CENTRE *for* ECONOMIC
PERFORMANCE

CEP Discussion Paper No 1226

June 2013

The Effects of Resources Across School Phases: A Summary of Recent Evidence

Stephen Gibbons and Sandra McNally

LSE THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE ■

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Abstract
This report provides an overview and discussion of the past decade of academic evidence on the causal effects of resources in schooling on students' outcomes. Early evidence lacked good strategies for estimating the effects of schools resources, leading many people to conclude that spending more on schools had no effect. More recent evidence using better research designs finds that resources do matter, but the range of estimates of the impacts is quite wide. The review devotes special attention to differences across the early years, primary and secondary phases. Theoretical work has indicated that interventions early in a child's life may be more productive than interventions later on. However, although there are more examples of good quality studies on primary schooling, the existing body of empirical work does not lead to a conclusive case in favour of early interventions.

Keywords: education, school resources, government policy, pupil premium, education funding, inequality, OECD

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

Introduction, Background and Scope

By 2009, Britain was well above the average amongst OECD countries in terms of the share of national income spent on primary and secondary schooling, spending about 4.5% of GDP compared to an OECD average of 4.0% and the EU average of 3.8% (OECD, 2012). At the same time, Britain's performance in the OECD PISA international student assessment was broadly similar to the rest of Europe and the OECD as a whole (OECD, 2011)[1]. As Figure 1 shows, real total public expenditure in England increased substantially in both the primary and secondary phases up to 2009, rising from around £13 billion in 2005 up to around £14.5 billion in 2009 for each phase (all in 2005 prices). Prior to this period there were even more dramatic increases in public funding, with primary and secondary per-pupil spending increasing in real terms by over 47% in real terms between 2000 and 2007[2]. Since 2009, expenditure has levelled off but still stands at around £15 billion in each of the primary and secondary phases (expenditure on nursery education is considerably less, because coverage is not universal; see Figure 1 below).

A key question then, especially in times of austerity in the wake of the 'Great Recession', is whether this is a good use of resources. A superficial look at academic results at the end of compulsory education suggests considerable payoff to expenditure increases over the decade, with the proportion achieving 5 or more A*-C GCSEs increasing from 49% in 2002 to 64% in 2008 (DfE figures[3]). Primary school performance has shown less improvement with the proportion reaching the target Level 4 grade rising by only 5% from 75-80% after 2002, and higher achievement at Level 5 staying roughly constant at around 30%. However, many commentators raise doubts over to what extent any gains in these national tests represents a genuine improvement in standards, given that UK students showed almost no change in performance in the OECD international student assessment (PISA) reading, science tests or maths tests between 2006 and 2009[4]. Businesses and universities also remain perennially disappointed in the capabilities of students leaving the educational system, and the public perception is of low standards despite the good league table results.

A more nuanced consideration is to what extent the allocation of resources across the education phases is optimal in terms of benefits for student outcomes, and whether rather than increasing resources overall, a redistribution across the phases would generate benefits. According to OECD figures, spending per pupil is fairly well balanced across the primary and secondary phases ($US 9000 and $US 10000 respectively in 2009). Compared to other OECD countries, spending per pupil on primary education is high in Britain relative to spending other phases. We spend 10% more per pupil on secondary education than primary education, whereas the OECD average is 20%. We spend 25% less per pupil on nursery education than on primary education, compared to 11% less in other countries. The general

---

[1] In 2009, the UK was 1 point above the OECD average and 2 points above the western Europe average in Reading (494, 493, 492, respectively), 4 points below the OECD average and 8 points below the western Europe average in Maths (492, 496 and 500 respectively), and 13 points above the OECD average and 11 points above the western Europe average in Science (514, 501 and 503 respectively). Averaging across all the subjects, the UK was 1 point above the western Europe average and 3 points above the OECD average (500,497,499 respectively). The standard deviation of scores across countries is 25.5.

[2] http://www.education.gov.uk/researchandstatistics/statistics/allstatistics/a00196759/spending-per-pupil-in-cash-terms

[3] http://www.education.gov.uk/researchandstatistics/statistics/allstatistics/a00196904/gcse-attainment-by-eligibility-for-free-school-mea

[4] Reading scores were 3 points above the OECD average at 495 in 2006, and 1 point above the OECD average in 2009. Maths scores were 3 points below the OECD average at 495 in 2006, and 4 points below the OECD average at 492 in 2009. Science scores were 15 points above the OECD average at 515 in 2006 and 13 points above the OECD average in 2009.

patterns are illustrated in Figure 2, which plots nursery and secondary education expenditure per pupil against primary expenditure. Red labels show nursery spending, and the majority of labels below the 45 degree line indicate that countries spending less on nursery education (per pupil) than primary education. Blue labels show secondary spending, with the majority above the 45 degree line. Evidently there is considerable variety amongst countries in terms of expenditure policy across the phases.

It is of course inappropriate to draw inferences about the impacts of expenditure or other policy variables by simply comparing changes in expenditure with changes in performance over time, or by simple comparisons across countries. Changes in student achievement within countries over time, and differences between countries arise through a myriad of other channels. Deeper research is needed. With these questions in mind, this report provides a summary of recent academic evidence on the causal effects of resources in schooling on students' outcomes.[5] The review and discussion makes special reference to the differences in effects at different phases of schooling from early years, through primary to secondary schooling.

The survey looks at empirical work on the impact of additional resources on the outcomes of students, in terms of school achievements and qualifications. We also consider longer run outcomes, including continuation to higher education and subsequent labour market earnings, where evidence is available. By 'educational resources' we mean increases in general expenditure per student, or resource-based educational interventions and policy changes which explicitly involve additional resources, rather than changes in pedagogic methods. These resource-based interventions can involve additional spending on technology, infrastructure or other material resources. But usually, 'resources' are linked to staff costs either through employment or pay. In particular, general increases in expenditure per pupil are most commonly linked to reductions in pupil/teacher ratios or class sizes, so the survey will also describe the extensive literature on the effects of class-size reduction, which has been a fertile area for research.

The evidence presented in the report is drawn mainly from the literature in economics and economics of education. Researchers in this field have an interest in the costs and benefits of education policy. Therefore, estimation of the impacts of spending on outcomes, both for students and the wider economy has been a natural focus of research attention. A natural take off point for a survey of recent evidence in this discipline is the debate featured in a special issue of the Economic Journal in 2003. The centrepiece of this issue was two articles, one by Eric Hanushek and one by Alan Krueger (Hanushek, 2003; Krueger, 2003), which set out what was, at the turn of century, and is still, a central question over the interpretation of the evidence on the link between resources and outcomes in schooling. One view that has become widespread amongst economists and some other social scientists (epitomised by Hanushek's work based on meta-analyses of earlier studies) is that the available evidence, on balance, shows no benefits from additional spending in schools, at least at the margin available to policy makers in developed economies. The alternative view (set out by Krueger) is that much of the traditional statistical evidence is of poor quality, and that reliable estimates require careful experimental or experiment-like research designs. Emerging evidence from studies of this type is much more favourable to resource-based school interventions like class-size reductions. These general arguments still shape the landscape of research on school resources, and it is from this point that our survey picks up the debate, covering evidence that has emerged since then and spanning the decade since the Krueger/Hanushek articles which summarised the state of play in 2003.

---

[5] The review was commissioned by Ofsted as part of a project on evaluating the effectiveness of educational policy across all schooling phases.

Recent research in this field has paid close attention to issues of 'causality' i.e. establishing a causal connection running from changes in spending to changes in outcomes, rather than correlations and other statistical associations, because of the importance of causal estimates in informing policy. There have been many advances in the understanding of the challenges in measuring these causal effects of school resources and these advances have led to a better appreciation of the limitations of early evidence. In response, there has been more widespread adoption of improved estimation methods and research designs in empirical research. Greater data availability has also assisted this development. The review is set against this background of improvements in practice and data, which also help define its scope. Our review is selective in the sense that it highlights evidence based on what we consider good practice in empirical methods aimed at estimating the causal link between spending in schools and student outcomes. We make use of the many existing surveys that summarise the state of the literature recently, and prior to 2002. In addition, we highlight specific recent studies in more depth where we judge these examples to be of particular relevance or importance. In some cases we highlight studies which, while not exhibiting best practice in terms of data and methods, may be of interest because they look at special issues not covered elsewhere, or cover contexts where better data and methods are simply not available.

The structure of the report is as follows. The next section outlines the methods used in the empirical work that is covered by the review, in order to highlight both the limitations of some types of study, and the strengths of others, to assist the reader in making their own evaluations of the evidence. The main body of the survey splits the material into sections covering the effects of expenditure in two school phases: early years and primary school in Section 3 and secondary school in Section 4. Within each section, the material is organised into sub-sections relating to geographical coverage, firstly the UK, and then international evidence from other developed countries. Section 5 looks briefly at evidence from developing and less developed countries, where there has been a surge in evidence from policy changes and field experiments. Section 6 outlines the literature on the links between school resources and aggregate (country, regional) outcomes like growth and GDP, although given the limitations of this line of research we do not discuss specific studies in depth. Following this outline of the literature, Section 7 draws together the findings to summarise the limitations of the evidence, and what policy makers can learn given the current state of knowledge in this field, paying particular attention to the relative benefits of expenditure at the various educational phases.

Figure 1: Total real expenditure in England by phase, 2005 to 2011 (£millions, in 2005 prices)
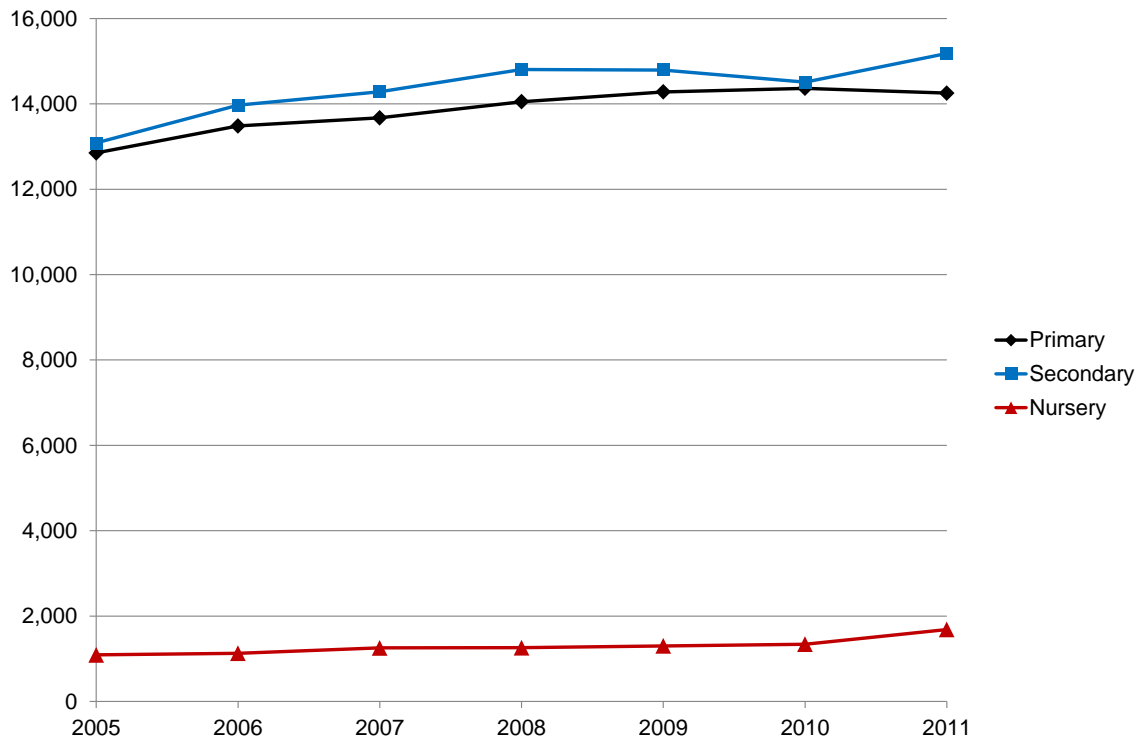
Source: Authors' elaboration from DfE LA summary budget data 2005-2011, Academies Spend Data 2011 and CPI. Academies spending data unavailable prior to 2011. Nursery spending is LA maintained nursery schools only.
http://www.education.gov.uk/schools/adminandfinance/financialmanagement/schoolsrevenuefunding/section251/archive/b0068383/section-251-data-archive/budget-data---summary-level
http://www.education.gov.uk/schools/performance/academies.html

Figure 2: Nursery, Primary and Secondary spending per pupil in OECD countries compared, 2009 (USD, purchasing power parity). Blue labels refer to secondary spending; pink labels nursery spending
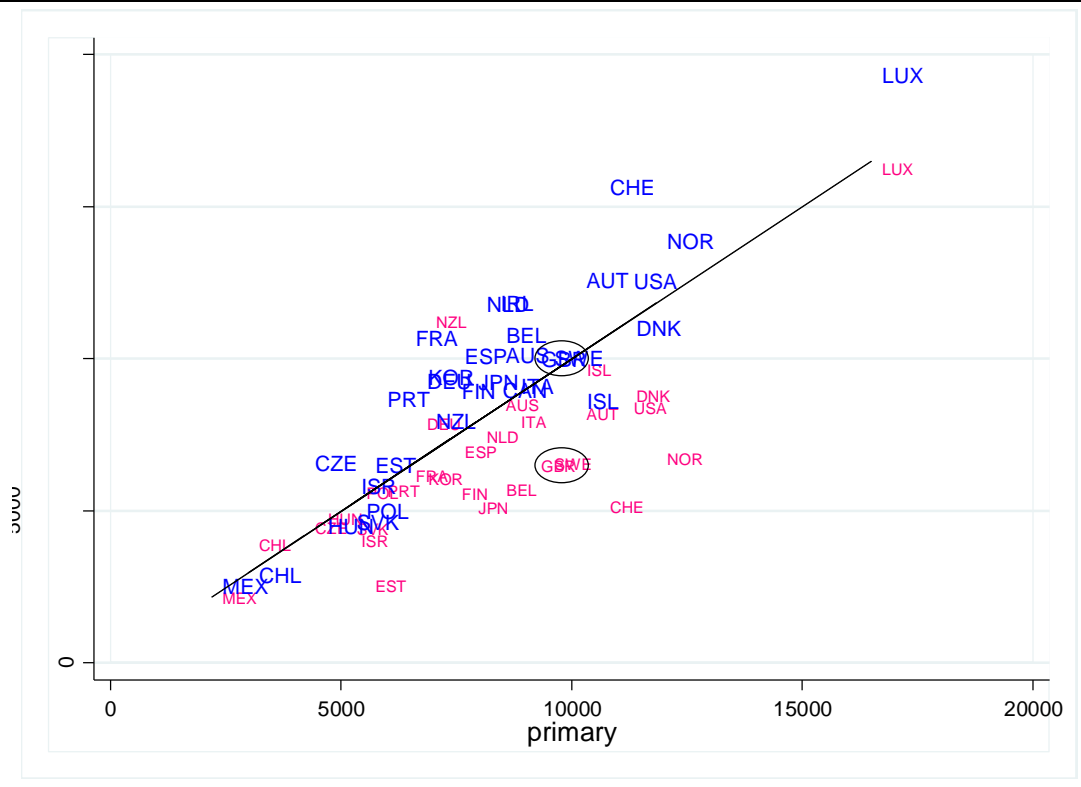
Source: OECD Education at a Glance, 2012.
Graph plots secondary (large blue labels) or nursery (small pink labels) spending against primary expenditure. GBR is in centre of plot, coinciding with Sweden (SWE) and is shown highlighted. The codes refer to standard OECD country codes www.oecd.org/dataoecd/31/15/34107835.xls

2. Methods Used in Empirical Studies of Schooling and Resources

All studies that we consider in this survey investigate the effects of school resources on student outcomes by measuring the statistical association between measures of school resources devoted to schooling and student outcomes – typically test scores or qualifications. As is well known, correlation need not imply causation, so the main challenge to this research program is establishing that these estimated statistical associations are the result of a causal link between resources and student outcomes. By a 'causal estimate' of the effect of resources on e.g. test scores, researchers usually mean an estimate of the change in test scores that would be expected (on average) in a group of students, from a change in the resources spent in school on their education. The challenge in empirical work is to determine whether a statistical association between school resources and student performance occurs because a change in school resources causes a change in student performance – the policy-relevant causal question - or whether a change in student performance leads to more or less resources, or whether changes in student performance and resources are simultaneously affected by something else.

The basic research design used to answer this question is a regression analysis of student outcomes on a chosen measure of resource expenditure. Micro-level analyses use data at the

student, class, school level for estimation. The underlying problem with this is that resource expenditure is potentially correlated with student performance for many reasons other than a causal link running from resources to student outcomes. In general, the issue is that pupil and school characteristics (e.g. ability, teacher quality) in schools or classes with more resources are not necessarily comparable to pupil and school characteristics in schools or classes with fewer resources. These pupil and school characteristics may have a direct effect on achievement, and are not necessarily observable to the researcher in their data. These omitted or confounding factors lead to upward or downward biases in the estimation of causal resource impacts. The principal reasons behind these differences in characteristics between high and low-resource schools are documented in Appendix A, but primarily relate to dependency of funding on school and student characteristics, and the sorting of students and teachers of different types into different schools:

To overcome these problems, research designs need to ensure that comparisons are made between students or schools that face different levels of resources, but are otherwise comparable on a like-for-like basis. The traditional method of achieving this was to use statistical regression techniques to adjust for differences in observable (i.e. recorded in the data) characteristics of students and schools. In these regressions, student outcomes are the 'dependent' variable, measures of school resources are the main explanatory variables, and a wide range of school, student and teacher characteristics are included as additional control variables. These control variables are intended to represent schooling inputs, other than financial resources, but which may be correlated with financial resources. These regression models are called 'educational production functions'. Often these production functions are estimated using test score or achievement gains – value-added – as a dependent variable, or (roughly equivalently) use test scores as a dependent variable and control for prior test scores as an additional control variable. The point of this value-added approach is to ensure that comparison are made between outcome achievements of students with similar levels of prior achievement (which is partly determined by ability and background), thus improving the chances of making valid like-for-like comparisons between students or schools.

The key underlying limitation of the general application of education production function regression models of this type is that they do not specify any explicit source of variation in resources. Estimates are implicitly based on residual differences in resources between students that arise due to factors other than those represented by the control variables in the regressions (school, student characteristics etc.). The problem then is that the results will tend to vary according to the set of control variables available, and it is never clear whether a sufficient number of factors, or too many factors are included. Too few control variables increases the probability of bias due to omitted/confounding factors. But as more and more control variables are added it becomes less and less clear why the students being compared are receiving different resources. For example, a regression of mean student test scores on expenditure per pupil using school level data might control for a wide range of family background characteristics. But if expenditure per pupil is determined by funding rules that depend primarily on students' family background then comparing schools with the same average family background makes little sense, because these schools will have similar levels of resources. There are also common cases of mis-specification where researchers include various interrelated measures of resource differences in the regressions – e.g. expenditure per pupil and pupil teacher ratios. Causal interpretation of these estimates is difficult, because the regression estimates of the effect of the pupil-teacher ratio imply the effect of this increase holding expenditure per pupil constant. Given expenditures per pupil depend heavily on pupil teacher ratios, this implies that the comparison being made is between schools with high pupil teacher ratios and more spending on non-teacher inputs, and schools with low pupil teacher ratios and less spending on non-teacher inputs (in order to hold expenditure per pupil

constant). This is clearly not a 'like for like comparison'. These issues are discussed in greater detail in Todd and Wolpin (2003).

In order to try to overcome these problems, modern research still applies regression techniques, but aims to identify specific sources of variation in resources or class sizes across schools and/or make the comparisons that are being made more explicit. The emphasis has been on finding potentially random differences in resources or class sizes, which are uncorrelated with other student and school characteristics. This means that no, or very few, control variables are required and it is easier for people reading the research to assess whether to have confidence that the estimated relationship is causal. One ideal way to achieve this is to set up a dedicated experiment to investigate the effects of resources or class sizes, randomly assigning children to classes with different levels of resourcing. Randomisation ensures that students getting different levels of resources are, on average, comparable on other dimensions. The Tenessee Project STAR experiment is the most often cited example (see below Section 3.3) of this type of randomised control trial. Rockoff (2009) reports on the findings of a number of other class size experiments conducted prior to World War II. Such experiments are, however, rare and costly to implement, and have their own specific set of problems related to the generalizability of the experiment to the real world, and the possibility that participation in the experiment causes behavioural changes that would not be replicated outside the experiment. Therefore researchers more commonly look for existing settings and contexts where differences in resources can be considered random, due to policy design or natural variation that arises through demographic processes. These 'quasi-experimental' or 'natural experiments' are the backbone of modern applied research in this field.

There are some common standard strategies in the school resource literature. Firstly, natural variation in birth rates generates changes from year to year in school enrolment, which in turn leads to variation in class sizes (assuming teachers are not hired and fired to maintain class sizes constant). Thus, studies can compare outcomes in the same school in different years as enrolment numbers and class sizes change. Hoxby (2000) is usually credited with this idea. The most basic way of doing this is to set up a panel dataset with the same set of schools observed in multiple years. A regression of changes in outcomes from year to year within a school on changes in class sizes from year to year within a school, then provides an estimate of the effect of class size changes taking account of fixed-over-time differences between schools, including those induced by permanent differences in teaching quality and student sorting. This kind of 'fixed effect' design is widely applied, and has been used to look at the effects of expenditure changes (Holmlund, McNally and Viarengo, 2010), variation in peer group quality (Lavy and Schlosser, 2012), student mobility (Gibbons and Telhaj, 2011) and many other factors. Another common method uses class size rules (so called 'Maimonides rules', Angrist and Lavy, 1999) that impose strict limits to class sizes. This means that as enrolment increases from zero, a single class can vary in size up to a maximum threshold, beyond which it is split into two classes. Thus two schools with very comparable enrolmentsms e.g. 28 and 32 will have widely different class sizes (28 and 16) when the maximum class size is 30. Designs of this type are called 'regression discontinuity' designs, because they are based on a discontinuity in class sizes at some enrolment threshold. In the class size studies, this design is typically implemented by predicting class sizes using the class size rules, and using these predictions as a source of random variation in class sizes (an 'instrumental variables' procedure). Other common designs make use of specific policy interventions that allocated more money or teachers to particular sets of schools, comparing outcomes of students in these schools with similar schools not subject to the intervention. One study (Gibbons, McNally and Viarengo, 2012) exploits geographical variation in funding rules, and compares similar close neighbouring

schools that are face similar demographics and prices, but are in different funding districts so get different incomes. Others use political voting shares to predict funding differences based on local government political control, on the assumption that variation in funding is caused by local party policy preferences (Jenkins, Levacic and Vignoles, 2006). This strategy is employed in applied papers in many fields, but is subject to the criticism that the left-right balance of political control tends to be related to underlying demographics in the area, so is not self-evidently unrelated to the demographics in the schools concerned. There are various methods of implementing all these 'quasi-experimental' methods in practice – 'fixed effects' and 'regression discontinuity' designs as discussed above, or 'instrumental variables' methods that predict funding differentials from explicit sources of (putatively) random variation, but the details need not concern us here. A more detailed overview is available in Barrow and Rouse (2005).

Another strand of the school resource literature takes a more 'macro' approach and uses data at a geographically aggregated level e.g. district, state, or country. These studies face a slightly different set of problems, although the basic principles and potential solutions are the same. Sorting and selection due to individuals moving classes and schools in response to quality and resource differences may be less of a challenge, assuming that people do not choose which country or state to live based on schooling. However, countries are different from each other on a whole range of unobserved, confounding dimensions which may influence both resources and student outcomes. These differences can be difficult to control for. Most studies of this type use a fixed-effects panel research design in which regions are observed over time in multiple periods, which allows researchers to control for fixed regional differences between regions in the same way as fixed effect panel studies of schools control for unobserved differences between schools (see above).

In the sections that follow, we present an overview of results from key papers over the past decade. One issue is how to present and compare the size of the resulting estimates across different studies. Where possible we report the magnitudes scaled in terms of standard deviations of the outcome variables. For the reader unfamiliar with this scaling convention, a discussion is provided in Appendix B.

3. Primary Phase and Early Years

3.1 Pre-school investments
It has become increasingly common to argue for the efficacy of early investment in children over investment at a later stage. The Nobel-laureate, James Heckman, has many papers arguing this point. A brief summary of the argument is as follows (in Heckman, 2004): 'early environments play a large role in shaping later outcomes. Skill begets skill and learning begets more learning. Early advantages cumulate; so do early disadvantages. Later remediation of early deficits is costly, and often prohibitively so, though later investments are also necessary since investments across time are complementary. Evidence on the technology of skill formation shows the importance of early investment….'. Heckman's work makes the case that investments in the early years offer higher returns than investments in later years, and that investment is needed in both cognitive and non-cognitive skills. These conclusions are drawn from a dynamic theoretical model i.e. one in which events in one period depend on events in previous periods. From this model, there is an optimal sequence of interventions over the life cycle that delivers the highest returns in terms of later life earnings, which depends on the functional form and parameters of the function that transforms monetary investments (e.g. home or school investments) in to human capital, and hence earnings (Cunha and Heckman, 2007 & 2008). More specifically, the optimal sequence depends on the

extent to which early and late investments in human capital are substitutes or complements in the production of skills, that is whether late investments can fully compensate for lack of early investments, or whether both early and late investments are needed. It also depends on the extent to which early investments make later investments more productive (a 'skill multiplier'). Appendix C illustrates this model in more detail. Cunha and Heckman (2008) provide predictions about the relative benefits of investments at different ages, by calibrating this kind of model using statistical evidence on the associations between parental investments in children, measures of cognitive and non-cognitive skills during child development, and adult earnings. For example, they conclude that early investments in cognitive skills are around twice as productive in early years (age 6-7) than later (age 8 and beyond). [6] On the other hand, investments in non-cognitive skills appear to have maximum payoffs slightly later at age 8-9. Parental investments are measured by things like number of books at home, access to a musical instrument and newspapers, trips to museums, additional lessons and interactions with teachers. The model does, however, make quite a lot of theoretical and empirical assumptions, and it is quite a big step to conclude from this work that expenditures or interventions outside or inside the home are actually effective at changing the path of children's development over the life cycle.

These arguments are part of the rationale behind the substantial investment in early childcare settings in the recent past (in the UK, as in other countries). The evidence is clear that gaps in skills between children from different backgrounds open up at a very early age, as documented in Cunha and Heckman (2007) for the US and for the UK by Feinstein (2003) and Hansen and Hawkes (2009). For example, Hansen and Hawkes (2009) show that family background factors are the strongest predictors of age 3 vocabulary scores for children in the Millenium Cohort Study. Their Table 3 indicates that a child from parents with less than five A*-C GCSEs, has pre-school vocabulary scores almost 0.5 standard deviations below a similar child from a family where both parents are degree educated. These estimates are for families that are otherwise similar in terms of age, composition, ethnicity, employment and childcare arrangements.

The hope has been that early education would help close these gaps in skills identified between children from different backgrounds at the start of school. Whether these early investments have helped to close these gaps , and the effect of resources and quality of provision in pre-school settings, is the subject of on-going research. Economic evidence on specific programmes is often about whether children get access to pre-school relative to a situation where families get no formal help (or full-day versus half day pre-school care). Furthermore, specific programmes are often directed to disadvantaged families and not to all families (e.g. Head Start, the Perry Pre-School programme in the US; Sure Start in the UK). Also, the focus of evaluation is whether participation in the programme has an impact on subsequent outcomes and not on whether variation in class size or expenditure has an impact (unlike in the school resources literature).

To give one example, UK's Sure Start local programme has been the subject of a major evaluation looking at age 5 outcomes, comparing outcomes for children in eligible (treatment) and non-eligible (control) areas. Control areas are matched to treatment areas to make them comparable (Institute for the Study of Children, Families and Social Issues, 2010). The main positive effects for children relate to health - those in the treatment group had lower BMI and better physical health. There are mixed findings on outcomes relating to maternal wellbeing and family functioning. There are no differences between the treatment and control groups on seven measures of cognitive and social development from the Foundation Stage Profile (the teacher assessment carried out on entry to school). There is also

---

[6] For example, Table 17a, in Cunha and Heckman (2008) indicates that a 10% increase in investment in cognitive skills raises earnings by 12.5% if invested at age 6-7, but by only 5.5% if invested at age 10-11.

a reduction of the proportion of children living in families where no parent was in paid work in the treatment group relative to the control group. The report monetises the effects arising from the fact that parents in eligible areas move into work more quickly (£279-£557 per eligible child), which does not compare favourably to the costs of around £4,860 over the period from birth to the age of four. Of course, these figures do not capture the direct benefits on the child through health, cognitive and non-cognitive skills, and the report emphasises that these benefits may not become apparent for 10-15 years. However, it is not clear why these benefits are not evident in the Foundation Stage Profile, and hence where these large gains in educational attainment are expected to come from.

Given the difficulty in comparing the evidence on pre-school interventions, with the evidence on expenditure and class size during school years, the focus of this review is from age-5 onwards (US kindergarten, UK Year 1) – where we do know something about the effect of school expenditure or class size. However, the evidence is not sufficiently strong to allow one to distinguish between the effects of resources at early versus late stages of primary education.

3.2 UK evidence

Reviews of the literature for the UK in the early 2000s found there to be few methodologically strong studies (Blatchford et al., 2002; Levăcić and Vignoles, 2002). Furthermore, similarly to many studies in the international literature, little or no relationship was found between class size/school resources and measures of educational attainment. Since that time, available data has become much richer, enabling several studies of higher quality than in the past. In particular, the National Pupil Database (NPD) for England contains pupil-level information on all pupils attending schools in the state system as they progress through education. It is possible to link these data with school-level information on expenditure.

Two studies that have used the NPD to look at the relationship between expenditure and attainment in primary school are by Holmlund et al (2010) and Gibbons et al (2011). The former use data between the early- and late-2000s - a period in which school expenditure increased by about 40%. They look at the relationship between expenditure and pupil attainment at the end of primary school in the Key Stage 2 tests. Their strategy involves controlling for characteristics of pupils and schools – including 'school fixed effects' and allowing for school-specific time trends in attainment. They find evidence for a consistently positive effect of expenditure across the different subjects (English, Maths and Science). The magnitude corresponds to about a 0.03-0.05 standard deviation increase in attainment for an extra £1,000 in per pupil expenditure. In the context of the literature, this is a small effect – although similar to comparable studies looking at secondary schools.

Gibbons et al (2011) use the same data set over a similar time period to look at the same question. However, they confine attention to schools in urban areas that are close to Local Authority boundaries and compare neighbouring schools on different sides of these boundaries. The percentage of economically disadvantaged children in these schools is much higher than the national average (28% are eligible to receive free school meals, compared to 16% nationally). The strategy uses the fact the closely neighbouring schools with similar pupil intakes can receive markedly different levels of core funding if they are in different education authorities. This is because of anomaly in the funding formula which provides an 'area cost adjustment' to compensate for differences in labour costs between areas whereas in reality teachers are drawn from the same labour market and are paid according to national pay scales. The study shows that schools on either side of Local Authority boundaries receive different levels of funding and that this is associated with a sizeable differential in pupil achievement at the end of primary school. For example, for an extra £1,000 of spending, the effect is equivalent to moving 19% of students currently achieving the expected level (or

grade) in Maths (level 4) to the top grade (level 5) and 31% of students currently achieving level 3 to level 4 (the expected grade at this age, according to the National Curriculum). The magnitude of the effect is much higher than in the study by Holmlund et al (2011). Whereas the latter found a £1,000 increase to lead to an increase in age 11 attainment of about 0.03-0.05 standard deviations, Gibbons et al find an increase of around 0.25 standard deviations. The main reasons for this difference are two-fold. Firstly, the sample Gibbons et al are using refers to schools in urban areas with many disadvantaged pupils whereas Holmlund et al use all schools in England. Even in the Holmlund et al study, effect sizes were higher for disadvantaged children (by 50-100%). Secondly, the methodology is very different. The Gibbons et al. study has the stronger methodology and shows that without use of a credible identification strategy, estimates for the effect of pupil expenditure show severe downward bias. This is because a high component of how resources are distributed to schools is compensatory. Thus, it is likely that the strategy used by Holmlund et al does not go far enough to remove this source of bias.

There are fewer good studies that look at the impact of class size in the UK. The NPD does not allow one to observe pupils at classroom level (only the number in the year group). Hence the data is not appropriate for investigating this issue. A number of studies have looked at the relationship between class size and later outcomes using cohort studies and found there to be little impact (e.g. Dearden et al., 2002). Iacovou (2002) re-examined this issue using the National Child Development Survey (which relates to a cohort of children born in 1958). She shows that class size and school size are positively related and that for any given size of school, average class sizes in infant schools are larger than in 'combined' primary schools (i.e. schools catering for wider age range). She argues that the interaction between school type and school size can be used as for a predictor of class size at age 7. Her findings show a strong relationship between class size and reading. The estimated effect is around 0.29 standard deviations for a reduction in class size of eight pupils. This is in line with the higher end of effects found in the international literature (discussed below – Angrist and Lavy, 1999; Krueger, 1999). On the other hand, she did not find there to be a significant relationship between class size and maths scores.

Blatchford et al (2002) also investigate the relationship between class sizes in early years (i.e. reception) and age 7 outcomes. However, their study relates to more recent cohorts (starting school in 1996 and 1997). They use multi-level models, which involves examining the relationship between class size and educational attainment after controlling explicitly for potentially confounding factors and taking account of the hierarchical structure of the data. They find class size effects which they view as impressive – particularly for children of low ability. Interpreting the magnitude of effects (with some further extrapolation), they suggest that a decrease of class size of 10 to below 25, is associated with a gain of about one year's achievement for the lowest achieving group and about 5 months for other pupils. This estimate relates to literacy, although the authors also find there to be a strong relationship between class size and attainment in maths. With regard to the estimates cited here, the authors say that they are rough and should be treated with caution.

Studies which make use of 'natural experiments' to uncover the relationship between resources and attainment often give useful insights. One such study by Machin et al (2007) looks at the impact of ICT funding per pupil on average attainment at the end of primary school. They make use of a change in the rules governing ICT funding at the Local Authority level. Their results show that a doubling of ICT funding per pupil led to an increase in the proportion of students reaching level 4 or above in English and Science by 2.2 and 1.6 percentage points (whereas the effects were only 0.2% and not statistically significant for Maths). Although the paper cannot shed light on theoretical channels through which ICT funding raises student achievement, they argue that the reason for these fairly large impacts is

that there was a significant redistribution of ICT funding (as well as an overall increase) to more efficient Local Authorities. This stands in contrast to many international studies looking at the relationship between ICT and attainment, which often show no effect.

A broad range of disparate programmes in England under the umbrella title of the National Strategies have received evaluations of various types (in the early years, primary and secondary phases), but none offer a general quantitative assessment of the benefits of the programmes compared to the costs, or the gains per unit of expenditure. Many of the interventions under these programmes are pedagogic or organisational and the resource implications are unclear. One specific strand – the School Improvement Budget paid to LAs – cost £363 million in the last year of the programme, around £50 per pupil. [7] The general summary of subsequent evaluations of the National Strategies from 2007-2011 (DfE, 2011) makes very bold claims that 'investment in the National Strategies has paid major dividends'. However, the general story is told by simply referring to changes in trends in outcomes, and the underlying reports are based on case studies, small scale surveys and qualitative evidence, with no serious attempts to understand the causal impact of the policies. As noted by Ofsted (2010) there is 'little evidence of systematic robust evaluation of specific National Strategies' (p.5). A review of this work is beyond the scope of the current report. One rigorous study (Machin and McNally, 2008) looks at the Literacy Hour component of the National Literacy Project, which was an early pilot of the National Literacy Strategy. They find a significant impact, with a 2-3 percentile (0.06-0.08 standard deviation) improvement in the reading and English skills of primary school children exposed to the policy, compared to children in appropriately selected comparison schools. The policy involved introducing a dedicated, structured hour for literacy and costs of this policy were very low, estimated at only £25 per pupil per year.

Another English primary school programme that has been evaluated is the London/City Challenge programmes which targeted a number of interventions at three metropolitan areas – London, Manchester and the Black Country (see also Section 4.1 below), mostly aimed at low-performing schools. Hutchings et al (2012) report improvements in performance (both on tests and Ofsted inspection ratings) in these metropolitan areas relative to other metropolitan areas and the national averages. They also report positive gains for the schools within each area that were targeted as low-performing. However, their quantitative analysis is fairly unsophisticated in terms of ensuring that treated and non-treated schools are compared on a like for like basis, there is no indication of the costs of the programme, and overall there are few lessons to be learned about the impacts of resources more generally.

We now consider the international evidence more broadly.

3.3 International evidence from developed economies

There is a huge volume of work about the effects of school resources on pupil attainment – particularly in the US, but increasingly in Europe. There are very different views about how to interpret the evidence. This was strikingly portrayed in the papers published by Eric Hanushek and Alan Krueger in the Economic Journal (published in 2003). Hanushek's (2003) view is based on a meta-analysis of 89 studies published prior to 1995. He argues that, taken a whole, the literature suggests little or no impact between increasing resources (measured in various ways) and improving educational attainment. This view has been very influential and is commonly cited in the literature. On the other hand, others (such as Krueger, 2003) argue against the 'vote counting' used in this methodology and suggest that most attention should be given to studies with the strongest methodological design.

---

[7]

http://www.education.gov.uk/schools/adminandfinance/financialmanagement/schoolsrevenuefunding/section25
1/archive/b0068383/section-251-data-archive/budget-data---summary-level

However, there are studies with a strong methodological design that have found completely opposite findings (although in different contexts). So there is no end to the controversy as to how one should interpret the literature and weight the different studies. However, there is only one study with the 'gold standard' randomized design. This is the Tennessee 'STAR' (Student/Teacher Achievement Ratio) experiment, which was a large scale randomized trial of lower class sizes for pupils during their first four years in school. The first phase of the study ran from 1985-1989. In this study students and teachers were randomly assigned to a group of 'regular size' (22-25 students, to another group of regular size including a teaching assistant or to a small group (13-17 students)

There have been many papers about the experiment and Schanzenback (2007) provides a good summary of the findings. It was found that students benefited greatly from being allocated to the smaller class compared to the regular-sized class (a reduction of about 8 pupils). The effects were about 0.15 standard deviations in terms of average maths and reading scores (measured after each grade for the four years). The effects were much greater for black than for white students (i.e. 0.24 standard deviations versus 0.12 standard deviations) and this was primarily driven by a larger treatment effect for all students in predominantly black schools. There was also a differential (although less stark) between disadvantaged students (i.e. eligible to receive a free lunch) and other students. In third grade, students eligible for a free lunch gained about 0.055 standard deviations more than other students. In fourth grade, all students went back to regular sized classes. In grades 4 to 8, there continued to be a positive impact of initial assignment to a small class. However the magnitude of the gain reduced to one-third to one half of the initial effect. Again, the impact remained stronger with black and disadvantaged students.

Recently Chetty et al (2011) look at much longer term effects of the STAR experiment. They link the original data to administrative data from tax returns, allowing them to follow 95% of the STAR participants into adulthood. They find that students assigned to small classes are 1.8 percentage points more likely to be enrolled in college at age 20 (a significant improvement relative to the mean college attendance rate of 26.4% at age 20 in the sample). They do not find significant differences in earnings at age 27 between students who were in small and large classes (although these earnings impacts are imprecisely estimated). Students in small classes also exhibit statistically significant improvements on a summary index of other outcomes (home ownership, savings, mobility rates, percent college graduates in ZIP code and marital status).

There has been no class size experiment as thoroughly investigated as the STAR experiment. However, there have been various other credible strategies used to identify class size effects and they do not always come up with results that are consistent with this evidence. One of the strategies used has been to use demographic variation across year groups within a school to identify class size effects. Hoxby (2000) was the first to use this strategy. Specifically, she exploits the idea that (after controlling for a trend) cohort sizes within school districts can be larger or smaller in some years than in others. Using data on elementary school pupils in the state of Connecticut, she is able to rule out even modest effects of class size on pupil attainment. Rivkin et al (2005) employ a similar approach to look at schools in Texas. While they find small class size effects, the magnitude varies across grades and specifications. Cho et al (2012) apply Hoxby's method for students in Minnesota (grades 3 and 5). They also find very small effects. They estimate that a decrease of ten students would increase test scores by only 0.04 to 0.05 standard deviations. Another fairly recent paper (Sims, 2009) looks at the effect of class size reductions in California. However, in this case, he uses a quasi-experiment, where some schools were forced to increase the class sizes of later grades to facilitate the required reduction in class sizes in earlier grades. His estimates for grade 5 are closer to the high end of estimates in the literature, but much smaller

(half the size) for grade 4. Jepsen and Rivkin (2009) look at the effects of reducing class size in California on the cohorts directly affected (rather than those at later grades within the school). In their analysis, a ten student reduction in class size is estimated to raise average achievement in maths and reading by 0.10 and 0.06 standard deviations respectively. However, these effects can be completely negated by the effect of having an inexperienced teacher (i.e. a first year teacher as opposed to a teacher with at least two years' experience). This finding points to a consequence of extensive class size reductions in the real world (where other things are not held constant) − a lot more teachers needed to be hired in California to facilitate a class size reduction across the state by roughly ten students per class. On the other hand, this could be just a transitionary problem of moving to lower pupil-teacher ratios, and not one that would persist in the medium term.

There have been several studies about the effects of class size (in primary schools) outside the US. Perhaps the best known is by Angrist and Lavy (1999) for Israel. These authors were to first to use rules on maximum class size to conduct a 'quasi-experiment' about the effect of lower class sizes on pupil attainment. In Israel, a maximum class cannot exceed 40 pupils. Variation in the size of an enrolment cohort generates discontinuities in the class size attended by students. They find large effects of class size on the educational attainment of students in the fourth and fifth grade. They compare their results with the Tennessee STAR experiment by calculating the effect size for a reduction in class size of eight students. Their estimates suggest impacts of about 0.13 and 0.18 standard deviations for test scores in grades four and five respectively. Piketty (2004) uses a similar methodological strategy for France. In the French case, when second-grade enrolment goes beyond 30, another class is opened (in most cases). Hence, the two new classes have an average size of 15 pupils. Piketty uses this discontinuity as an instrumental variable (i.e. predictor of class size differences). He finds that a reduction in class size induces a significant and substantial increase in mathematics and reading scores, and that the effect is larger for low-achieving students. Bressoux et al (2009) use administrative rules in France to argue that effects of class size can be estimated in a sub-sample of relatively inexperienced teachers. They find effect sizes that are close to those found in the Tennesse STAR experiment. Furthermore, they find that the effect size is higher for classes with a low initial achievement and also in areas of high socio-economic deprivation.

Lindahl (2005) estimates class size effects for 5th grade students in Sweden. He tries to identify the effect of class size on achievement by taking the difference between school and summer period changes in test scores. He finds that reducing class size by one pupil gives rise to an increase in test scores by at least 0.4 percentiles. He also find that immigrants' children benefit more from smaller maths classes. He argues that the magnitude of the class-size effect and the result that some disadvantaged groups benefit more from smaller classes are in line with the results of Angrist and Lavy (1999) for Israel and the STAR experiment for the US (Krueger, 1999).

A number of recent papers have looked at the effect of school funding (rather than class size) in primary schools. In most cases, funding per pupil has to be measured at the level of the school (or district) rather than at the level of the class. Guryan (2001) examines the effectiveness of public school spending the context of an effort to equalize funding across school districts within Massachusetts. In particular he looks at variation in funding caused by two aid formulas. He finds that districts that received large increases in state aid as a result of the equalisation scheme had a fairly large increase in 4th and 8th grade test scores. Specifically, his point estimates suggest that a $1,000 increase in per pupil spending (about one standard deviation) is associated with a 0.3-0.5 standard deviation increase in test scores (although 8th grade results are more sensitive to specification). Chaudhary (2009) investigates the impact of a school finance reform in Michigan. He looks at the effects on test scores in

the 4th and 7th grade and finds that effects are only significant for the 4th grade. The estimates suggest that a 10% increase in spending ($580 on average) would increase 4th grade maths scores by 0.10 standard deviations. With the use of another measure, he suggests that a 60% increase in expenditures would increase scores regarded as 'satisfactory' by one standard deviation. Chaudhary suggests that an explanation for the differential effects across grades could be due to targeting of resources within schools or that younger students are more responsive to changes in inputs. However, these data do not allow further exploration of this issue. The findings here are consistent with the earlier study by Papke (2005) about finance reforms in Michigan. She also looks at students in 4th grades and finds large effects of expenditure on the pass rate in the Maths test. She says that a rough rule of thumb would be that 10% more real spending increases the pass rate by between one and two percentage points, and more for initially underperforming schools.

While the above studies about the effects of school finance reforms suggest a positive impact of school resources, this is not always the conclusion in the recent literature. For example, Matsudaira et al (2012) is the most recent of many studies looking at the effect of 'Title 1' (i.e. the biggest US Federal government programme targeted towards primary and secondary education). As they discuss, most of the literature on the effects of this programme finds no effect of increasing resources attributable to this funding stream. However, Gordon (2004) found that state and local governments adjust their funding levels in response to the federal grant. Matsuidaira et al (2012) re-examine this issue in a large urban school district. They also find that the federal grant is partly offset by a decrease in funding from other sources. They say that 'given the high variation in per pupil expenditures even among very similar schools, however, Title 1 eligibility results in no statistically significant increase in total direct expenditures'. They also find that Title 1 has no impact on overall school-level test scores and suggest that this is unsurprising given the small amounts of money involved. However, they do not find any impact in the subgroups of students most likely to be affected.

Lavy (2012) is unusual for linking expenditure to particular classes (in 5th grade). He uses a particular experiment in Israel about changes in funding rules and shows that this is linked to the length of the school week and with instructional time in different subject areas (Maths, Science and English). His results suggest a modest effect of the policy. For example, increasing instructional time in each subject by one hour per week increases the average test scores in these subjects by 0.053 standard deviations. The effects on students with parents who have below average education are twice as large in maths, 25% higher in science, but 25% smaller in English compared to those with higher than average education. Lavy argues that providing two or three additional hours of maths instruction per week to the low ability group would go a long way to narrowing the gap between socio-economic groups.

Finally, Leuven et al (2007) investigate the impact of two specific subsidies that were targeted at primary schools with large proportions of disadvantaged students in The Netherlands. One subsidy provided extra resources to improve teachers' working conditions and another subsidy provided additional funding for computers and the internet. The authors make use of discontinuities in the entitlement of schools to receive this funding in order to identify the effects. They did not find positive effects of the subsidies in either case. The personnel subsidy was mainly spent on extra payments for current teachers and hiring extra teachers. Leuven et al. interpret the negligible effect of this policy as attributable to (1) the fact that the extra payment was not conditioned on performance, and (2) schools targeted by the personnel subsidy may have already had sufficient numbers of teachers in place (the pupil-teacher ratio was already below 14 in such schools). With regard to the computer subsidy, they argue that traditional instruction methods might be more effective that methods using computers. They point to several other economic studies that also come to this conclusion (Angrist and Lavy, 2002; Goolsbee and Guryan, 2006; Rouse and Krueger, 2004).

4. Secondary Phase

4.1 UK evidence

A survey of evidence for the UK (and internationally) was provided by Vignoles at el (2000). This report summarised the state of play of evidence at the turn of the century, using evidence drawn mainly from the 1990s. Much of research cited in the review addresses slightly different questions regarding the cost effectiveness of different school types, or of different experimental pedagogic interventions. Many of the studies on the general effects of resources that are cited use methods that do not meet appropriate criteria in terms of establishing causal links.

The most reliable British research they cite that provides direct evidence on the impacts of additional resources is based on student level data from British birth cohort studies (National Child Development Study – NCDS and British Cohort Study – BCS). The schooling data relates to age 16 during the 1970s. Despite the potential lack of contemporary relevance, the advantages of these cohort studies is that they allow investigation of impacts on schooling on later life outcomes, such as earnings. The summaries in the review indicate that only one of the four studies finds stable evidence that lower pupil-teacher ratios (at secondary school level) improve exam results and those that look at spending (at LEA level) find no effect. There are impacts from pupil-teacher ratios when considering differences across school types (private, grammar, comprehensive etc.) but not when considering differences in PTRs between schools of a given type (e.g. comprehensives). One of these articles (published as Dustman et al (2003) reports effects from pupil teacher ratio reductions on the probability of staying on at school (and hence on subsequent wages). Another study using the same data source also finds impacts of lower student-teacher ratios on wages, but only for women (Dearden at al, 2003). The wage effects are moderate: one less pupil per teacher (from a mean of 17) increasing wages by around 1%. A limitation of this work based on the NCDS is that, despite the rich data source and use of student level data, these are cross-sectional, educational production function based estimates, and the resource effects are not estimated from any specific policy driven difference or change in resources (although they control for a wide range of student background factors).

More recent evidence on secondary school expenditure or pupil-teacher ratios in the UK is relatively scarce. Two reports were produced for the then Department for Education and Skills, by Jenkins at al (2006a and 2005b). Like a lot of recent work on education in England, the authors use administrative data from the National Pupil Database in England. They look at effects of additional spending at school-level on student's performance at age 14 (key stage 3) and age 16 (GCSEs). Both studies find small positive effects from general spending (or PTRs) on attainment in science at both ages, effects on maths at GCSE, but no effects on English either age. The order of magnitude of these effects is around 5-6% of one standard deviation in test scores, for a one-standard deviation increase in expenditure (about £300-£400 in early 2000s prices) with little difference across pupil types. In forming these estimates, the authors predict spending from the political control of the Local Authority in which the school is located in order to try correct for the reverse linkages between school disadvantage and resourcing. This design (which is common in many fields of research) is potentially limited by the fact that voting behaviour may be related to student achievement through population demographics (and hence voting) rather than school expenditure. A very recent study uses unique information on siblings, imputed from address information in the National Pupil Database. By comparing outcomes for siblings exposed to different levels of education expenditure, Nicoletti and Rabe (2012) are better able to control for family background factors, and they find significant but very small impacts from expenditure on progress between Key Stage 2 test scores at age 11, and GCSE achievement in secondary

school. A permanent £1000 increase in expenditure per student raises achievement by about 0.02 standard deviations. A potential limitation of this approach is that 85% of siblings attend the same school, but in different years, so the study is effectively estimating the effect of marginal changes in expenditure from year to year within a school. As will become clear throughout this review, studies that adopt designs based on short run changes over time of this type tend to find small impacts. However, a useful feature of the design in this case is that the findings can be compared directly to those in Holmlund et al (2011) which looks at primary school achievements in the same educational system using the same data set. The estimates of the resource impacts in the two phases are of a similar order of magnitude when using the same methods (around 0.05 standard deviations for £1000 increase), although slightly smaller at the secondary phase when comparing siblings.

Class size effects in the UK are investigated by Denny and Oppedisano (2010) using PISA data on mathematics tests for 15 and 16 year olds, and exploiting year on year changes in school-specific cohort size as a source of random variation in class sizes. They also look at the US data in this analysis. Their findings are that mathematics scores are better for students in *bigger* classes, with one extra student in a class of 25 raising scores by 0.07 standard deviations. These results imply that expenditure on class size reductions is counterproductive. However, their research design assumes that the year to year changes in a school's enrolment are not partly determined by changes in school quality (and hence student scores).

As discussed in Section 2, the modern approach to investigating these kinds of questions requires explicitly defined changes or differences in resourcing from which to estimate the impacts, for example from resource-based policy interventions. For secondary schooling in England, Machin et al (2010) look at the effects of the 'Excellence in Cities' programme which allocated extra resources (about £120 per pupil per year in the early 2000s) to some secondary schools in disadvantaged urban areas in England. They find evidence of benefits from the programme in mathematics and on attendance at age 14, but not on English, but with variation in the effects across pupil and school types. The biggest effects are concentrated on medium to high ability pupils in the most disadvantaged schools. Bradley and Taylor (2010) in a study of the impacts of various policies on GCSE performance also report beneficial effects from Excellence in Cities in disadvantaged schools, with a 3 percentage point improvement in GCSEs for participating schools.

Another major programme in England is the Academy programme, in which new Academies were built to replace failing schools in disadvantaged areas The early stages (up to 2009) of this programme under the Labour government has been evaluated in Machin and Vernoit (2011). They compare average educational outcomes in schools that became academies and similar schools, before and after academy conversion took place. There are three main findings. Firstly, schools that became academies started to attract higher ability students. Secondly, there was an improvement in performance at GCSE exams – even after accounting for the change in student composition. Thirdly, to an extent, neighbouring schools started to perform better as well. This might either be because they were exposed to more competition (and thus forced to improve their performance) or it might reflect the sharing of academy school facilities (and expertise) with the wider community. The National Audit Office in a less sophisticated evaluation also found evidence of improving performance in the new academies (NAO 2007). However, all these results relate to the early phase of the programme which was targeted at disadvantaged areas and students, where the potential gains are greater (as documented elsewhere in this survey). The Academies programme has been significantly widened, with any state school in England now able to apply for academy status and it is as yet, too early to evaluate the impacts of this new model. Although this programme clearly involved a commitment of extra resources, it is not a simple resource-based intervention, and assessing the additional costs in terms of expenditure per pupil is not

possible, given that there are high initial capital costs and that the on-going expenditure differences between non-academy schools are not clear cut.

As for primary schools, Hutchings et al (2012) report improvements in performance for secondary schools in the London/City Challenge programme, but as discussed for primary schools (Section 3.2 above) there are few clear lessons about the effect of general resources from this study.

Slater, Davies and Burgess (2009) take a different approach that estimates the overall contribution of teacher quality to the distribution of children's achievements at GCSE. This study follows methods developed in the US literature – see section 4.2 below – and uses a dataset on Bristol students that is unique for the UK in providing linked student teacher data. Teacher 'quality' here is measured by teachers' persistent ability to achieve test score gains for different groups of children in different years. Slater Davies and Burgess find (like the US studies) that teacher quality accounts for some of the variation in student scores. A one standard deviation increase in teacher quality yields a 0.3 standard deviation increase in test scores according to their estimates. However, these differences in teacher quality are not explained by teacher salary, or differences in qualifications or experience, factors which generally determine pay, and hence teaching resource costs. The finding is thus consistent with other strands of evidence that finds little or no impact from financial resources on student achievement.

4.2 International evidence from developed economies

The international literature is extensive. Across countries, the OECD Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) are popular data sources for these analyses, but there are many other studies using country-specific administrative and survey data, with a large body of evidence for the US in particular. A popular approach is to use (arguably) random variation in school enrolment between cohorts (grades) to estimate the effect of average class size changes on achievement at secondary level. Woessmann and West (2006) find mixed evidence on mathematics and science scores of 13 year olds in a sample of 11 countries (excluding UK) in the TIMSS data with significant beneficial effects in Greece and Iceland, but zero or inconclusive results elsewhere. They argue that the lack of any general effects of class size reductions could be because the magnitude of the effects is dependent on the educational system. Another related study on a bigger sample of European countries from TIMSS (Woessmann, 2005) also finds a mix of effects, with Iceland again the only country showing clear beneficial impacts from smaller classes, but no evidence across countries in general that resources spent on class size reductions are productive. Similarly mixed findings emerge in Altinok and Kingdon (2012) from the TIMSS data, who investigate the effects of differences in subject-specific class sizes in a student's achievement across subjects. This method has the advantage of controlling for omitted pupil variables that are common across all subjects. Using this method they find evidence of significant but very small beneficial class size effects for a number of Eastern European and Developing countries, and in the Netherlands (where a 1 student reduction raises achievement by less than 0.01 standard deviations), but in general the results are zero and insignificant.

Turning to country specific studies, Heinesen (2010) looks at the effects of year to year variation in French class sizes in Danish schools, from grades 7 to 9 (13-15) in 2002-4. He argues that focussing on a single language subject mitigates the selection problems induced by parents switching schools in response to class size or quality differences. This design assumes that parents do not base such choices on single subjects, that students do not choose to study a particular a language based on anticipated class quality, and that variation in French class sizes over time within a school is due only to random variation in the numbers

of students choosing to study French rather than German. The evidence from Heinesan's work is that an extra student in a class reduces students' test scores by 0.03 standard deviations, and the effect is bigger for lower ability/disadvantaged students and boys. A series of 'placebo' results on the effects of French class sizes on other subjects is reassuring in showing no significant effects.

As in the primary school literature, class size rules have been used to implement regression discontinuity designs to estimate that effects of additional teaching staff resources following Angrist and Lavy (1999). Bonesronning (2003) uses 30 student class limits in Norway to find that students surveyed in lower secondary schools (age 13-16) did better on tests if they were in smaller classes, but only marginally. A 1 student reduction in class sizes increases test scores by 0.01 standard deviations, although there is some variation in response across different student types. Leuven Oosterbeek and Ronning (2008) provide related evidence from administrative data in Norway and exploiting the class size rules, and population variation over time, and find effects of a similar order of magnitude, although not statistically significant. A number of studies, some described in Section 3.3 above and in Gary-Bobo, Mahjoub and Badrane (2006), use similar methods on French data with moderate impacts on achievement in grade 9 reported from Piketty and Valdenaire (2006): a 10 student reduction in class size increases achievement by about 0.2 standard deviations. In their own research, Gary-Bobo et al look at the effects of class size on grade repetition in a sample of French students from grade 6 through to grade 9, they find moderate beneficial effects of class size reductions in primary school, but find no effect in junior high school (grades 8 and 9), suggesting that class size stops being so important for mature students. Bingley Jensen and Walker (2007) looks at the effect of uses the class size rules in Denmark on length of post-compulsory schooling and do find significant but small benefits from smaller class sizes. A one pupil reduction in class sizes (from a mean of 20) is linked to a 1% change in the length of compulsory schooling – which amounts to 8 days on average. Interestingly, they translate this into an economic return in terms of gains in lifetime earnings, and come to a figure of about £3500 for men and half this for women (30,000 DKR) and estimate that this is about equal to the costs per person of implementing such a reduction in class sizes. A number of papers in the US refer to mandated class size reduction (CSR) programs that induced sudden changes to class sizes, but only Chingos (2012) looks at secondary schooling. He estimates whether mandated class size reductions in Florida had any impact on test scores, by comparing districts with different initial class sizes, and hence different mandated class size reductions. The conclusion is that CSR in Florid had no positive effect on performance in achievement through grades 6-8.

Haegeland, Raaum, and Salvanes (2012) report relatively large positive effects from school expenditures and teacher hours in Norway, using tax revenues raised from hydro-electric plants as a source of quasi-experimental variation. The idea here is that school funding is drawn from the local tax base, and hydro-electric plants result in a bigger local tax base. Given that the geological processes that lead to an area being suitable for a hydro-electric plant are unlikely to have a direct effect on student achievement, this higher funding can be treated as a random. If families choose where to live in order to access schooling, then the 'sorting' of different families into different school districts might undo this randomness (e.g. if higher ability families pick the better funded districts), so the authors endeavour to show that this does not affect their results. A 30% increase in funding (NOK 18,000, about £1800) is associated with 0.28 standard deviations higher achievement at age 16, although the research cannot say through what mechanisms this improvement occurs. This contrasts quite sharply with the lack of evidence on class size (and implicitly resource effects) in Leuven et al (2008). As noted already in relation to other studies, the difference may stem from the fact that the Leuven et al design is based on changes over time in enrolment within the same

school (coupled with class size rules), whereas the Haegeland et al study is estimating from cross-sectional differences between municipalities. Interestingly, the Hegeland et al study using cross-district differences in funding in Norway arrives at similar estimates to Gibbons et al (2011) who investigate differences across schools in neighbouring districts in England.

A number of studies use school finance policy reforms as a source of variation in school expenditures. Chaudhary (2006) studies the impact of a finance reform in Michigan ('Proposal A') which increased teacher salaries and reduced class sizes. Despite evidence of effects on 4th grade (primary schooling) there is no evidence of effects at 7th grade (secondary, age 12-13). Although not exactly an expenditure related intervention, Hakkinen , Kirjavained and Uusitalo (2003) investigate the impacts of large changes in school expenditure during the 1990s recession in Finland using a school fixed effects design (again based on changes over time in expenditure within schools) and but report no significant effects from teaching expenditures on senior secondary test scores. Guryan (2001), described in detail in Section 3.3, looks at the impact of Massachusetts finance equalisation policy on achievement in grade 8. In contrast to the effects found in primary school, he concludes that there are no effects at this secondary school grade.

There are no recent explicit experiments in class sizes reductions or general resource-based interventions in secondary school that are comparable to the Project STAR experiment described in Section 3.3 (at least as far as we have been able to ascertain). A number of early experiments were carried out in the US in the 1920s and 1930s to assess the potential impacts of class size increases, and have been recently uncovered and described in Rockoff (2009). The general findings of these experiments, which included 5 studies of around 1800 students at secondary school, was that class size had neglible effects on student performance, in sharp contrast to the findings of Project STAR. Potential reasons for this difference are discussed in Rockoff (2009), and may include weaker methodology in the early studies, that the early experiments excluded the kindergarten age group, or that the relationship between class sizes and achievement has changed over time.

A related strand of research has shown that variation in teacher quality, within schools is one of the more important factors affecting students' achievements (Hanushek and Rivkin, 2012 & 2010b; Rivkin, Hanushek and Kain, 2005). These studies typically show that the variation in teacher components represents about 0.10 to 0.2 standard deviations of the student test score distribution (implying that teachers account for 1-4% of the variance in student scores). As with other literature on school resources, it turns out to be hard to pin down specific resource-related factors that explain these differences between teachers. For example, Aaronson, Barrow and Sander (2007) find that a one standard deviation increase in teacher quality raises 9th grade student achievements by about 0.25 standard deviations in Chicago schools, but that teacher experience, qualifications, tenure and demographics (which often determine pay, and hence resource costs) do not explain these differences between teachers. Koedel (2007) has similar results for secondary school students in San Diego (as do Koedel and Betts, 2007 for primary schools in San Diego). A strong criticism of these teacher quality findings is, however, that it does not control for the possibility that some teachers are persistently assigned the best or worst performing children over the lengths of the sample, and test score gains (value-added) are potentially a bad way to evaluate individual teachers (Rothstein, 2010).

5. Work on Less Developed and Developing Countries

Similarly to his work for developed countries, Hanushek (2006) has conducted meta-analyses of studies about the effects of resources on educational attainment in developing countries. A

comprehensive review of work from 1990 to 2010 is provided by Glewwe, Hanushek, Humpage and Ravina (2011). This analyses suggests a similar inconsistency of estimated resource effects as that found in the US and other developed countries. However, a major concern with work on developing countries is study quality, as many researchers do not have access to longitudinal data on individuals, and methods are sometimes below the currently accepted standards. Starting from an initial pool of 253 papers which estimated the impacts of school and teacher characteristics, Glewwe et al end up with only 43 which they consider 'high quality' on the basis of the methods and data used. A potential strength of the work on developing countries is that there are many examples of field experiments – randomized controlled trials to investigate the effects of specific interventions (see Kremer, 2003 and Glewwe et al, 2011). These RCTs have a strong advantage over studies based on non-experimental data in dealing with omitted variable biases of the type discussed in Section 2. However, the range of interventions investigated is diverse – e.g. flip charts, computers and computer assisted learning, school meals - and the findings mixed, meaning it is again impossible to draw any general conclusions, especially in relation to policy lessons for developed countries.

One study that is similar to those for developed countries is by Urquiola (2006) and relates to educational attainment of third grade students in Bolivia. He is interested in the effects of class size and estimates effects using two strategies. The first approach focuses on variation in class size in rural areas with fewer than 30 students and hence only one classroom per grade. The second approach is similar to that used by Angrist and Lavy (1999) for Israel. This exploits regulations that allow schools with more than 30 students in a given grade to obtain an additional teacher (thus generating discontinuities in the class size variable). Both strategies suggest a sizeable impact of class size – particularly the effect estimated from the discontinuity (which gives rise to large changes in classes). In this case, a one standard deviation reduction in class size (approximately 8 students) raises scores by up to 0.3 standard deviations. He suggests that the effect of class size could be non-linear and that large effects might be attributable to the fact that hiring an extra teacher enables tracking (streaming/setting) in schools. In contrast, a study based on class-size rules in Bangladesh (Asadullah, 2005) finds that smaller class sizes are inefficient, leading to lower secondary school grades. Urquiola and Verhoogen (2009) present evidence for Chile that suggests beneficial effects of small 4th grade class sizes for test scores, using maximum class size rules. However, free choice amongst schools in Chile (which has a voucher system) and the strategic behaviour of schools in limiting enrolment to multiples of the maximum class size leads to sorting of children across schools with different enrolments, implying that the schools with the bigger classes end up enrolling students from more educated parental backgrounds. This violates the intentions of the research design, and Urquiola and Verhoogen conclude that the class size impacts in Chile are not necessarily causal.

Ultimately, it is quite difficult to draw conclusions from developing countries with regard to the impact of school resources in developed countries, because of the major differences in institutional context. In a review of the literature about randomised experiments, Kremer and Holla (2009) argue that supplying more of existing inputs, such as teachers or textbooks, often has a limited impact on student achievement because of distortions in developing country education systems. These distortions include elite-orientated curricula and weak teacher incentives, as manifest through a high level of absenteeism. They argue that pedagogical innovations (e.g. technology assisted learning or standardised lessons) that work around these distortions can improve student achievement at low cost. In a broader review about education in developing countries, Glewwe and Kremer (2006) suggest that the most effective forms of spending are likely to be those that respond to inefficiencies in schooling systems. For example, remedial education may be extremely effective in an environment in

which many students fall behind and are no longer able to follow teachers' lessons; providing computer-based education may be effective when teachers attend irregularly. They caution against a false dichotomy between the view that schools in developing countries need more money and the view that they need to reform. The two are not mutually exclusive.

6. Country, State and Regional Level Studies

A vast literature on economic growth has demonstrated relationship between education in the workforce (e.g. average years of schooling) and subsequent economic growth at the country, state and regional level. More recent evidence suggests that the level of cognitive skills in the workforce (e.g. average test scores in secondary school) does an even better job of predicting subsequent growth literature. This is a broad literature in its own right, and a full survey is beyond the scope of this paper, but the key literature and evidence based on the cognitive skills measures from the OECD PISA data is described in Hanushek and Woessman (2008). This link between education and growth at the aggregate level raises the possibility that the quality of schooling – and not just the quantity – could be an important driver of growth, although in itself this does not provide evidence on whether investments in schooling have any effect on school quality.

Another strand of literature does try to tackle this question directly, by estimating the impacts of educational spending on the economy, aggregate income and on growth. An advantage of research that looks for impacts at the aggregate level is that they can capture 'macro' level benefits to the economy that go beyond the sum of the individual benefits that would be measured in micro-level studies. As an example, a general increase in educational quality in a country or region that benefits an individual student through improvements in their own education and pay, may have additional benefits in terms of labour market earnings and income working through spillovers from other individuals in the local labour market or wider economy. These effects are not easily detected from micro level studies into the effects of spending on individual students or schools, rather than the local region or economy as a whole.

These types of study typically find positive associations between spending on primary and secondary education, and growth rates. However, a limitation of this line of research, in common with similar research that tries to estimate the effects of of government investment and infrastructure spending, is that it is very difficult to disentangle the direction of causality. It is generally harder to come up with research designs that can easily separate out the causal links running from education spending and economic performance, from more general changes that affect both of these simultaneously (e.g. if higher income growth leads to a greater public spending on schooling). These studies should therefore be interpreted with some caution, and we do not review them further here.

7. Discussion, Synthesis and Conclusions

Research on the effectiveness of additional resources in schools has moved on some way since the early 2000s, with a far greater number of high-quality research designs, a better understanding of the challenges to causal estimation, and better data. Ten years ago, the two prevailing interpretations of the evidence were succinctly articulated by Hanushek (2003)- 'resource policies have not led to discernible improvements in student performance' - and Krueger (2003) – '…reanalysis of the literature suggests a positive effect of smaller class sizes on student achievement, although the effect is subtle and easily obscured…'.The

position now regarding the more recent evidence is somewhat more in favour of the second interpretation, with a far greater number of studies finding evidence of positive resource impacts although even studies with good research designs and high quality data still produce a variety of sometimes conflicting results on the effects of resources.

Whether these impacts are small or large is arguable, because it depends what scale of impact one is expecting. The scale of the effects appears small when judged against the overall variation in student achievement. The usual benchmark figure is from the Project STAR programme, where a class size reduction of 8 students (and the resources this entails) is associated with around 15-20% of one standard deviation improvement in test scores. A comparable figure comes from our own work on English primary schools, where we find that a £1000, or 28% increase in expenditure (which could fund a 6-7 student reduction in class sizes) is associated with a 25% standard deviation improvement in test scores (Gibbons et al, 2011). Other studies find similar, or smaller effects. These impacts are not enormous compared to the overall variance in achievements, but it should be remembered that countless studies have demonstrated that most (more than 90%) of the variation in student test scores is due to family background, parental inputs, natural student abilities and purely random variation, so it should not be too surprising if the impacts of resource changes are relatively small by comparison. Where approximate cost benefit analyses have been carried out, these are usually favourable. For example, Machin and McNally (2008) in their evaluation of the Literacy Hour strategy in England, estimate a labour market return of 0.42% to a one percentile increase in test scores at age 10 (using data from children raised in the 1970s and 1980s), implying that a 0.20 standard deviation change in test scores would raise earnings by about 2.4%. Using the 2011 median earnings (£400 per week) and employment rate (70%) implies that average earnings are around £15000, so this change in test scores is worth around £6200 in present value terms at age 10 (assuming an additional £360 is earned in each year of a person's working life from age 16-65, and discounting back to age 10 using a discount rate of 3.5%). Therefore any investment that raises child achievement by 0.2 standard deviations at a cost of less than £6200 per child (in present value terms, over all years of schooling up to age 10) is worthwhile in terms of future labour market earnings alone. This is equivalent to a £800 per pupil increase in spending in each year between ages 4 and 10.

One limitation of the existing research is that it is primarily about cognitive skills as measured by in-school test scores. A few studies have looked at longer run outcomes like staying on rates and earnings, but this research is data intensive, requiring linked data on student schooling, post school education, and labour market outcomes. Relatively few studies have looked at the long run impacts of school resources, because this analysis can only be applied on life cycle panel data such as the cohort and household panel studies in Britain and the US. This is an important area of future investigation. The recent long-term study of Project STAR is important in this respect (Chetty et al, 2011).

There are some general patterns which have emerged from this review of the recent evidence on resource impacts. A first notable point is that increases in resourcing are usually, though not ubiquitously, found to be more effective in disadvantaged schools and/or on disadvantaged students at all phases (e.g. where students are entitled to free meals, or have low parental education). This may indicate that disadvantaged students are genuinely more responsive to resource based interventions and reductions in class sizes, implying that it is more efficient (as well as equitable) to target resources at these students. However, it is possible that this phenomenon sometimes arises because the potential gains at the top end are limited by the design of many school testing systems (e.g. the English key stage tests).

A second common pattern is that different research designs, although ostensibly asking the same questions, are using very different sources of variation in resources and class sizes, and tend to come to slightly different conclusions. Studies that use variation in class sizes and

resources over time, arising from population variation in cohort sizes, are working from marginal changes in class sizes and implied resources per student. These studies generally struggle to find large or even significant resource effects, compared to those that look at the large class size differences generated by maximum class size rules, or other large differences in resources between schools and classes – e.g. those induced by the STAR experiment.

One reason for the difference between the findings of these types of studies could be that responses of students to small changes are different from the responses to large changes i.e. there is some non-linearity in the response, and the impact of large changes cannot be inferred from the response to small changes in resources. A related explanation might be the schools, teachers, students and parents involved in the educational process naturally adapt to marginal changes in resources from year to year, and accommodate these changes by adjusting effort and engagement in the educational process. On the other hand, adaptation to large resource differences is less likely. This is one reason sometimes put forward for the lack of convincing evidence of resource impacts is that the individuals – teachers, students, parents – involved in the educational process change their behaviour in response to resource differences, and these changes in behaviour are unobserved to the researcher. In many cases these changes in behaviour will be compensatory, with individual substituting their own inputs as resources are withdrawn, and so tend to 'crowd out' and mask the impacts of resource differences. This is a fundamental limitation to any research based on human behaviour, when it is infeasible to make the participants 'blind' to the interventions or resource differences under investigation. For instance parents may compensate for a lack of school resources by paying for private tuition or devoting more of their own time; teachers may undo class size impacts by exerting more effort in larger classes than in smaller classes; pupils themselves may respond with greater or lesser effort. Unless researchers have detailed data on these kinds of inputs (which in the case of 'effort', is practically impossible), estimates of the effects other inputs for which they are substitutes are very likely to be downward biased. Experimental studies are not immune to these kinds of substitution effects, when participants are aware of their involvement in an experiment (so called Hawthorn effects). Although this crowding out is a problem in theory, it is unclear whether it matters in practice for educational interventions. In one study of the issues, Datar and Mason (2008) suggest that parents responded to larger classes in the Tenessee STAR experiment by more financial input, more school involvement, but less child interaction, although controlling or not for these changes does not appear to affect the STAR experimental results on class size reductions.

Whether these responses should be accounted for or not when assessing the effectiveness of policy is an open question. On the one hand, the behavioural responses of teachers, parents and pupils need to be taken into account in evaluating the causal effect of policy, because there is no value in devoting additional public resources if this simply crowds out individual effort and private investments. On the other hand, if these behavioural responses only apply to the small resource changes that underpin practical estimation strategies then this is something to be concerned about – e.g. teachers may be able to easily accommodate changes of one or two students in a class without any impact on achievement, but would respond very differently to a halving or doubling of class size. Behavioural responses to small changes may simply be masking the potential impacts of potentially larger policy-driven resource interventions. Further work is needed in this area to assess the threat that compensatory changes in behaviour imposes to empirical work on educational resources.

Other potential explanations for the differences in estimates from different studies is that some designs are simply better able to control for selection issues and omitted variables than others. It may be that designs using changes in resources over time within schools are genuinely better at comparing like with like (because they are looking at changes within a

school rather than differences between schools) and hence arrive at lower estimates than other designs. Comparing results across different countries and education systems is also problematic, given that the response to resource changes is likely to be context-dependent. This is particularly true of the studies that look at general funding changes and differences, because obviously the effect of more resources will depend crucially on how these resources are used, and research generally lacks sufficiently data to answer questions about the impact of expenditures on specific items (except where they investigate specific interventions like ICT or teaching expenditure). None of the studies we looked had sufficiently detailed data to investigate the impact of changes in multiple categories of resource expenditure, and it is hard to devise research designs that estimate the causal effects of multiple categories of resource intervention simultaneously (although Gibbons et al, 2011 provide indirect evidence on how different categories of expenditure respond to exogenous differences in school income).

Although we have highlighted this variation across studies, it is worth noting that the studies we reviewed that found evidence of statistically significant impacts, found effects in a similar *order* of magnitude, even if they differed in their exact conclusions. The smallest non-zero effects were in the order of 2-5% of one standard deviation in achievement for a 30% increase in expenditures (equivalent to roughly a 6 student reduction in class sizes from a mean of 25). The largest impacts were in the order of 25-30% of one standard deviation for a similar resource impact. The experimental evidence from the STAR experiment is somewhere in the middle of this range. Clearly, from a policy perspective, this range is very wide, as the cost-benefit implications are very different at the top and bottom ends of the range, so more work is needed to try to narrow this down e.g. by a statistical meta-analysis of recent studies. More experimental research using randomised control trials would also help provide confidence in these figures, although experimental work on class sizes and resources in schools requires large scale experiments and are clearly liable to be controversial. Given the current state of knowledge, it is sensible to treat the current estimates as upper and lower bounds to the potential impacts of resource changes.

The key question at the outset of this review was whether the evidence indicated that resources invested in early years and primary education were more effective than resources allocated to secondary education and later years, justifying a transfer of resources from later to earlier educational stages. The background to this line of reasoning is the work by James Heckman and others that appears to support early interventions. The fact that many of the most well known and most reliable studies – particularly the STAR experiments – are on younger children may also have led to popular impression of the greater efficacy of early interventions. Our reading of the evidence is, however, that there is no completely compelling case to support a transfer from later to early stages of education given the current state of knowledge. Certainly there is evidence that differences in achievement open up early in a child's life, and subsequent achievements are closely linked to early achievements. This in turn implies that there is a theoretical advantage in addressing disparities in achievement early on, so that these disparities are not propagated to, and amplified in, later stages in the life cycle. This is essentially the evidence on which the Heckman line of reasoning is based. The problem is that it is not obvious from the empirical evidence on the effectiveness of resources that it is any easier to address the small disparities early on in life through policy interventions than it is to address disparities later in child development, so it would be premature to advocate a shift of resources give the current information available.

On balance there are probably more studies finding positive resource impacts in primary school and early years than in secondary school. This may be in part because there have been more studies focussing on primary schooling, and the research designs have typically been better. However, where comparable designs are available in same economic and education

context (e.g. the studies using the National Pupil Database in England, Holmlund et al, 2010; Jenkins et al, 2005 & 2006; Nicoletti and Rabe, 2012), the effect sizes at different phases seem comparable. Moreover, a closer reading of the Heckman literature on investments over the life cycle (Cunha and Heckman, 2007) suggests that a balanced approach with investments throughout the lifecycle is preferable to interventions at any one stage. The benefits of investments at an early age, although potentially offering higher returns, erode during later phases of childhood unless topped up with subsequent investments.

8. References

Altinok, N. and G. Kingdon (2012), New Evidence on Class Size Effects: A Pupil Fixed Effects Approach, Oxford Bulletin of Economics and Statistics, 74 (2) 203-34.

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

Angrist, J. and V. Lavy (2002), New Evidence on Classroom Computers and Pupil Learning, Economic Journal, 112: 735-65.

Asadullah, M. Niaz (2005) The Effect of Class Size on Student Achievement: Evidence from Bangladesh, Applied Economics Letters, 12 (4) 217-21.

Barrow, L. and C. Rouse (2005), Causality, Causality Causality: the View of Education Inputs and Outputs from Economics, Federal Reserve Bank of Chicago, Working Paper Series WP05-05-15.

Blatchford, P., H. Goldstein, C. Martin and W. Browne (2002), A Study of Class Size Effects in English School Reception Year Classes, British Educational Research Journal, 28(2): 169-85.

Bradley, S. and J. Taylor (2010), Diversity, Choice and the Quasi-market: An Empirical Analysis of Secondary Education Policy in England, Oxford Bulletin of Economics and Statistics, 72(1) 1-26.

Bressoux, P., F. Kramarz and C. Prost (2009), Teachers' Training, Class Size and Students' Outcomes: Learning from Administrative Forecasting Mistakes, Economic Journal, 119: 540-61.

Chaudhary, L. (2009), Education Inputs, Student Performance and School Finance Reform in Michigan, Economics of Education Review, 28: 90-98.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. Whitemore Schanzenbach, and D. Yagan (2011), How Does Your Kindergarten Classroom Effect Your Earnings? Evidence from Project STAR, Quarterly Journal of Economics, 126(4): 1593-660.

Cho, H., P. Glewwe, and M. Whitler (2012), Do Reductions in Class Size Raise Student's Test Scores? Evidence from Population Variation in Minnesota's Elementary Schools, Economics of Education Review, 31: 77-95.

Cunha, F. and J. J. Heckman (2007), The Technology of Skill Formation, American Economic Review, 97 (2) 31-47.

Cunha, F. and J. J. Heckman (2008), Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation, Journal of Human Resources, 43 (4) 738-82.

Dearden, L., J. Ferri and C. Meghir (2002), The Effect of School Quality on Educational Attainment and Wages, Review of Economics and Statistics, 84(1): 1-20.

Denny, K. and V. Oppedisano (2010), Class Size Effects Using a New Estimation Technique, Geary Institute University College Dublin Working Papers 201051, December.

DfE (2011), The National Strategies 1997–2011, A Brief Summary of the Impact and Effectiveness of the National Strategies, DfE Report 00032-2011.

Feinstein, L. (2003), Inequality in the Early Cognitive Development of British Children in the 1970 Cohort, Economica, 70, 73-97.

Gary-Bobo, R. J. and M. B. Mahjoub (2006), Estimation of Class-Size Effects, Using 'Maimonides' Rule': The Case of French Junior High Schools, Centre for Economic Policy Research Discussion Paper 5754.

Gibbons, S., S. McNally, and M. Viarengo (2011), Does Additional Spending Help Urban Schools? An Evaluation Using Boundary Discontinuities, SERC Discussion Paper No.90, August; CEE Discussion Paper No.128, September.

Gibbons, S and S. Telhaj (2011), Pupil Mobility and School Disruption, Journal of Public Economics, 95 (9–10), 1019-1236.

Glewwe, P. and M. Kremer (2006), Schools, Teachers and Education Outcomes in Developing Countries. In Eric. A. Hanushek and Finis Welch (eds.), *Handbook of the Economics of Education*, Volume 2, Elsevier.

Glewwe, P., E. Hanushek, S. Humpage and R. Ravina (2011) School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010, NBER Working Paper 17554, October.

Goolsbee, A. and Guryan, J. (2006), The Impact of Internet Subsidies in Public Schools, Review of Economics and Statistics, Review of Economics and Statistics, 88(2): 336-47.

Gordon, N. (2004), Do Federal Grants Boost School Spending? Evidence from Title 1. Journal of Public Economics, 88: 1771-792.

Guryan, J. (2001), Does Money Matter? Regression Discontinuity Estimates from Education Finance Reform in Massachusetts, NBER Working Paper 8269, May.

Haegeland, T., O. Raaum and K. Salvanes (2012), Pennies from Heaven? Using Exogenous Tax Variation to Indentify Effects of School Resources on Pupil Achievement, Economics of Education Review, October 2012 31(5) 601-14.

Hakkinen, I., T. Kirjavainen and R. Uusitalo (2003), School Resources and Student Achievement Revisited: New Evidence from Panel Data, Economics of Education Review, 22 (3) 329-35.

Hansen, K. and D. Hawkes (2009), Early Childcare and Child Development, Journal of Social Policy, Vol. 38, 211-39.

Hanushek, E. and S. Rivkin (2010), Generalizations About Using Value-added Measures of Teacher Quality, American Economic Review, 100 (2) 267–71.

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

Hanushek, E. A. (2003), The Failure of Input-Based Schooling Policies, Economic Journal, 113: F64-F98.

Hanushek, E. A. and L. Woessmann (2008), The Role of Cognitive Skills in Economic Development, Journal of Economic Literature, 46 (3) 607-68.

Hanushek, E. A. and S. Rivkin (2012), The Distribution of Teacher Quality and Implications for Policy, Annual Review of Economics, 4, September 2012, 131-57.

Hanushek, E. A. (2006), School Resources. In Eric A. Hanushek and Finis Welch (eds.), *Handbook of the Economics of Education*, Volume 2, Elsevier.

Heckman, J. S., R. Moon, P. Pinto, P. Savelev and A. Yavitz (2010), The Rate of Return to the HighScope Perry Preschool Program, Journal of Public Economics, 94, 114-28.

Heinesen, E. (2010), Estimating Class-Size Effects Using Within School Variation in Subject-Specific Classes, Economic Journal, 120 (545) 737-60.

Holmlund, H., S. McNally, and M. Viarengo (2010), Does Money Matter for Schools? Economics of Education Review. 29 (6) 1154-64.

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

Hutchings, M., C. Greenwood, S. Hollingworth, A. Mansaray and A. Rose, with S. Minty and K. Glass (2012), Evaluation of the City Challenge programme, Department for Education Research Report DFE-RR215.

Iacovou, M. (2002), Class Size in the Early Years: Is Smaller Really Better? Education Economics, 10(3): 261-90.

Institute for the Study of Children, Families and Social Issues (2010), The impact of Sure Start Local Programmes on Five Year Olds and their Families. The National Evaluation of Sure Start (NESS) Team, Institute for the Study of Children, Families and Social Issues, Birkbeck University of London. DFE-RR067.

Jenkins, A., R. Levacic and A. Vignoles (2006), Estimating the Relationship Between School Resources and Pupil Attainment at GCSE, Department for Education and Skills Research Report RR727.

Jepsen, C. and S. Rivkin (2009), Class Size Reductions and Student Achievement: The Potential Trade off between Teacher Quality and Class Size, Journal of Human Resources, 44(1) 223-50.

Kremer, M. (2003), Randomized Evaluations of Educational Programs In Developing Countries: Some Lessons, American Economic Review, Vol. 93, 102–06.

Kremer, M. and A. Holla (2009), Improving Education in the Developing World: What Have We Learned from Randomised Evaluations? Annual Review of Economics, 1. 513-42.

Krueger, A. (1999), Experimental Estimates of Education Production Functions, Quarterly Journal of Economics, 114(2): 497-532.

Krueger, A. (2003), Economic Considerations and Class Size, Economic Journal, 113: F34-F63.

Lavy, V. and A. Schlosser (2011), Mechanisms and Impacts of Gender Peer Effects at School, American Economic Journal: Applied Economics, Vol. 3(2), 1–33.

Lavy, V., D. Paserman and A. Schlosser (2012), Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom, Economic Journal, 122(559): 208-37.

Lavy, V. (2012), Expanding School Resources and Increasing Time on Task: Effects of a Policy Experiment in Israel on Student Academic Achievement and Behaviour, NBER Working Paper 18369, September.

Leuven, E., M. Lindahl, H. Oosterbeek and D. Webbink, (2007), The Effect of Extra Funding for Disadvantaged Pupils on Achievement, The Review of Economics and Statistics, 89(4): 721-36.

Leuven, E., H. Oosterbeek and M. Ronning (2008), Quasi-Experimental Estimates of the Effect of Class Size on Achievement in Norway, Scandinavian Journal of Economics, 110 (4) 663-93.

Levačić, R., A. Jenkins, A. Vignoles, F. Steele and R. Allen (2005), Estimating the Relationship Between School Resources and Pupil Attainment at Key Stage 3, Department for Education and Skills Research Report RR679.

Levačić, R. and A. Vignoles (2002), Researching the Links between School Resources and Student Outcomes in the UK: A Review of Issues and Evidence, Education Economics 10(3): 313-31.

Lindahl, M. (2005), Home Versus School Learning: A New Approach to Estimating the Effect of Class Size on Achievement, The Scandinavian Journal of Economics, 107(2): 375-94.

NAO (2007), The Academies Programme. Report by the Controller and Auditor General, National Audit Office, The Stationery Office, London.

Nicoletti, C. and B. Rabe (2012), The Effect of School Resources on Test Scores in England, ISER Working Paper 2012-13, July.

Machin, S, S. McNally and C. Meghir, (2010), Resources and Standards in Urban Schools, Journal of Human Capital, 4(4) 365-93.

Machin, S., S. McNally and O. Silva (2007), New Technology in Schools: Is there a Payoff? Economic Journal,117(522):1145-67.

Machin, S. and S. McNally (2008), The Literacy Hour, Journal of Public Economics, 92, 1441–62.

Matsudaira, J. D., A. Hosek and E. Walsh (2012), An Integrated Assessment of the Effects of Title 1 on School Behaviour, Resources and Student Achievement, Economics of Education Review, 31: 1-14.

OECD (2011), PISA 2009 results Annex B1 http://www.oecd.org/pisa/pisaproducts/pisa2009/pisa2009resultswhatstudentsknowandcandostudentperformanceinreadingmathematicsandsciencevolumei.htm

OECD (2012), Education at a Glance (2012), OECD Indicators, Organisation of Economic Cooperation and Development, Paris.

Ofsted (2010), The National Strategies: A Review of Impact, Office for Standards in Education, Manchester.

Papke, L. E. (2005), The Effects of Spending on Test Pass Rates: Evidence from Michigan, Journal of Public Economics, 8: 821-39.

Piketty, T. and M. Valdenaire (2006), L'impact de la taille des classes sur la reussite scolaire dans les ecoles, collμeges et lycees francais: Estimations μa partir du panel primaire 1997 et du panel secondaire 1995, Ministμere de l'Education Nationale, Les Dossiers n173.

Piketty, T. (2004), L'impact de la taille des classes et de la ségrégation sociale sur la réussite scolaire dans les écoles françaises : une estimation à partir du panel primaire 1997, mimeo.

Rivkin, S., E. Hanushek and J. Kain (2005), Teachers, Schools, and Academic Achievement, Econometrica, 73 (2) 417–58.

Rockoff, J. (2009), Field Experiments in Class Size from the Early Twentieth Century, Journal of Economic Perspectives, 23 (4) 211-30.

Rothstein, J. (2010), Teacher Quality in Educational Production: Tracking, Decay, and

Student Achievement, <u>The Quarterly Journal of Economics</u>, 125(1), 175-214.

Rouse, C. E. and A. B. Krueger (2004), Putting Computerized Instruction To The Test: A Randomized Evaluation Of A 'Scientifically Based' Reading Program, <u>Economics of Education Review</u>, 23(4): 323-38.

Schanzenback, D. (2006), What Have Researchers Learned from Project STAR? Brookings Papers on Education Policy: 205-28.

Sims, D. P. (2008), Crowing Peter to Educate Paul: Lessons from a Class Size Reduction Externality, <u>Economics of Education Review</u>,28: 465-73.

Slater, H., N. Davies and S. Burgess (2012), Do Teachers Matter? Measuring the Variation in Teacher Effectiveness in England, <u>Oxford Bulletin of Economics and Statistics</u>, 74(5) 629-45.

Urquiola, M. (2006), Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia, <u>The Review of Economics and Statistics</u>, 88(1): 171-77.

Vignoles, A., R. Levačić, S. Machin, D. Reynolds and J. Walker (2000), The Relationship Between Resource Allocation and Pupil Attainment in the UK: A Review, CEE Discussion Paper 02, September.

Woessmann, L. and M. West (2006), Class Size Effects From Around the World: Evidence from between Grade Variation in TIMSS, <u>Economics of Education Review</u> 50(3) 695-736.

9. Appendix A

Reasons for lack of comparability between schools with different levels of resources

As discussed in Section 2 the key problem in estimating the causal effect of resources on student achievement is that pupil and school characteristics (e.g. ability, teacher quality) in schools or classes with more resources are not necessarily comparable to pupil and school characteristics in schools or classes with fewer resources. These pupil and school characteristics may have a direct effect on achievement, and are not necessarily observable to the researcher in their data. These omitted or confounding factors lead to upward or downward biases in the estimation of causal resource impacts. The principal reasons behind these differences in characteristics between high and low-resource schools are:

1) Resources devoted to schooling are determined by centralised policy in a way that compensates disadvantaged areas, schools or students by providing extra resources. A negative correlation between resources and achievement is then driven by the compensatory funding formula, which tends to work against finding any positive effect of expenditure on achievement.

2) Resources devoted to schooling are raised from local sources through taxation or charity, so expenditure depends on the tax base, local incomes and local demographics which are in turn directly related to child outcomes through parental background. A positive correlation between resources and achievement could then arise because higher achieving students come from backgrounds which are conducive to generating more expenditure, and again there is not necessarily any causal link between expenditure and outcomes.

3) The resources available to schools (and the way they are used) depend on the governance and leadership of the school, which may also have a direct influence on achievement. For example, a motivated an effective headteacher/principal may raise achievement through recruitment and organisation, and be effective at raising finance, but this does not imply that it is the financial resources which make a difference to outcomes.

4) Class sizes and hence spending per student may depend directly on parent and pupil choices, given that parents and children can choose which school to attend (or where to live in order to access a school) and this decision is likely to be related to school quality. For example, a school (or class) that is known to perform well may attract a high enrolment leading to large class sizes, high pupil teacher ratios and low expenditure per pupil. Conversely a poor performing school (or class) may lose students.

5) 'Sorting' and 'selection' effects arise through mobility of pupils and teachers across schools, and result in high resource schools differing in their composition from low-resource schools, with different types of teacher and different types of pupil. For example, if the school choices of higher ability students from better off backgrounds are more sensitive to class size or expenditure differences, or these families better positioned to exercise choice, then higher ability pupils may end up in the better-resourced schools or classes. In addition, school policies may assign more able pupils or pupils with educational needs into smaller classes, so again small and large classes are not comparable in terms of student composition. Similar arguments apply to teachers, if, for example higher quality teachers end up choosing better resourced schools. The recent literature worries about sorting and selection a lot.

6) Responses by teachers and pupils/parents can work against or with resource differences leading to upward and downward biases. For example, parents may hire private tuition or put in their own time to compensate for low resources/large class sizes in school. Teachers may respond to class size and resource differences by varying the amount of effort (mental effort in the class, out of work hours etc.) that they put in. Both of these compensatory behaviours tend to attenuate estimates of the effects of the resource differences Alternatively, parents,

students and teachers may disengage from the education process in response to resource cuts and engage more in response to resource increases, amplifying the effects from resource changes. Estimates of the effects of resource differences that do not take these responses into account are still causal, as they show what would be expected to happen in response to resource changes. However, they are not estimates of the impact of resources holding everything else including teacher, pupil and parent effort constant, which is often the intended research goal.

10. Appendix B: Benchmarking and Comparing the Size and Strength of Effects Across Different Studies

Different studies on resource effects in education use different measures, even for what is conceptually the same outcome. For example, when looking at achievement, some studies report effects on test scores, some studies report effects on the proportion achieving certain qualifications or standards, or dropping out of college. Even when using a broadly similar indicator like test scores in different contexts, the scales are not always comparable due to the different designs and scale of the tests. Resource variables too are often not easily compared across different studies. A trivial case is when studies look at the effects of expenditure, but expenditure is measured in different currency units. In these cases some simple currency conversions can help. More difficult cases to compare are, for example, class size changes and expenditure changes. For this reason, researchers often try to standardise the size of the effects they report, and to provide comparisons with other benchmark studies. In the class-size literature, it has become common to relate the size of effects to the class size impacts found in the Tennessee STAR experiment. More generally, researchers usually report 'effect sizes' in terms of standard deviations of the outcome variable (e.g. test scores).

Where possible, we too have reported the results of the studies using this convention, stating for example that a 10 student reduction in class sizes, or a £1000 increase in expenditure raises student achievement by x% of one standard deviation. For statisticians and applied researchers, standard deviations (sd) are a natural way to think about effect sizes, but for others they are not necessarily intuitive and may need some explanation. The standard deviation is a measure of the variability in a score around the mean (average). For the most common distributions of test scores, around 60-70% of students have test scores within +/-1 standard deviation of the mean. A student who is one-standard deviation above the mean, is therefore just inside the top 15-20% of students.

The easiest way to understand these orders of magnitude is to imagine ranking all students taking a test and assigning the top 1% a score of 100, the next 1% a score of 99 and so on until the bottom 1% who get a score of 1 (i.e. assign them to percentiles in the distribution). If we say a given resource change raises scores by 10% of one standard deviation (0.1 s.d.) , this is like moving the average achieving student ranked at 50, up around 3 points on this scale (3 percentiles) to 53. This may not seem like a big effect, although in educational intervention terms it would be considered quite a large effect, and there is no general way of deciding whether an effect is big or small, without reference to the effects of feasible alternative interventions. In education terms a 0.1 s.d. change in student scores is a big change because a lot of the variation in student achievement is due to natural ability, family background and luck, and other factors that have so far seemed out of reach of feasible education-related interventions. The effect size typically quoted for the Tennessee STAR class size experiment is around 0.15 standard deviations for an 8 student class size reduction.

Another way to benchmark an effect measured in standard deviations is to think in terms of some familiar level of achievement, like the probability of achieving Level 4 in Key Stage

2 tests or the probability of achieving a C-A* in GCSEs. This is difficult, because it depends on the way that the underlying distribution of test scores converts into these levels of achievement. In some of our work (Gibbons et al 2011) we looked at the conversion to Key Stage 2 maths levels. One standard deviation in the maths score distribution in 2009 was 23 points (on a 0-100 scale), therefore a 0.1 standard deviation change is 2.3 points. About 20% of students were at Level 3 or below, and the threshold mark between Level 3 and Level 4 was around 45. Around 15% of these students had marks above 42.7, so a 2.3 mark improvement would have put them over the 45 threshold. In other words, a 0.1 s.d. improvement in maths test scores would push 15% of students currently at Level 3 to Level 4. However, given that only 20% of students are at Level 4, this implies only a 3 percentage point increase the probability of achieving Level 4 in maths (0.2*0.15).

Sometimes it is also necessary to report the change in resources in terms of standard deviations too, perhaps because the resource has no natural scale. An example is 'teacher quality', where studies often state that a one standard deviation improvement in teacher quality leads to a 0.1 (or 10%) standard deviation increase in student scores. Referring to the above, this would imply that a teacher who is one of the best 15-20% of teachers raises average student scores by 3 points (percentiles) on this uniform scale from 0-100. Again this doesn't look big, but these teacher impacts are considered some of the biggest impacts in the education economics literature.

Another complication is that different interventions may have different implications in terms of total economic benefits, because of the number of students affected. For example, in the teacher case above, if an intervention could be found to raise a teacher's teaching quality by 1.s.d. permanently, then all students in the class benefit (perhaps 25 students) over the entire remainder of the teachers career, which could amount to 1000 students. For a proper comparison of resource-based interventions, some form of cost benefits or cost effectiveness analysis is required. A cost effectiveness estimate might compare the costs of achieving a 0.1 s.d. improvement in student test scores through intervention A (e.g. reducing class sizes by 10 students), with the costs of achieving a 0.1 s.d. improvement in student test scores through intervention B (e.g. simply allocating £1000 more resources per pupil to schools). A cost benefit analysis would attempt to monetise the benefits, though this is a very difficult and assumption-laden exercise, and practical examples are typically limited to working out the labour market returns to individual differences in achievement, and so providing an estimate of the benefits in terms of lifetime earnings.
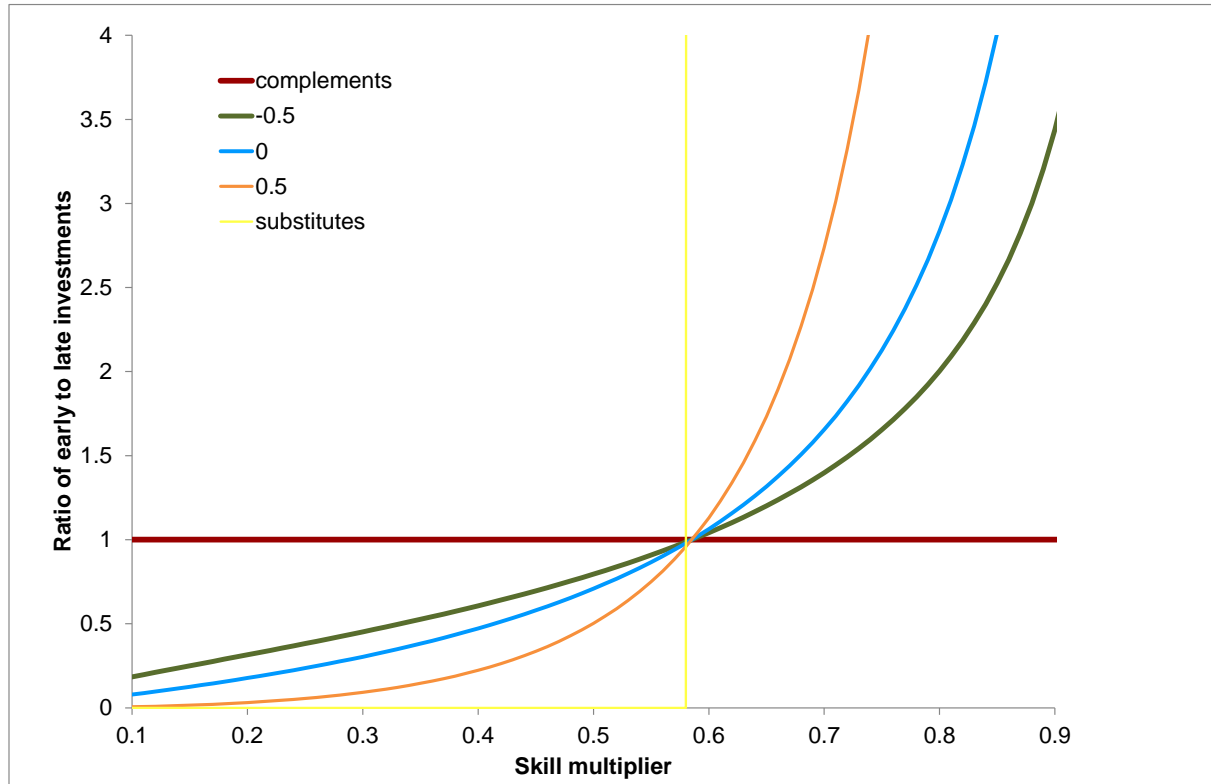

11. Appendix C

The figure illustrates the theoretical optimal ratio of early (period 1) to late investments (period 2) in the two period model described in Cunha and Heckman (2007). The figure is derived from their Equation 9 and is similar to their Figure 2, but adapted to allow for a non-zero discount rate, which is set to 3.5% per year, and assuming a 10 year interval between periods 1 and 2.

The horizontal axis shows the skill multiplier, which is a parameter which represents the effect of early investments on the productivity of later investments. The vertical axis shows the predicted ratio of early (period 1) to late investments (period 2). The different curves show the relationship for different assumptions about the substitutability of investments in early and later periods in producing the adult stock of skill: from perfect complements through to perfect substitutes (the number in the legend is the value of a parameter that determines this complementarity). In general, for low levels of the skill multiplier, investment is optimal in later periods (the ratio of early to late investments is below 1). For high levels of

the skill multiplier, more investment is optimal in later periods (the ratio of early to late is above 1). For the special case where early and late investments are perfect complements, the ratio is 1 regardless of the skill multiplier. This is because perfect complementarity means both early and late investments are needed in equal quantity. For the special case where early and late investments are substitutes, investment is in either the early or late periods but never in both. As can be seen, the theory alone cannot determine the optimal investment sequence without knowledge of the skill multiplier and the degree of complementarity in the production of skills.

Technical note: the ratio of investments is given by ratio = $(c/(1-c)*(1+r))^{1/(1-s)}$, where c is the skill multiplier, r is the discount rate and s is the complementarity parameter.

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1225	Cornelius A. Rietveld David Cesarini Daniel J. Benjamin Philipp D. Koellinger Jan-Emmanuel De Neve Henning Tiemeier Magnus Johannesson Patrik K.E. Magnusson Nancy L. Pedersen Robert F. Krueger Meike Bartels	Molecular Genetics and Subjective Well-Being
1224	Peter Arcidiacono Esteban Aucejo Patrick Coate V. Joseph Hotz	Affirmative Action and University Fit: Evidence from Proposition 209
1223	Peter Arcidiacono Esteban Aucejo V. Joseph Hotz	University Differences in the Graduation of Minorities in STEM Fields: Evidence from California
1222	Paul Dolan Robert Metcalfe	Neighbors, Knowledge, and Nuggets: Two Natural Field Experiments on the Role of Incentives on Energy Conservation
1221	Andy Feng Georg Graetz	A Question of Degree: The Effects of Degree Class on Labor Market Outcomes
1220	Esteban Aucejo	Explaining Cross-Racial Differences in the Educational Gender Gap
1219	Peter Arcidiacono Esteban Aucejo Andrew Hussey Kenneth Spenner	Racial Segregation Patterns in Selective Universities
1218	Silvana Tenreyro Gregory Thwaites	Pushing On a String: US Monetary Policy is Less Powerful in Recessions
1217	Gianluca Benigno Luca Fornaro	The Financial Resource Curse
1216	Daron Acemoglu Ufuk Akcigit Nicholas Bloom William R. Kerr	Innovation, Reallocation and Growth
1215	Michael J. Boehm	Has Job Polarization Squeezed the Middle Class? Evidence from the Allocation of Talents

1214	Nattavudh Powdthavee Warn N. Lekfuangfu Mark Wooden	The Marginal Income Effect of Education on Happiness: Estimating the Direct and Indirect Effects of Compulsory Schooling on Well-Being in Australia
1213	Richard Layard	Mental Health: The New Frontier for Labour Economics
1212	Francesco Caselli Massimo Morelli Dominic Rohner	The Geography of Inter-State Resource Wars
1211	Stephen Hansen Michael McMahon	Estimating Bayesian Decision Problems with Heterogeneous Priors
1210	Christopher A. Pissarides	Unemployment in the Great Recession
1209	Kevin D. Sheedy	Debt and Incomplete Financial Markets: A Case for Nominal GDP Targeting
1208	Jordi Blanes i Vidal Marc Möller	Decision-Making and Implementation in Teams
1207	Michael J. Boehm	Concentration versus Re-Matching? Evidence About the Locational Effects of Commuting Costs
1206	Antonella Nocco Gianmarco I. P. Ottaviano Matteo Salto	Monopolistic Competition and Optimum Product Selection: Why and How Heterogeneity Matters
1205	Alberto Galasso Mark Schankerman	Patents and Cumulative Innovation: Causal Evidence from the Courts
1204	L Rachel Ngai Barbara Petrongolo	Gender Gaps and the Rise of the Service Economy
1203	Luis Garicano Luis Rayo	Relational Knowledge Transfers
1202	Abel Brodeur	Smoking, Income and Subjective Well-Being: Evidence from Smoking Bans