *Capitalism and Freedom*, Chicago, University of Chicago Press.

Fuller B. (1987), 'What School Factors Raise Achievement in the Third World?', *Review of Educational Research*, 57 (3), pp. 255-92.

Fuller B. and Clark P. (1994), 'Raising School Effects While Ignoring Culture? Local Conditions and the Influence of Classroom Tools, Rules and Pedagogy', *Review of Educational Research*, 64 (1), pp. 119-57.

Glewwe P. (2002), 'Schools and Skills in Developing Countries: Education Policies and Socioeconomic Outcomes', *Journal of Economic Literature*, June, 40 (2), pp. 436-82.

Glewwe P., Ilias N. and Kremer M. (2002), 'Teacher Incentives', working paper, Harvard University, November.

Glewwe P., Kremer M. and Moulin S. (2002), 'Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya', working paper, Harvard University, November.

Glewwe P., Margaret Grosh, Hanan Jacoby, and Marlaine Lockheed (1995), 'An Eclectic Approach to Estimating the Determinants of Achievement in Jamaican Primary Education', *World Bank Economic Review*, 9 (2), May, pp. 231-58.

Glewwe P., Kremer M., Moulin S. Zitzewitz E. (2004), 'Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya', *Journal of Development Economics*, 74 (1), June, pp. 251-69.

Goldhaber D. D. and. Brewer D.J. (1997), 'Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Education Productivity', *Journal of Human Resources*, 32 (3), Summer, pp. 505-23.

Gundlach E., Woessmann L. and Gmelin J. (2001), 'The Decline of Schooling Productivity in OECD Countries', *Economic Journal*, 111 (471), May, pp. 135-47.

Gupta S., Verhoeven M. and Tiongson E. (1999), 'Does Higher Government Spending Buy Better Results in Education and Health Care?', Working Paper No. 99/21, International Monetary Fund, February.

Hanushek E. A. (1979), 'Conceptual and Empirical Issues in the Estimation of Educational Production Functions', *Journal of Human Resources*, 14 (3), Summer, pp. 351-88.

Hanushek E. A. (1986), 'The Economics of Schooling: Production and Efficiency in Public Schools', *Journal of Economic Literature*, 24 (3), pp. 1141-77.

Hanushek E. A. (1989), 'Expenditures, Efficiency, and Equity in Education: The Federal Government's Role', *American Economic Review*, 79 (2), pp. 46-51.

Hanushek E. A. (1994), 'Money Might Matter Somewhere: A Response to Hedges, Laine, and Greenwald', *Educational Researcher*, 23 (4), pp. 5-8.

Hanushek E. A. (1995), 'Interpreting Recent Research on Schooling in Developing Countries', *World Bank Research Observer*, 10 (2), August, pp. 227-46.

Hanushek E. A. (1996 a), 'A More Complete Picture of School Resource Policies', *Review of Educational Research*, 66, pp. 379-409.

Hanushek E. A. (1996 b), 'School Resources and Student Performance' *in* G. Burtless (ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, Washington, DC, Brookings Institution.

Hanushek E. A. (1997), 'Assessing the Effects of School Resources on Student Performance: An Update', *Education Evaluation and Policy Analysis*, 19 (2), Summer, pp. 141-64.

Hanushek E. A. (1998), 'Conclusion and Controversies about the Effectiveness of School Resources', *Economic Policy Review*, Federal Reserve Bank of New York, 4 (1), March, pp. 11-28.

Hanushek E. A. (1999 a), 'The Evidence on Class Size', in S. E. Mayer and P. E. Peterson (eds.), *Earning and Learning: How Schools Matter*, Washington, D.C., National Center for Education Statistics, pp. 185-95.

Hanushek E. A. (1999 b), 'Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects', *Educational Evaluation and Policy Analysis*, 21 (2), Summer, pp. 143-63.

Hanushek E. A. (2002 a), 'Evidence, Politics, and the Class Size Debate' *in* L. Mishel and R. Rothstein (eds.), *The Class Size Debate*, Washington, DC, Economic Policy Institute, pp. 37-65.

Hanushek E. A. (2002 b), 'Publicly Provided Education', ch. 30 *in* A. J. Auerbach and M. Feldstein (eds.), *Handbook of Public Economics*, 4, Elsevier, pp. 2045-141.

Hanushek E. A. (2003), 'The Failure of Input-based Schooling Policies', *Economic Journal*, 113 (485), February, pp. 64-98.

Hanushek E. A. and Kimko D. D. (2000), 'Schooling, Labor-Force Quality, and the Growth of Nations', *American Economic Review*, 90 (5), December, pp. 1184-208.

Hanushek E. A. and Luque J. A. (2003), 'Efficiency and Equity in Schools around the World', *Economics of Education Review*, 22 (5), October, pp. 481-502.

Hanushek E. A. and Rivkin S. G. (2003), 'Does Public School Competition Affect Teacher Quality?' *in* C. Minter Hoxby (ed.), *The Economics of School Choice*, Chicago, University of Chicago Press.

Hanushek E. A., Kain J.F. and Rivkin S.G. (1999), 'Do Higher Salaries Buy Better Teachers?', Working Paper No. 7082, NBER, April.

Hanushek E. A., Kain J.F. and Rivkin S.G. (2004), 'Why Public Schools Lose Teachers', *Journal of Human Resources*, 39 (2), Spring, pp. 326-54.

Heady C. (2000), 'What Is The Effect of Child Labour on Learning Achievement? Evidence from Ghana', Innocenti Working Papers No. 79, Florence, UNICEF Innocenti Research Centre, October.

Heckman J. J. and Rubinstein Y. (2001), 'The Importance of Noncognitive Skills: Lessons from the GED Testing Program', *American Economic Review*, 91 (2), May, pp. 145-9.

Heckman J. J., Lochner L. J. and Todd P.E. (2003), 'Fifty Years of Mincer Earnings Regressions', Working Paper No. 9732, NBER, May.

Hedges L. V. and Greenwald R. (1996), 'Have Times Changed? The Relationship Between School Resources and Student Performance' *in* G. Burtless (ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, Washington, DC, Brookings Institution.

Hedges L. V., Laine R. D. and Greenwald R. (1994), 'Does Money Matter? A Meta-analysis of Studies of the Effects of Differential School Inputs on Student Outcomes', *Educational Researcher*, 23 (3), pp. 5-14.

Heyneman S. P. and Loxley W. (1983), 'The Effect of Primary School Quality on Academic Achievement across Twenty-Nine High and Low Income Countries', *American Jounal of Sociology*, 88, May, pp. 1126-94.

Hoxby C. M. (2000 a), 'The Effects of Class Size on Student Achievement: New Evidence from Population Variation', *Quarterly Journal of Economics*, 115 (4), November, pp.1239-85.

Hoxby C. M. (2000 b), 'Does Competition among Public Schools Benefit Students and Taxpayers?', *American Economic Review*, 90(5), December, pp. 1209-38.

Hoxby, C. M. (ed.) (2003), *The Economics of School Choice*, Chicago, University of Chicago Press.

James E., King E. M. and Suryadi A. (1996), 'Finance, Management, and Costs of Public and Private Schools in Indonesia', *Economics of Education Review*, 15 (4), October, pp. 387-398.

Jepsen C. and Rivkin S. (2002), 'What is the Tradeoff Between Smaller Classes and Teacher Quality?', Working Paper No. 9205, NBER, September.

Jimenez E. and Paqueo V. (1996), 'Do Local Contributions Affect the Efficiency of Public Primary Schools?', *Economics of Education Review*, 15 (4), October, pp. 377-86. .

Jimenez E. and Sawada Y. (1998), 'Do Community-Managed Schools Work? An Evaluation of El Salvador's EDUCO Program', Impact Evaluation of Education Reforms Paper No. 8, World Bank, February

Kingdon G. G. (1996 a), 'The Quality and Efficiency of Private and Public Education: A Case-Study of Urban India', *Oxford Bulletin of Economics and Statistics*, 58 (1), February, pp. 57-82.

Kingdon G. G. (1996 b), 'Student Achievement and Teacher Pay: A Case-Study of India', Discussion Paper n° 74, STICERD, London School of Economics, August.

Kingdon G. G. (1996 c), 'Private Schooling in India: Size, Nature and Equity-Effects', *Economic and Political Weekly*, 31 (51), December 25, pp. 3306-14.

Kingdon G. G. (2002), 'The Spread of Private Schooling in India', working paper, Department of Economics, University of Oxford.

Kingdon G. and Teal F. (2003), 'Does Performance-Related Pay for Teachers Improve Student Performance? Some Evidence from India', working paper, Department of Economics, University of Oxford, November.

Kremer, Michael (1995), 'Research on Schooling: What We Know and What We Don't: A Comment on Hanushek', *World Bank Research Observer*, 10 (2), August, pp. 247-54.

Kremer M. (2003), 'Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons', *American Economic Review*, 93 (2), May, pp. 102-6.

Kremer M. and Sarychev A. (2000), 'Why Do Governments Operate Schools?', working paper, Harvard University, October.

Kremer M., Miguel E. and Thornton R. (2003), 'Incentives to Learn', working paper, Harvard, October.

Kremer M., Moulin S. and Namunyu R. (2003), 'Decentralization: A Cautionary Tale', working paper, Harvard University, March.

Kremer M., Moulin S., Namunyu R. and Myatt D. (1997), 'The Quantity-Quality Tradeoff in Education: Evidence from a Prospective Evaluation in Kenya', working paper, Harvard University.

Krueger A. B. (1998), 'Reassessing the View that American Schools Are Broken', *Economic Policy Review*, Federal Reserve Bank of New York, 4 (1), March, pp. 29-46.

Krueger A. B. (1999), 'Experimental Estimates of Education Production Functions', *Quarterly Journal of Economics*, 114 (2), May, pp. 497-534.

Krueger A. B. (2002), 'Understanding the Magnitude and Effect of Class Size on Student Achievement' *in* L. Mishel and R. Rothstein (eds.), *The Class Size Debate*, Washington, DC, Economic Policy Institute, pp. 7-35.

Krueger A. B. (2003), 'Economic Considerations and Class Size', *Economic Journal*, 113 (485), February, pp. 34-63.

Krueger A. B. and Lindahl M. (2001), 'Education for Growth: Why and For Whom?', *Journal of Economic Literature*, 39 (4), December, pp. 1101-36.

Krueger A. B. and Whitmore D. M. (2001), 'The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR', *Economic Journal*, 111, January, pp. 1-28.

Krueger A. B. and Whitmore D. M. (2002), 'Would Smaller Classes Help Close the Black-White Achievement Gap?' *in* J. Chubb and T. Loveless (eds.), *Bridging the Achievement Gap*, Washington, DC, Brookings Institute Press, 2002.

Ladd H. F. (2002), 'School Vouchers: A Critical View', *Journal of Economic Perspectives*, 16 (4), Fall, pp. 3-24.

Lavy V. (1998), 'Disparities between Arabs and Jews in School Resources and Student Achievement in Israel', *Economic Development and Cultural Change*, 47 (1), October, pp. 175-192.

Lavy V. (2003), 'Paying for Performance: The Effects of Teachers' Financial Incentives on Students' Scholastic Outcomes', Working Paper No. 022, BREAD, February.

Lazear E. P. (2001), 'Educational Production', *Quarterly Journal of Economics*, 116 (3), August, pp. 777-803.

Lee J-W. and Barro R. J. (2001), 'Schooling Quality in a Cross-Section of Countries', *Economica*, 38 (272), November, pp. 465-88.

Lopez-Acevedo G. (2004), 'Professional Development and Incentives for Teacher Performance in Schools in Mexico', Policy Research Working Paper No. 3236, World Bank, March.

McMahon W. (1999), *Education and Development: Measuring the Social Benefits*, Oxford, Oxford University Press.

Miguel E. and Kremer M. (2004), 'Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities', *Econometrica*, 72 (1), January, pp. 159-218.

Mocan H. N., Scafidi B. and Tekin E. (2002), 'Catholic Schools and Bad Behavior', Working Paper No. 9172, NBER, September.

Neal D. (1997), 'The Effect of Catholic Secondary Schooling on Educational Attainment', *Journal of Labor Economics* 15 (1), January, pp. 98-123.

Neal D. (2002), 'How Vouchers Could Change the Market for Education', *Journal of Economic Perspectives*, 16 (4), Fall, pp. 25-44.

Newman J., Rawlings L. and Gertler P. (1994), 'Using Randomized Control Designs in Evaluating Social Sector Programs in Developing Countries', *World Bank Research Observer*, 9 (2), July, pp. 181-201.

Patrinos H. A. (2002), 'A Review of Demand-Side Financing Initiatives in Education', working paper, World Bank, August.

Pritchett L. (2001), 'Where Has All the Education Gone?', *World Bank Economic Review*, 15 (3), August, pp. 367-91.

Pritchett L. (2002 a), 'Ought Ain't Is: Midnight Thoughts on Education', working paper, Harvard University, January.

Pritchett L. (2002 b), '"When Will They Ever Learn?": Why *All* Governments Produce Schooling', working paper, Harvard University, February.

Pritchett L. (2002 c), 'Inculcation of Beliefs Is Not Third-party Contractible', working paper, Harvard University.

Pritchett L. (2003), 'Basic Education Services', ch. 7 in World Bank, *World Development Report 2004: Making Services Work for Poor People*, Washington, DC, World Bank.

Pritchett L. and Filmer D. (1999), 'What Education Production Functions *Really* Show: A Positive Theory of Education Expenditures', *Economics of Education Review*, 18 (2), pp. 223-39.

Ravallion M. and Wodon Q. (2000), 'Does Child Labour Displace Schooling? Evidence on Behavioral Responses to an Enrolment Subsidy', *Economic Journal*, 110, pp. 158-75.

Rawlings L. B. and Rubio G. M. (2003), 'Evaluating the Impact of Conditional Cash Transfer Programs: Lessons from Latin America', Policy Research Working Paper No. 3119, World Bank, August.

Rivkin S. G., Hanushek E. A. and Kain J. F. (2002), 'Teachers, Schools and Academic Achievement', working paper, Stanford University, July. Revised version of NBER Working Paper No. 6691, August 1998.

Rosenzweig M. R. and Wolpin K. I. (2000), 'Natural "Natural Experiments" in Economics', *Journal of Economic Literature*, 38 (4), December, pp. 827-74.

Rutter M., Maughan B., Mortimore P. and Ouston J. (1979), *Fifteen Thousand Hours: Secondary Schools and their Effects on Children*, London, Open Books.

Schultz T. P. (1995), 'Accounting for Public Expenditures on Education: An International Panel Study' *in* T. Paul Schultz (ed.), *Research in Population Economics*, 8, Greenwich, CT, JAI Press, 8.

Schultz T. P. (2004), 'School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program', *Journal of Development Economics*, 74 (1), June, pp. 199-250.

Strauss J. and Thomas D. (1995), 'Human Resources: Empirical Modeling of Household and Family Decisions', ch. 34 *in* J. R. Behrman and T. N. Srinivasan (eds.), *Handbook of Development Economics*, 3, Amsterdam, North Holland, pp. 1883-2023.

Tan J-P., Lane J. and Lassibille G. (1999), 'Outcomes in Philippine Elementary Schools: An Evaluation of Four Experiments', *World Bank Economic Review*, 13 (3), August, pp. 493-508.

Teddlie C. and Reynolds D. (2000), *The International Handbook of School Effectiveness Research*, London, Falmer Press.

Todd P. E. and Wolpin K. I. (2003), 'On the Specification and Estimation of the Production Function for Cognitive Achievement', *Economic Journal*, 113 (485), February, pp. 3-33.

Tyler J. H., Murnane R.J. and Willett J. B. (2000), 'Estimating the Labor Market Signaling Value of the GED', *Quarterly Journal of Economics*, 115 (2), May, pp. 431-68.

Unnever J. D. (2001), 'Book Review: *The International Handbook of School Effectiveness Research*', *Economics of Education Review*, 20, pp. 515-6.

Vegas E., Pritchett L. and Experton W. (1999), 'Attracting and Retaining Qualified Teachers in Argentina: Impact of the Structure and Level of Compensation', working paper, World Bank, April.

Vermeersch C. (2002), 'School Meals, Educational Achievement and School Competition: Evidence from a Randomized Experiment', working paper, Harvard University, November.

Vignoles A., Levacic R., Walker J., Machin S. and Reynolds D. (2000), 'The Relationship Between Resource Allocation and Pupil Attainment: A Review', working paper, Center for the Economics of Education, London School of Economics, September.

West E. G. (1997), 'Education Vouchers in Principle and Practice: A Survey', *World Bank Research Observer*, 12 (1), February, pp. 83-103.

Woessmann L. (2000), 'Schooling Resources, Education Institutions, and Student Performance: The International Evidence', Working Paper No. 983, Kiel Institute of World Economics.