Oxford Review of Economic Policy, Volume 30, Number 4, 2014, pp. 597–620

The role of experimentation in education policy

Sally Sadoff*

Abstract As the gold standard for programme evaluation, experimentation is gaining increasing attention from education researchers and policy-makers. In this article I discuss how experimentation can be used to shape education policy moving forward. I first discuss how well-designed experiments can both build upon and inform a general framework for the education production function. Experiments within this framework can be particularly powerful when they draw on a wide range of disciplines including child development, psychology, and behavioural economics. Insights from these areas can help identify underlying mechanisms of the education production function and inform the design of interventions in ways that increase (cost-) effectiveness. Additionally, I argue that there should be a rich array of experiments in education, ranging from lab-like basic research to policy-level efficacy trials. Finally, I discuss the policy-maker's role in integrating experimentation into policy decision-making, including how to address concerns commonly raised with experimentation in education.

Keywords: education, policy, field experiments, behavioural economics

JEL classification: I28, I21, C93, D03

I. Introduction

Experimentation is a simple yet extremely powerful tool. Particularly so in education, where it has great potential to inform policy and shape education reform. The key feature of experiments—random assignment to an intervention—is considered the 'gold standard' for programme evaluation and is increasingly being demanded by education policy-makers as an essential tool for decision-making (Coalition for Evidence Based Policy, 2003). Random assignment ensures that on average there are no underlying differences between the intervention 'treatment' group and the comparison 'control' group. Therefore, any differences in outcomes between these two groups can be attributed to the intervention itself, allowing us to *causally identify* the impact of a given programme.

This is crucial for the evaluation of education policy. First, it allows you to move one policy lever at a time. Generally, education policies move together, making it difficult

* University of California, San Diego, e-mail: ssadoff@ucsd.edu

I am grateful to the editors as well as two referees for providing extremely helpful comments. Participants in the UK Department for Education seminar provided particularly useful feedback from the policy-maker perspective. Finally, I would like to thank participants in the meeting on research in parental engagement in education for highlighting examples of innovative experiments in education.

doi:10.1093/oxrep/grv001

to isolate the question of interest. For example, successful schools are often characterized by a bundle of features including high-quality teachers, a specialized curriculum, engaged parents, and motivated students. Without experimentation, it is very difficult to identify the role each of these plays in educational achievement. Relatedly, experiments allow you to avoid the selection biases that plague many comparisons in education. Suppose you want to measure the effectiveness of a supplemental tutoring programme. The factors that drive where these programmes are available and who attends them are largely unobservable and are likely related to the outcomes you want to measure. For example, students willing to spend extracurricular time in an academic programme are likely more motivated (or their parents are) than students who do not attend the programme. Thus, differences in performance between tutored students and non-tutored students, could be either due to the programme or due to underlying differences in student motivation. More broadly, so much of educational achievement is driven by factors outside the school system (e.g. pre-schooling cognitive development, parental resources, etc.) that identifying the effect of a particular policy is difficult without experimentation.

Finally, experiments allow you to test interventions that do not yet exist—i.e. there is no naturally occurring data to analyse. This is the area where experiments have the greatest potential to shape education policy. As I discuss in more detail below, well-designed experiments can deepen our understanding of the education production function, which in turn allows us to develop more (cost-) effective interventions. Particularly, if they draw on insights from a broad range of disciplines including child development, psychology, and behavioural economics (Lavecchia *et al.*, 2014).

This article does not attempt to survey the increasingly large literature on experiments in education. Instead, I focus on examples that highlight how experimentation can be used to shape education policy moving forward. These examples are largely drawn from experiments in developed countries, though I do note a number of related studies implemented in developing country contexts. In section II, I discuss a general framework for the education production function, in which individuals invest in an education production technology that increases students' abilities and achievement that in turn have long-run lifetime returns for the individual (and society). I first discuss the production technology (section II(i)), then the investment decision (section II(ii)), and finally abilities, achievement, and long-run returns to education (section II(iii)). Section III discusses the policy-maker's role in both addressing challenges to experimentation in education (section III(i)) and integrating experimentation into policy decision-making (section III(ii)). Section IV concludes with recommendations for the role of experiments in education policy, including areas for future research.

II. Framework for the education production function

(i) The education production technology

National, state, and local district authority level

While probably having the largest impact on student outcomes, policies that take place at the national, state, or district level are the least likely to be tested experimentally. This is because it is not obvious how to randomize policies across school districts, let alone states or countries. For example, how can we evaluate the effect of school finance equalization across districts, a nationwide reform such as No Child Left Behind, state or nationwide requirements such as high school exit exams (e.g. the General Certificate of Secondary Education (GCSE) in the UK), or national college entrance exams (e.g. the ACT and SAT in the US or A-levels in the UK)?

Given the difficulty in estimating the causal impact of these policies, this is an area where providing space for experimentation could be extremely valuable. For example, proposals to equalize per-student spending across districts often face resistance from local politicians and parents who perhaps rightly question findings from non-experimental studies. Policy-makers could potentially diffuse concerns by running an experiment in which schools with low spending per student are randomly assigned to either receive increased resources or receive their current resources. In this case, no school is made worse off and the policy can be tested randomly. Of course there may still be concerns about fairness, which I discuss in section III(i).

A similar approach could be used to evaluate a nationwide reform such as the No Child Left Behind (NCLB) legislation in the US, which aims to improve achievement through school-based accountability standards. In response to push-back against NCLB, some states have been allowed to design alternative accountability systems. Policy-makers could require that, in order to opt out of the national requirements, states, districts, or schools must test their alternative policies using randomized experiments. That is, some schools or districts would remain under the national standards while others would be randomly assigned to the state's proposed alternative. This would allow a causal comparison of the two policies. In other cases, such as the Common Core reform recently implemented in the US, the policy lays out learning goals but allows flexibility in how these goals are met. States that follow the Common Core can then run randomized experiments across schools and districts in order to understand how to implement the policy most effectively.

Even in the absence of district-level randomization, experimental methods can shed light on the effectiveness of district-level policies. For example, does increased school choice improve school quality and student outcomes via the forces of competition? Choice policies can take a number of forms. Some allow students to choose among traditional public schools rather than being required to attend their neighbourhood school. Others allow students to opt out of traditional public schools to attend an alternative public school—such as a charter school, magnet school, small school, or vocational school—which receives public funding but has greater instructional flexibility. An alternative model gives students vouchers that serve as a credit towards tuition at private schools.

It is difficult to evaluate the effect of these policies on overall school quality because it is difficult, though not impossible, to randomly assign some school districts to a choice policy and others to no choice. However, a substantial number of studies have evaluated these policies at the school and individual level. That is, what is the effect of attending a school of choice, an alternative public school, or receiving a voucher? These studies take advantage of random assignment of vouchers to students or randomized admissions lotteries to schools of choice to evaluate the impact of these programmes (e.g. Cullen *et al.* (2013) discuss the evidence on public school choice programmes; Chingos and Peterson (2012) and Wolf *et al.* (2013) estimate the impacts of voucher programmes). Taken together, choice schools do not on average seem to improve student achievement. However, these studies have been able to identify certain types of schools or programmes that are particularly effective.

For example, several studies examine the differential effects of charter schools in order to identify the critical school characteristics correlated with larger impacts (Hoxby and Murarka, 2009; Angrist *et al*., 2013; Dobbie and Fryer, 2013). Building on these findings, Fryer (2011*a*) designs a prospective experimental intervention in which the key components of charter schools' success are incorporated into traditional public schools. The early results are promising, with increases in test performance comparable to those in highly successful charter schools. This type of research exemplifies how experiments should be used in education policy. It draws on both experimental estimates of causal effects as well as secondary analysis of experimental data, using evidence from a large number of students and schools. These results inform the design of a new intervention which is first tested at a programmatic scale (in nine schools with 7,000 students). The findings from the initial study can then guide if and how the intervention should be scaled up at a policy level.

School, class, and student level

The most common experiments in education policy are what are often referred to as programme evaluations. They usually measure the effect of an intervention that is randomly assigned at the school, class, or individual level. And they have been used to evaluate a wide range of programmes and policies including: instructional models, curriculum, and coursework; added resources such as classroom aides, extended instructional time, physical materials, or technological tools; and, providing supplementary instruction through individualized interventions, mentoring, tutoring, guidance counsellors, or extracurricular activities such as after school or summer programmes. Policymakers have begun to synthesize these studies for practitioners in databases such as the What Works Clearing House (an initiative of the US Department of Education's Institute of Education Sciences, http://ies.ed.gov/ncee/wwc/) and the UK Education Endowment Fund's Toolkit (available at: http://educationendowmentfoundation.org.uk/toolkit/).

Below I discuss seminal studies in this area and highlight recent experiments that move beyond traditional programme evaluation. These studies are richest when they track a wide range of long-term outcomes, examine heterogeneous effects in order to motivate the design of future interventions, and incorporate insights from behavioural economics into their design.

One of the longest-term follow-ups comes from two model preschool programmes tested in the 1960s and 1970s. The Perry Preschool and the Abecedarian programmes both randomized disadvantaged children in the US into intensive early childhood interventions. Follow-up studies have measured differences between the treatment and control groups, not only on school performance but also on a wide range of long-term outcomes. Interestingly, while treatment effects on test scores faded over time (with treatment and control students performing similarly by third grade), the early childhood interventions had a large and significant impact on longer-term outcomes, increasing educational attainment and earnings while decreasing drug use, reliance on welfare programmes, rates of teen parenthood, and criminal activity (Duncan and Magnuson (2013) review the evidence on early childhood interventions). These results suggest that tests of academic achievement and 'cognitive' skills may not fully capture abilities that are important for longer-term success. As I discuss in section II(iii), these findings have motivated greater interest in the role education plays in building 'non-cognitive' abilities, such as executive function, social-emotional skills, persistence, and self-control.

One concern with these studies is the extent to which these small-scale, resource-intensive programmes run by researchers can be scaled up at a policy level (Perry and Abecedarian each enrolled fewer than 60 children and cost about $16,000–$20,000 per year in 2011 dollars). In their meta-analysis of early childhood programmes from 1960 to 2005 in the US, Duncan and Magnuson (2013) find that programme impacts are, indeed, smaller for larger-scale programmes implemented by policy-makers and practitioners rather than researchers. The largest of these is HeadStart, which now serves almost a million 3- and 4-year olds each year. While the programme dates back to the 1960s, the first randomized assessment of the programme did not begin until 2002. The short-term programme impacts were significant but smaller than in Perry and Abecedarian and faded out more quickly. The question remains whether longer-term benefits will emerge when these children reach adolescence and adulthood. Findings from non-experimental evaluations of HeadStart suggest that, similar to findings for Perry and Abecedarian, treated students experience improved lifetime outcomes. Longer-term follow-up of the experimental cohorts will allow us to compare the causal impacts of the small-scale model programmes to larger-scale policy-level interventions.

A seminal example of experimentation at the policy level is the Tennessee Student/Teacher Achievement Ratio (STAR) study, which randomly assigned students both to teachers and to classes of different size. The initial results found that lowering class size by a third increases student achievement by about two-tenths of a standard deviation and that African-American and low-income students benefited most (Krueger, 1999). Longer-term follow-up has found positive impacts on educational attainment, home ownership, savings rates, and marital status. In parallel to the short-term results, the long-term effects are largest for African-American students and those attending the poorest third of schools (Chetty *et al*., 2011; Dynarski *et al.*, 2013). Because the STAR programme involved over 11,500 students in 79 schools, the results are likely to generalize across settings. For this reason, the study has served as an important benchmark against which the (cost-) effectiveness of other educational interventions is often measured.

Perhaps the most important input at the school or class level is the quality of the personnel. The impact of a one standard deviation improvement in teacher quality on student achievement is equivalent to reducing class size by a third. Even more impressive are the long-term impacts, which include improved educational attainment, increased earnings, higher savings rates, and decreased probability of teen parenthood (Chetty *et al.*, 2014). While more difficult to measure, the estimated effects of the quality of school principals on student achievement are similar to those for teachers (Branch *et al.*, 2012).

There are two primary mechanisms for improving the human capital side of the education production technology: through selection or through increasing ability and performance. Selection mechanisms include screening tools for hiring, performance-based criteria for retention, and incentives to enter the profession or work in high-need schools. Interventions aiming to increase ability and performance have traditionally focused on professional development activities, such as training and certification. More recently, policy-makers have become increasingly interested in the use of incentives to improve performance.

While this is a critical area for policy, there are few experimental studies examining interventions aimed at principals and teachers. Observational studies have found that existing policies are largely ineffective at screening teachers, using performance criteria for retention, or increasing quality through professional development (e.g. Boyd *et al.*, 2007; Rockoff *et al.*, 2011; Jacob, 2011), though recent experimental evidence suggests that personalized teacher coaching can improve teacher quality and student achievement (Campbell and Malkus, 2011; Allen *et al.*, 2011). Analyses of teacher incentive programmes have found mixed results and are often subject to the concerns about causal identification discussed above (Neal (2011) reviews the literature on teacher incentives). And the few experimental studies of performance pay have found little impact in developed countries (Springer *et al.*, 2010, Fryer, 2011*b*; Goodman and Turner, 2013).[1] As a consequence, teachers unions are often resistant to incorporating such policies into their contracts.

Clearly, there is room for the design and experimental evaluation of new interventions. Two recent studies of incentives are promising. A growing literature demonstrates the difficulty of recruiting and retaining high-quality teachers at disadvantaged schools, where they could potentially have the largest impact (Lankford *et al.*, 2002; Clotfelter *et al.*, 2007). In response to this pressing policy challenge, the US Department of Education's Institute of Education Sciences (IES) contracted with a team of researchers who worked with 10 school districts to design and test incentives offered to high-performing teachers for teaching in low-performing schools (Glazerman *et al.*, 2013). The programme offered the top 20 per cent of teachers (as measured by their value added) $20,000 over 2 years if they transferred to and remained in a high-need school. The impact of the programme on elementary school achievement was equivalent to the effects of class size reduction and at lower cost (there were no significant effects on middle school performance). Longer-term follow-up will track teachers and their students to examine whether, after the incentives end, high-quality teachers remain in low-performing schools and continue to improve student achievement.

One challenge of the transfer incentive programme is that only 5 per cent of eligible teachers took up the incentives. If teachers are reluctant to transfer schools, even with large incentives to do so, it may be difficult to scale up the programme to sufficiently meet the needs of low-performing schools. An alternative to using incentives to induce teachers to change schools, is to incentivize performance among existing teachers. However, as discussed above, traditional merit pay programmes have demonstrated limited evidence of success. A recent study tests a new performance pay design, which draws on insights from behavioural economics. The design exploits *loss aversion*, a well-demonstrated finding in laboratory settings that individuals are more responsive to protocols framed as losses rather than gains. Fryer *et al*. (2012) design loss-framed incentives in which teachers receive bonus payments at the beginning of the year that they must pay back at the end of the year if their students do not improve sufficiently on a standardized test. They find that the impact of the loss-framed incentives on maths achievement is equivalent to improving teacher quality by a standard deviation. Similar to previous studies, they find that standard gain-framed incentives distributed at the end of the year do not have a significant impact on performance. These findings provide

[1] Results from teacher incentive studies in developing countries are more promising (Glewwe *et al.*, 2010; Muralidharan and Sundararaman, 2011; Duflo *et al.*, 2012).

evidence that small changes in design can have a large impact on the effectiveness of policy interventions. Future research could build on these incentive designs, examining their effect on longer-term teacher selection and retention, as well as applying them in related contexts such as teacher training and principal compensation.

(ii) The investment decision

Understanding the investment decision is where insights from behavioural economics can shed the most light on the education production function. In the standard framework, individuals invest in education based on the discounted returns to schooling net of costs. This framework typically makes the following assumptions: (i) individuals have perfect information and face negligible transactions costs; (ii) students and parents act as a unitary household (i.e. a single actor); (iii) individuals discount the returns to schooling exponentially; and (iv) investment can be measured by years of schooling—i.e. ignoring investment of time and effort while in school. One challenge of the standard framework is that under these assumptions it is difficult to explain why educational investment among many students is so low, given the high returns to schooling. Below, I discuss how recent behavioural interventions have relaxed these assumptions in order to better understand students' decision-making and ultimately increase educational investments.

Information and transactions costs

The standard framework assumes that individuals make their investment decisions with full information and negligible transactions costs. That is, they are able to acquire, process, and optimally act on information, and the cost of doing so is low relative to the returns. In contrast, research in psychology and behavioural economics demonstrates that, even for high-stakes decisions, slight changes to how information or choices are presented can have a significant impact on choices (Thaler and Sunstein, 2008).

In education, among the highest-stakes decisions students make is whether (and where) to attend college. A growing policy debate has arisen in the US about the extent to which this decision is determined by rising tuition costs. Some argue that many low-income students can no longer afford college, while others respond that these students are eligible for substantial financial aid to defray tuition costs. It is puzzling then that a significant portion of students who qualify for financial aid fail to take it up. In a recent study, Bettinger *et al*. (2012) examine the extent to which the financial aid application process itself may be a barrier to take-up and, as a consequence, college enrolment. They test their intervention among over 25,000 clients of H&R Block, a US tax preparation company, which provided eligible families with information, assistance, and a streamlined process for filling out the Free Application for Federal Student Aid (FAFSA). They find that the intervention increases not only FAFSA submission rates but also has a long-term impact on college enrolment, attendance, and persistence.

This study demonstrates that a simple, low-cost intervention implemented at a policy-level scale can have a significant impact on the ultimate educational outcome we care about: final attainment. Similar policy-level experiments in the US have shown that providing low-income families with accessible information about their schooling options increases the likelihood that they will choose to attend higher-performing schools at the elementary, secondary, and post-secondary levels (Hastings and Weinstein, 2008;

Hoxby and Turner, 2013). Building on these results, an informational intervention aimed at high-performing secondary students in low-achievement schools is currently being tested in a UK context (Sanders and Chande, 2015).[2]

A second strand of this literature also examines informational interventions, but rather than focusing on the enrolment decision, aims to increase parents' and students' investment in time and effort while enrolled. Fryer (2013) tests the effects of providing students with text messages about the returns to schooling. He finds that the intervention increases students' estimates of schooling returns as well as self-reported effort, but has no impact on achievement.

A related set of recent studies examines the effect of providing information to parents about their child's activities and performance in school. Researchers, policymakers, and educators have long recognized the role of parents in shaping student achievement. A large body of observational studies documents the strong relationship between family background and educational outcomes (Sirin, 2005), but to date there have been very few experimental studies in this area (Avvisati *et al.*, 2010). Thus, there is little causal evidence on, first, whether it is possible to move parental behaviour and engagement in children's schooling, and, second, whether increased parental investment has an impact on student behaviour and achievement. Providing parents with information also addresses concerns that there may be informational frictions in the household—i.e. parents and students cannot be treated as a single agent making aligned decisions.

Broadly, studies in this area find evidence that providing parents with information through teacher phone calls, written communication, or text messages can increase parental engagement and student effort (Bergman, 2012: Kraft and Dougherty, forthcoming; Kraft and Rogers, in preparation). For example, Bergman (2012) tests the impact of providing parents with text messages about their child's performance. He finds that the intervention increased engagement among parents of high school students in the US and that students in turn exerted greater effort, which improved their grades and test performance in maths (there were no gains in English). A similar intervention is currently testing the impact of parental text messages at a scaled-up policy level in the UK (Education Endowment Fund, 2014*a*).[3]

In related work,Avvisati *et al.* (2014) implement a randomized intervention in disadvantaged French middle schools that invited parents to participate in parent–school meetings. The programme offered information on the transition from primary school to middle school and advice on how to support and monitor children with schoolwork. They find that the intervention improved both parental engagement in school-related activities and student behaviours, including reduced rates of truancy and better classroom reports from teachers (there were no effects on test performance). Interestingly, the impact spilled over to the classmates of treated students, who also demonstrated

[2] Similarly, Castleman and Page (2013) show that information provided to low-income students either in person or through text messages during the summer before college increases enrolment rates. Harackiewicz *et al.* (2012) find that providing parents with information about the value of science, technology, engineering, and mathematics (STEM) increases secondary school student enrolment in these courses.

[3] In a developing country context, Nguyen (2008) and Jensen (2010) examine the effect of providing information about the returns to schooling to students and parents; Bursztyn and Coffman (2012) examine parental demand for information about children's attendance at school; Banerji *et al.* (2013) measure the impact of maternal literacy and training classes on child learning.

improved behaviours, even though their parents did not participate in the programme. Future studies can build on these findings to design interventions aimed at generating positive peer effects.

Incentives

There has been an explosion of academic and policy interest in incentive-based education programmes in recent years. These are in part motivated by findings from psychology and behavioural economics that children and adolescents tend to exhibit high discount rates and have difficulty planning for the future (e.g. Gruber, 2001; Bettinger and Slonim, 2007; Steinberg *et al.*, 2009). These studies challenge the assumption of the standard framework that individuals discount returns to schooling exponentially. If, instead, students have hyperbolic time preferences or are otherwise myopic in a way that future rewards are largely ignored, they are likely to underinvest in education where the returns are primarily experienced after a delay of years or even decades. Indeed, as discussed in the section below, several studies provide evidence that individuals with higher discount rates invest less in education. Research in this area also focuses on the importance of time and effort invested while in school, recognizing that enrolment alone does not ensure student learning and human capital accumulation.

Incentive programmes, therefore, generally aim to increase student effort by reducing the time delay in experiencing returns. They usually do so by offering near-term rewards based on student achievement. For example, in a series of experiments in several US school districts, Fryer (2011*c*) tests the impact of offering students incentives for various performance measures, including attendance, behaviour, homework, grades, test scores, and reading books. He finds that incentives for reading books have the largest effects, followed by incentives for attendance, behaviour, and homework. The test-based and grade-based incentives have little or no effect on achievement. Similar randomized experiments have offered elementary, middle, and secondary students in the US incentives based on attendance (Dee, 2011), test performance (Bettinger, 2012; Levitt *et al.*, 2012), completing maths problems (Fryer, 2012), overall achievement (Levitt *et al.*, 2010; List *et al.*, 2012), and post-secondary application or enrolment (Rodriguez-Planas, 2012; Carrell and Sacerdote, 2013). Other programmes include rewards for passing the high school exit exam in Israel (Angrist and Lavy, 2009); as well as credit- or grade-based incentives for college students in Canada (Angrist *et al.*, 2009, 2014; MacDonald *et al.*, 2009), Italy (De Paola *et al.*, 2012), the Netherlands (Leuven *et al.*, 2010, 2011) and several US cities (Patel *et al.*, 2013; Barrow *et al.*, 2014).[4] While the effects of performance-based incentives vary, the studies tend to find small overall impacts with larger effects in particular sub-groups.

Several studies build on the basic idea of offering performance-based incentives to examine the effect of varying the incentive and intervention design. These experiments often explore whether there are important complementarities in the education production function—that is, whether interventions may be more powerful in combination than they are separately. For example, several incentive programmes combine incentives

[4] There are also a growing number of non-experimental evaluations of incentive programmes not discussed here, as well as experimental studies examining educational incentives in developing countries (e.g. Schultz, 2004; Behrman *et al.*, 2005; Angrist *et al.*, 2006; Kremer *et al.*, 2009; Sharma, 2010; Barrera-Osorio *et al.*, 2011; Berry, 2014).

with mentoring or academic support services (Angrist *et al.*, 2009, 2014; MacDonald *et al.*, 2009; Rodriguez-Planas, 2012; Carrell and Sacerdote, 2013). Evaluations of such programmes are richest when they test both the separate and combined impact of these interventions. In a randomized experiment among first-year college students in Canada, Angrist *et al.* (2009) compare offering students grade-based financial incentives, academic services, or both. They find that the combined intervention (financial incentives and services) has the largest impact, particularly among women. Carrell and Sacerdote (2013) find similar results among high school students in the US.

A related study explores complementarities among the various individuals that contribute to student achievement. List *et al.* (2012) examine the effect of varying the recipient of a performance-based incentive. They offer an equivalent reward to the student only, the parent only, the teacher only, or split the reward among these three. They find little evidence for complementarities in their context, with the individual rewards outperforming the rewards offered in combination. However, this is clearly an area that deserves further exploration.

Another recent strand of the literature explores whether incorporating insights from behavioural economics can increase the (cost-) effectiveness of incentives. Like List *et al.* (2012), Levitt *et al.* (2010) test the effect of varying the recipient of a monthly achievement incentive, either the parent or the student. As in the information interventions discussed above, this feature of the design builds on the behavioural insight that parents and students may not act as a unitary household. It also explores whether parental rewards can increase parental engagement as well as overall household investment in education. The study also varies the structure of the reward. Building on the insight from behavioural economics that people often overvalue small probabilities of receiving larger rewards, the design compares the effects of a fixed reward to a lottery reward of equivalent expected value. They find that the incentives have a modest effect on performance, but do not find differential impacts across features of the design.

In a related study, Levitt *et al.* (2012) incorporate insights from behavioural economics into the design of incentives aimed at increasing student effort on standardized tests. They first explore the importance of high discount rates by varying the timing of rewards, which they distribute either immediately after the incentivized task or with a small delay (of 1 month). They also examine the relevance of loss aversion (discussed in section II(i)), comparing rewards framed as losses to those framed as gains. Finally, they build on a growing literature on the motivational power of non-financial rewards, offering both financial and non-financial incentives. They find that several of the design features increase the cost-effectiveness of rewards. For example, non-financial incentives can be as effective as cash rewards at a fraction of the cost. They also find that immediate rewards have a much larger impact than rewards distributed with a small delay. This result provides an important lesson for our understanding of the education production function more generally. Given that a month-long delay in experiencing rewards dramatically reduces student effort, the years-long delay in experiencing returns to education likely leads to underinvestment among many students.

While several studies have demonstrated that near-term incentives can increase student effort and performance, there is often concern about longer-term effects after the incentives end. In particular, if offering students extrinsic rewards for achievement crowds out intrinsic motivation to learn, once the incentives are removed student performance may suffer (Kohn, 1999; Deci *et al.*, 1999). To address this concern, most incentive interventions continue to track students after the programme ends. These studies find little evidence of crowding out, with some programmes demonstrating positive post-treatment effects, particularly among the sub-group of students who experience the largest impacts during treatment (Angrist *et al.*, 2009; Leuven *et al.*, 2010; Levitt *et al.*, 2010). For the most part, however, the effects of incentive programmes tend to fade out after the interventions end.

A key challenge, then, is whether incentives can be used to promote habit formation in ways that lead to long-term learning and human capital accumulation. A first step is to better understand how individuals respond to incentives, as well as how effort 'inputs' map into achievement and learning 'outputs'. Surprisingly little is known about either of these questions. In fact, recent work has argued that many incentive programmes fail because students (and parents) do not understand the production function—i.e. how best to exert effort in response to performance-based incentives (Fryer, 2011*c*). For example, students may be motivated by a reward for improving their score on a final exam but may not know what to do during the year to achieve this goal. If this is the case, it may be more effective to offer incentives for effort inputs rather than performance outputs. That is, rather than reward students for achievement on the final exam, offer incentives for studying and learning the material throughout the year.

The challenge with this approach is that policy-makers, educators, and researchers have yet to identify which are the key effort inputs. Recent studies have demonstrated that student effort has a significant impact on performance (e.g. Levitt *et al.*, 2012; Metcalfe *et al.*, 2011) but do not identify the specific types of effort students are exerting. On this question, Barrow and Rouse (2013) use detailed time use surveys to examine specific effort responses to incentives, but do not link this response to achievement. Future research should build on these studies to identify the mechanisms of the education production function—linking incentives, the investment decision, and types of effort to learning and achievement. This can help shed light on the ultimate policy goal, which aims to understand when, why, and for whom incentives are effective at increasing educational investments and long-term human capital accumulation (Gneezy *et al.*, 2011).

(iii) Abilities, achievement, and the long-run returns to education

As discussed in the previous section, the optimal investment in schooling depends on the (discounted) long-run returns to education. These returns have traditionally been measured in terms of lifetime earnings. More recently, researchers have recognized that education improves not only labour market outcomes, but also a wide range of important non-labour market outcomes including health, savings rates, criminal behaviour, marital status, and self-reported well-being (Oreopoulos and Salvanes, 2011).

At the same time, most evaluations measure the effectiveness of an intervention using standardized assessments or school administrative data (e.g. enrolment, attendance, grades, and behaviour reports). The assumption is that these proxy students' abilities and knowledge so that improving these near-term measures will move the long-term outcomes of interest. Several recent studies provide support for this approach. As discussed in section II(i), Chetty *et al.* (2014) find that having a high-quality teacher as measured by the impact on a student's test scores improves a wide range of long-term outcomes. Similarly, the long-term follow-up studies of the Tennessee STAR experiment find that a student's kindergarten test score is highly correlated with long-term outcomes (Chetty *et al*., 2011; Dynarski *et al.*, 2013).

In contrast, studies of early childhood interventions find that while effects on test scores fade, treatment impacts re-emerge in adulthood. These results suggest that tests of academic achievement and 'cognitive' skills may not fully capture abilities that are important for longer-term success. This has motivated greater interest in the role of education in building 'non-cognitive' abilities, such as executive function, social-emotional skills, persistence, and self-control.[5] Recent research has focused on determining what the critical non-cognitive abilities are, how to measure them, and their link to later lifetime outcomes (Borghans *et al.*, 2008).

Research in psychology and behavioural economics has identified several personality traits that are correlated with long-term outcomes. Some of the strongest evidence comes from studies of time preferences and ability to delay gratification. These find a negative correlation between hyperbolic discount rates and educational outcomes (Kirby *et al*., 2002, 2005; Castillo *et al.*, 2011). Similarly, Mischel *et al.* (1989) find that measures of ability to delay gratification in early childhood are predictive of longer-term academic achievement. More recently, psychologists and economists have begun to focus on beliefs and attitude. Duckworth *et al.* (2007) highlight the relationship between 'grit'—defined as perseverance and passion for long-term goals—and educational achievement. Barón and Cobb-Clark (2010) find a positive relationship between internal locus of control (an individual's belief that she has control over her fate) and educational attainment. Similarly, Heckman *et al.* (2006) show that self-esteem and locus of control are related to earnings, incarceration, and teen pregnancy.

If these traits are important, the question then is whether educational interventions can move these abilities in ways that then affect longer-term outcomes. In one of the few studies to examine this question experimentally, Blackwell *et al.* (2007) first demonstrate a positive correlation between students' belief that intelligence is malleable and their mathematics achievement. They then build on their observational findings to design an intervention among US middle school students aimed at shifting beliefs about intelligence. They find that teaching students that intelligence is malleable improves motivation and mathematics grades. Future experiments should take a similar approach, building on findings from correlational studies to design prospective interventions examining the long-term causal impact of non-cognitive abilities.

In this vein, Heller *et al.* (2013) test a non-cognitive intervention among US middle and secondary students that are at risk for violence and dropping out of school. The intervention builds on techniques from cognitive behavioural therapy (CBT) to teach adolescents social-cognitive skills such as self-control and anger management. They find that programme participation dramatically reduces crime rates and increases engagement in school. When combined with small-group intensive tutoring, the programme also increases test scores (Cook *et al.*, 2014). This intervention is an important example for the future direction that experimentation in education should take.

[5] In related work, Heckman and Rubinstein (2001) find that individuals who receive a high school equivalence General Educational Development (GED) certificate earn less than peers of equal cognitive ability but who receive a high school diploma. They argue that the gap is due to unobserved differences in non-cognitive ability.

It targets a group of students who are largely unresponsive to programmes aimed at the broader student population. And, it then designs an intervention matched to their particular abilities and needs. Future experiments should explore how to better individualize interventions and match students to differentiated programmes. For, the more targeted interventions are, the more (cost-) effective they will be.

III. The policy-maker's role

(i) Addressing concerns and limitations of experiments in education

As discussed above, experimentation can shed light on questions that are difficult to examine using observational data and can inform the design of key educational programmes and policies. Despite its promise, however, experimentation often meets resistance. Given the large stakes, it is understandable that many educators and parents are cautious about using this relatively new and unfamiliar tool in schooling. A key role for policy-makers is to communicate the importance of experimentation and address stakeholders' concerns. Below, I discuss responses to four common challenges to experimentation in education.

Experimentation is unethical or unfair

Policy-makers and educators sometimes face the objection that, 'it is wrong to experiment on children'. This objection can be especially strong if the intervention being tested is unconventional in education, such as offering students financial incentives. A related concern is that it is not fair to treat people differently. That is, you should not offer a programme to some (i.e. the treatment group) and not to the others (i.e. the control group). Or alternatively, that it is unethical to implement a programme that may not work.

There are several responses to this concern. First, the alternative is to not experiment, which means continuing to spend resources on programmes that are ineffective while failing to develop programmes that could improve educational outcomes and ultimately the lives of millions of children. Second, experimentation is already occurring in unrecognized ways. Changes are constantly taking place at all levels of education policy, from national reforms to classroom instruction. We should implement these changes in ways that allow us to measure their impact and thus inform the direction that policy should take. Third, while the terms 'experimentation' and 'randomization' often raise concerns, educators, parents, and students are familiar and comfortable with the idea that a new programme has limited space or resources and therefore who receives the programme is determined by random lottery. Similarly, education policies are often rolled out gradually, for example to inform logistics or programme development. Such gradual rollouts can be made more informative (and arguably *more* fair) if within the eligible population the order of implementation is determined by random assignment.

Finally, it is sometimes helpful to consider the medical model in which a new vaccine, drug, or therapy is tested experimentally before it can be introduced widely. Several principles guide this model. First, do no harm. It is important to recognize that those who do not receive an experimental intervention are made no worse off than they were under the *status quo*. And for those who do receive the intervention, the evaluation process should always include measures of potential negative impacts. For example, as discussed in section II(ii), studies of incentive interventions track student effort, motivation, and achievement in order to assess 'crowding out' of intrinsic motivation (to date there is little evidence of such detrimental effects). The other central principle is that interventions must demonstrate efficacy before becoming policy. This ultimately serves children by steering resources towards programmes that work and limiting students' exposure to those that do not.

Experiments are logistically difficult, disruptive, and costly

A second concern with experiments is that schools do not have the resources necessary to run them and that they will disrupt the regular functioning of the classroom. Here again it is important to highlight what is taking place in the absence of experimentation. Any programme (whether or not it is experimental) requires resources to run successfully, and schools should be given those. Typically, the only added logistics and costs of an experiment are randomization, which can be implemented cheaply and easily, and data collection of student outcomes which is usually already occurring (or should be, even in the absence of experiments).

It is important to recognize, however, that new interventions often require additional resources. Researchers should work closely with schools to ensure that new interventions are designed in ways that are practical to implement and minimize disruption. And, as I discuss in more detail in the section below, policy-makers should provide the resources necessary to support implementation, data collection, and analysis.

Finally, the costs of experiments are dwarfed by the resources necessary for policy implementation. By helping to optimally design interventions at a small scale and only implementing what works, experimentation ultimately saves money. It also avoids the much larger disruption that takes place when old policies are overhauled because they are discovered to be ineffective and new policies are installed without having been properly tested.

Most experiments fail

Perhaps the greatest frustration with experiments in education is that it is very difficult to identify interventions that are effective. In section II, I discussed several highly successful programmes, but for every one of these there are many more that find little or no impact. However, null results should not be considered a failure. Learning which programmes are not working is crucial to guiding policy decision-making. Again, it is important to highlight the alternative, which is to continue spending resources on programmes that do not work. And such experiments often inform the design of future interventions.

Still, it can be politically challenging to garner resources for programmes knowing that they may ultimately prove ineffective. Policy-makers and researchers need to manage expectations accordingly, emphasizing the importance of experimentation to learn. At the same time, they should recognize the temptation to demonstrate the success of programmes in which they are invested and take this into account when determining guidelines for policy evaluations of experimental results. Finally, as I discuss below, researchers should design experiments to be informative regardless of the overall programme impact.

Experiments can only answer one question at a time

A related frustration with experiments is that they can include a large population of students and take a long period of time (e.g. a school year) and then primarily answer only one question: what is the average effect of the tested intervention on the measured outcome(s)? While this is a critical question, we often want to answer a richer set of questions, such as: Who benefits most? *Why* is the programme effective (or ineffective)? How would the effects change if we altered the intervention or implemented it in a different context among a different population—i.e. how do the results generalize?

Researchers and policy-makers can address this concern by designing experiments to answer a richer set of questions. As discussed in the section below, experiments in education will become more valuable if they are motivated by and inform our broader understanding of the education production function. For example, an intervention that compares the effect of offering incentives to parents to the effect of offering incentives to students is not only testing the most cost-effective way to structure rewards, but is also exploring questions about the parent–child dynamic in the household. Experiments should also be supplemented by surveys and secondary data analysis. For example, surveys about what is taking place in the household (e.g. how parents use their rewards to motivate their children) can shed light on the dynamic driving the overall treatment effects (or lack thereof). Secondary data analysis can help identify differential treatment effects in the population. These methods can help inform the design of future interventions. For example, even if the average effect of a programme is close to zero, there may be a sub-group that particularly benefits. Researchers should seek to identify this population and then prospectively design experiments to test programme impacts in key sub-groups.

(ii) Decision-making through experimentation

Perhaps the most important step policy-makers can take is to shift the paradigm of how decisions in education are made. To do this, policy-makers and educators, as well as families, need to recognize first that they are constantly making choices, whether it is to implement a new policy or to continue with an existing one; second, that to every extent possible these choices should be informed by an understanding of the impact these policies have on students; and third, that randomized experiments are the 'gold standard' for measuring these impacts.

In order to then base policy decisions on experimentation, there needs to be a strong foundation of experimental evaluations to draw upon. Below I discuss in more detail how to develop this foundation. Broadly, policy-makers should incentivize and enable experimentation by rewarding or requiring that interventions be tested using randomized experiments and by providing researchers and schools with the necessary political, financial, human, and technological resources. These experiments should cast a wide net to include not only traditional programme evaluation but also interventions that inform our understanding of the education production function more broadly.

Encourage, facilitate, and reward experimentation

Policy-makers can start by giving more decision weight and resources to programmes that demonstrate effectiveness through rigorous evaluations using randomized experiments. When crafting new interventions or considering an education reform, policy-makers should incorporate knowledge built up from experimental evaluations, and, wherever possible, incorporate experimental evaluations into the rollout of any policy. Facilitating such experimentation includes communicating the importance of experimentation (and addressing the kinds of concerns discussed in the previous section), sufficiently funding new interventions, developing databases that longitudinally track student outcomes, and supporting the human resources necessary to design and test experimental interventions. Below, I discuss the policy-maker's role in each of these areas—noting that much of this is already taking place through national programmes such as Race to the Top, the Institute of Education Sciences (IES), and the Education Endowment Fund (EEF), along with states, local school district authorities, and private foundations.

In communicating the importance of experimentation, one area where there is probably the greatest need is in encouraging schools to test experimental interventions. Many school administrators are unfamiliar with experimentation as a tool in education (or may have some of the concerns discussed in the section above). Policy-makers are beginning to increase awareness through the kinds of experimental toolkits discussed in section II(i). A next step is to link experimentation to added resources. For example, if a school or district requests support for a new programme (or an existing programme that has not been rigorously evaluated), whenever possible such support should be conditional on incorporating a randomized experiment into the implementation. Similarly, if researchers or programme developers propose testing a policy intervention, priority should be given to those who will do so using a randomized experiment. Policy-makers can facilitate such experiments by connecting researchers to school administrators who are interested in testing new programmes.

On the funding side, it is critical that these programmes are able to enrol sample sizes large enough to ensure adequate statistical power. This is particularly important in education because, as discussed in the section above, most experiments 'fail'—that is, they do not demonstrate a statistically significant programme effect. If this occurs and a study is underpowered, it may actually be the case that the programme is effective but that the sample size was not large enough to detect its impact. It is then very difficult to use the experimental results to inform decision-making. Should the programme be abandoned or should it be re-tested with a larger sample? An added challenge is that because educational interventions are often randomized at the class or school level—i.e. an entire class or school receives the treatment—these studies require large numbers of students to be adequately powered (see List *et al*. (2011) for further discussion of optimal sample design).

Equally critical to experimentation is data collection, access, and management. These data are most valuable when they track individuals over time and across a wide range of outcome measures. As discussed in section II, the few studies able to do this have demonstrated the impact of education (stretching back to early childhood) on a host of adult outcomes, including employment, health, marriage, criminal behaviour, and self-reported well-being. The next step is to identify the mechanisms —i.e. what are students learning in school—that drive these effects. For example, recent studies have suggested that one key driver may be non-cognitive skills, which are not traditionally measured in standard school administrative data. Providing resources to collect such additional measures can make experimental evaluations more informative and can help shape the design of educational interventions going forward. More broadly, investment in good data systems is a shared resource that benefits both researchers and practitioners.

Finally, high-quality experimentation requires human resources that are not typically present in schools. Many educators welcome opportunities to experiment (particularly when it is tied to additional resources or new programmes). However, they lack expertise in experimental design and programme evaluation, and often worry that programme management will overburden their teachers and staff. Policy-makers can play a critical role by facilitating partnerships with researchers, who can then properly design, implement, and evaluate experiments.

Expand experimentation beyond traditional programme evaluation

The richest experiments not only evaluate the effectiveness of a particular intervention but also shed light on our understanding of the production function. Section II discusses several examples, including experimental designs that explore the role of information, complementarities, effort, non-cognitive abilities, and preferences. These studies can then inform policy decision-making more broadly. Databases that compile and report experimental results are an important first step in doing this. A next step is to organize these findings within a framework for the education production function, and then to use this framework to inform intervention design and policy decisions. I discuss below three ways that experimentation can serve this aim by moving beyond traditional programme evaluation.

(1) Incorporate insights from a wide range of disciplines including child development, psychology, and behavioural economics.

As discussed in section II, these insights should be incorporated into the standard framework for the education production function, and then inform the design of interventions. For example, insights from behavioural economics suggest that near-term incentives—both financial and non-financial—can increase student effort. An open question, however, is whether incentives can move educational investments in ways that impact longer-term achievement. To address this, a recent incentive programme for secondary students in the UK offers students rewards for effort 'inputs' with the aim of improving end of the year performance 'outputs' on a high-stakes exam; and within the programme, compares the effect of financial and non-financial incentives (Education Endowment Fund, 2014*b*). The findings from this study will advance our understanding of how student effort responds to various incentives, what types of effort have an impact on longer-term achievement, and the extent to which incentive programmes generalize across settings at a policy-level scale.

(2) Use experimentation to target and individualize interventions

As discussed in the previous section, experiments are more valuable when they not only identify average treatment effects but also shed light on the distributional impacts of a policy. This can increase the (cost-) effectiveness of interventions by allowing us to target individuals who benefit most (and limit exposure among those who may experience negative impacts). Going a step farther, experimentation can be used to design and evaluate individualized interventions, which are tailored to a particular child's abilities and needs.

Researchers should combine experimental and non-experimental methods to address the policy goal of successfully differentiating educational interventions among a diverse student population.

The use of experimentation to understand the impact of peers on student achievement provides an illustration. Non-random selection into peer groups makes it difficult to causally identify peer effects in observational data. Previous studies have examined average peer effects in naturally occurring data that contain random or quasi-random assignment of peer groups in various contexts, including classrooms, college housing, and military academy squadrons (Sacerdote (2011) provides a review). Carrell *et al.* (2013) take this literature a step further. Among randomly assigned groups at the US Air Force Academy, they analyse peer effects along different parts of the ability distribution. They then use this analysis to generate peer groups predicted to be 'optimal' for students of all abilities, and finally design a prospective experiment that compares randomly assigned peer groups to 'optimally' assigned peer groups. Interestingly they find that the 'optimal' composition groups do not perform as predicted because students sort themselves by ability into non-optimal sub-groups. While the intervention was not ultimately successful, this study provides a useful model for how experimentation can be used to understand distributional effects in ways that inform the design of future interventions. It also sheds new light on the dynamics of sorting and its impact on peer effects. More broadly, it illustrates two important lessons for policy-makers. First, that 'failed' experiments are still informative. And second, that interventions suggested by secondary data analysis (of either observational or experimental data) should be tested experimentally as they may perform very differently than intended.

(3) As in the medical model, include not only tests of programme effectiveness but also 'basic research' as well as evaluations of efficacy at a policy level.

Rather than test a policy ready intervention, basic research focuses on identifying particular mechanisms in the education production function. Because these mechanisms generalize across contexts, the knowledge gained from such experiments can then be applied to a wide range of policy questions. For example, we know that most education policies will fail if students do not exert effort. In traditional programme evaluation, however, it is very difficult to disentangle student effort from student ability. For example, if a student performs badly on a test, is it because she does not understand the material or because she was not motivated to answer the questions correctly? And if lack of motivation is a factor, what can be done about it—for example, how can we increase student engagement in class or student effort on assignments? As discussed in section II(ii), Levitt *et al.* (2012) address these questions by isolating the role of effort (apart from ability) and then examine students' effort response to various incentive designs. Their study demonstrates that effort is often sensitive to the timing, size, and type of rewards that students experience; and it also shows how incentive effects vary by context and student characteristics. While not itself a programmatic intervention, basic research like the Levitt *et al.* (2012) study can help educators identify when students may lack motivation and how best to increase student engagement and effort.

Basic research may also use non-traditional methods such as lab-in-the-field type experiments. For example, the research discussed in section II(iii) that explores the relationship between preferences, non-cognitive abilities, and achievement often uses techniques first developed in laboratory studies. These methods are now being incorporated into a wide range of programme evaluations aimed at understanding the effect of interventions on non-cognitive measures, as well as how programme impacts vary based on underlying student abilities.

One advantage these studies have is that they tend to be less costly and less disruptive than traditional programme evaluation. This allows researchers to experiment with a wide range of intervention designs that can then be narrowed down before testing them in the field. For example, as discussed in section II(i), there is great interest in designing incentives to select high-quality teachers. One constraint on this research is that it is very difficult to experiment with teacher pay, particularly over the time and space required to examine the impact on selection into the profession. Before implementing a policy intervention, therefore, it is important to better understand the decision-making process of individuals who are considering becoming teachers. Toward this end, a recent study among university students in the UK performed an online survey that used experimental methods to examine the relationship between students' willingness to enter teaching and their underlying preferences, abilities, and responsiveness to various incentive designs (Dolan *et al.*, 2012). Their findings will then inform the ultimate design of incentive interventions tested in the field.

More broadly, the insights gained from basic research can be used to design policy programmes, which, as in the medical model, should be tested at a smaller scale for efficacy and, if promising, scaled up to examine generalizability, feasibility, and (cost-) effectiveness at a policy level. Just as policy-makers should create a greater role for basic research, they should also ensure that there is support and funding for studies at the other end of the spectrum, which examine replicability and scalability. While it is tempting to fully implement interventions that have demonstrated success in smaller-scale programme evaluations, it is critical for policy-makers to understand how an intervention operates in a wide range of settings when implemented by practitioners rather than researchers. In section II, I discuss several areas where this is taking place, including policy-level evaluations of early education programmes in the US as well as incentive and informational interventions currently being tested at a policy level in the UK.

IV. Conclusion

The quality of a child's education has long-term consequences for almost every important life outcome. Currently, large numbers of students—particularly those from disadvantaged backgrounds—are failing to acquire the educational attainment and basic skills necessary to succeed in today's economy. Experimentation can help us evaluate which policies are working for children and which are falling short, identify areas that need improvement, and develop new interventions to address these needs.

There are several critical policy areas where experimentation can potentially make great progress. These include: implementing national and statewide education reforms effectively; improving teacher and principal quality through training and incentives; increasing positive parent and peer engagement; designing incentives or other interventions that promote habit formation; understanding what the critical non-cognitive skills are, how to move them, and their causal impact on later life outcomes; and targeting and individualizing interventions.

The design of experiments addressing these questions should draw on a wide range of disciplines including child development, psychology, and behavioural economics. Insights from these areas can help identify underlying mechanisms of the education production function and inform the design of interventions in ways that increase (cost-) effectiveness. Finally, there should be a rich array of experiments in education, ranging from lab-like basic research to policy level efficacy trials. Together, this can help researchers, educators, and policy-makers develop new interventions to address pressing policy goals such as increasing educational investments and closing achievement gaps.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., and Lun, J. (2011), 'An Interaction-based Approach to Enhancing Secondary School Instruction and Student Achievement', *Science* **333**(6045),1034–7.

Angrist, J. D., and Lavy, V. (2009), 'The Effect of High-stakes High School Achievement Awards: Evidence from a Group-randomized Trial', *American Economic Review*, **99**(4), 1384–414.

— Bettinger, E., and Kremer, M. (2006), 'Long-term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia', *American Economic Review*, **96**(3), 847–62.

— Lang, D., and Oreopoulos, P. (2009), 'Incentives and Services for College Achievement: Evidence from a Randomized Trial', *American Economic Journal: Applied Economics*, **1**(1), 136–63.

— Oreopoulos, P., and Williams, T. (2014), 'When Opportunity Knocks, Who Answers?', *Journal of Human Resources*, **49**(3), 572–610.

— Pathak, P. A., and Walters, C. (2013), 'Explaining Charter School Effectiveness', *American Economic Journal: Applied Economics*, **5**(4), 1–27.

Avvisati, F., Besbas, B., and Guyon, N. (2010), 'Parental Involvement in Schools: A Literature Review', *Review d'Economie Politique*, **120**(5), 761–78.

— Gurgand, M., Guyon, N., and Maurin, E. (2014), 'Getting Parents Involved: A Field Experiment in Deprived Schools', *Review of Economic Studies*, **81**(1), 57–83.

Banerji, R., Berry, J., and Shotland, M. (2013), 'The Impact of Mother Literacy and Participation Programs on Child Learning: Evidence from a Randomized Evaluation in India', Working Paper, available at http://www.povertyactionlab.org/publication/impact-mother-literacy-and-participation-programs-child-learning-evidence-randomized-eva

Barón, J. D., and Cobb-Clark, D. A. (2010), 'Are Young People's Educational Outcomes Linked to their Sense of Control?', IZA Discussion Papers 4907, Institute for the Study of Labor (IZA).

Barrera-Osorio, F., Bertrand, M., Linden, L. L., and Perez-Calle, F. (2011), 'Improving the Design of Conditional Transfer Programs: Evidence from Randomized Education Experiment in Colombia', *American Economic Journal: Applied Economics*, **3**, 167–95.

Barrow, L., and Rouse, C. E. (2013), 'Financial Incentives and Educational Investment: The Impact of Performance-based Scholarships on Student Time Use', NBER Working Paper 19351.
— Richburg-Hayes, L., Rouse, C. E., and Brock, T. (2014), 'Paying for Performance: The Education Impacts of a Community College Scholarship Program for Low-income Adults', *Journal of Labor Economics*, **32**(3), 563–99.
Behrman, J. R., Sengupta, P., and Todd, P. (2005), 'Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico', *Economic Development and Cultural Change*, **54**(1), 237–75.
Bergman, P. (2012), 'Parent–Child Information Frictions and Human Capital Investment: Evidence from a Field Experiment', Working Paper, available at http://www.columbia.edu/~psb2101/BergmanSubmission.pdf
Berry, J. (2014), 'Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India', *Journal of Human Resources*, forthcoming.
Bettinger, E. (2012), 'Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores', *Review of Economics and Statistics*, **94**(3), 686–98.
— Slonim, R. (2007), 'Patience among Children', *Journal of Public Economics*, **91**(1–2), 343–63.
— Long, B. T., Oreopoulos, P., and Sanbonmatsu, L. (2012), 'The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment', *Quarterly Journal of Economics*, **127**(3), 1205–42.
Blackwell, L. S., Trzesniewski, K. H., and Sorich Dweck, C. (2007), 'Implicit Theories of Intelligence Predict Achievement across an Adolescent Transition: A Longitudinal Study and an Intervention', *Child Development*, **78**(1), 246–63.
Borghans, L., Duckworth, A. L., Heckman, J. J., and ter Weel, B. (2008), 'The Economics and Psychology of Personality Traits', *Journal of Human Resources*, **43**(4), 972–1059.
Boyd, D., Goldhaber, D., Lanjiford, H., and Wyckoff, J. (2007), 'The Effect of Certification and Preparation on Teacher Quality', *The Future of Children*, **17**(1), 45–68.
Branch, G. F., Hanushek, E., and Rivkin, S. (2012), 'Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals', National Center for Analysis of Longitudinal Data in Education Research, Working Paper 32.
Bursztyn, L., and Coffman, L. C. (2012), 'The Schooling Decision: Family Preferences, Intergenerational Conflict, and Moral Hazard in the Brazilian Favelas', *Journal of Political Economy*, **120**(3), 359–97.
Campbell, P. F., and Malkus, N. N. (2011), 'The Impact of Elementary Mathematics Coaches on Student Achievement', *The Elementary School Journal*, **111**(3), 430–54.
Carrell, S. E., and Sacerdote, B. (2013), 'Late Interventions Matter Too: The Case of College Coaching New Hampshire', NBER Working Paper 19031.
— West, J. E. (2013), 'From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation', *Econometrica*, **81**(3), 855–82.
Castillo, M., Ferraro, P., Jordan, J. L., and Petrie, R. (2011), 'The Today and Tomorrow of Kids: Time Preferences and Educational Outcomes of Children', *Journal of Public Economics*, **95**(11–12), 1377–85.
Castleman, B. L., and Page, L. C. (2013), 'Summer Nudging: Can Personalized Text Messages and Peer Mentor Outreach Increase College Going Among Low-income High School Graduates?', Center for Education Policy and Workforce Competitiveness Working Paper 9.
Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014), 'Measuring the Impacts of Teachers II: Teacher Value-added and Student Outcomes in Adulthood', *American Economic Review*, **104**(9).
— Hilger, N., Saez, E., Whitmore Schanzenbach, D., and Yagan, D. (2011), 'How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star', *Quarterly Journal of Economics*, **126**(4), 1593–660.
Chingos, M. M., and Peterson, P. E. (2012), 'The Effects of School Vouchers on College Enrollment', Brown Center on Education Policy at Brookings and Harvard Kennedy School.
Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., and Wheeler, J. (2007), 'High Poverty Schools and the Distribution of Teachers and Principals', *North Carolina Law Review*, **85**(June), 1345–80.

Coalition for Evidence-Based Policy (2003), 'Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User-friendly Guide', US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Cook, P. J., Dodge, K., Farkas, G., Fryer, Jr, R. G., Guryan, J., Ludwig, J., Mayer, S., Pollack, H., and Steinberg, L. (2014), 'The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results from a Randomized Experiment in Chicago', NBER Working Paper 19862.

Cullen, J. B., Levitt, S. D., Robertson, E., and Sadoff, S. (2013), 'What Can be Done to Improve Struggling High Schools?', *Journal of Economic Perspectives*, **27**(2), 133–52.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999), 'A Meta-analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation', *Psychological Bulletin*, **125**(6), 627–68.

Dee, T. S. (2011), 'Conditional Cash Penalties in Education: Evidence from the Learnfare Experiment', *Economics of Education Review*, **30**(5), 924–37.

De Paola, M., Scoppa, V., and Nisticò, R. (2012), 'Monetary Incentives and Student Achievement in a Depressed Labor Market: Results from a Randomized Experiment', *Journal of Human Capital*, **6**(1), 56–85.

Dobbie, W., and Fryer, R. G. (2013), 'Beneath the Veil of Effective Schools: Evidence from New York City', *American Economic Journal: Applied Economics*, **5**(4), 28–60.

Dolan, P., Metcalfe, R., and Navarro-Martinez, D. (2012), 'Financial Incentives and Working in the Education Sector', UK Department for Education Research Centre for the Understanding of Behaviour Change Research Report DFE-RR251.

Duckworth, A. L., Peterson, C., Matthews, M. D., and Kelly, D. R. (2007), 'Grit: Perseverance and Passion for Long-term Goals', *Journal of Personality and Social Psychology*, **92**(6), 1087.

Duflo, E., Hanna, R., and Ryan, S. P. (2012), 'Incentives Work: Getting Teachers to Come to School', *American Economic Review*, **102**(4), 1241–78.

Duncan, G. J., and Magnuson, K. (2013), 'Investing in Preschool Programs', *Journal of Economic Perspectives*, **27**(2), 109–32.

Dynarski, S., Hyman, J., and Schanzenbach, D. (2013), 'Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion', *Journal of Policy Analysis and Management*, **34**(2), 692–717.

Education Endowment Fund (2014*a*), accessed 1 September: http://educationendowmentfoundation.org.uk/projects/texting-parents-bristol-university-and-harvard-university/

— (2014*b*), accessed 1 September: http://educationendowmentfoundation.org.uk/projects/pupil-effort-programme/

Fryer, Jr, R. G. (2011*a*), 'Injecting Successful Charter School Strategies into Traditional Public School: Early Results from an Experiment in Houston', NBER Working Paper 17494.

— (2011*b*), 'Teacher Incentives and Student Achievement: Evidence from New York City Public Schools', NBER Working Paper 16850.

— (2011*c*), 'Financial Incentives and Student Achievement: Evidence from Randomized Trials', *Quarterly Journal of Economics*, **126**(4), 1755–98.

— (2012), 'Multitasking, Learning, and Incentives: A Cautionary Tale', NBER Working Paper 17752.

— (2013), 'Information and Student Achievement: Evidence from a Cellular Phone Experiment', NBER Working Paper 19113.

— Levitt, S. D., List, J., and Sadoff, S. (2012), 'Enhancing the Efficacy of Teacher Incentives Through Loss Aversion: A Field Experiment', NBER Working Paper 18237.

Glazerman, S., Protik, A., Teh, B., Bruch, J., Max, J., and Warner, E. (2013), 'Transfer Incentives for High-performing Teachers: Final Results from a Multisite Experiment', NCEE 2014–4003, Washington, DC, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Glewwe, P., Ilias, N., and Kremer, M. (2010), 'Teacher Incentives', *American Economic Journal: Applied Economics*, **2**(3), 205–27.

Gneezy, U., Meier, S., and Rey-Biel, P. (2011), 'When and Why Incentives (Don't) Work to Modify Behavior', *Journal of Economic Perspectives*, **25**(4), 191–210.

Goodman, S. F., and Turner, L. J. (2013), 'The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program', *Journal of Labor Economics*, **31**(2), 409–20.

Gruber, J. (2001), *Risky Behavior Among Youth: An Economic Analysis*, Chicago, IL, University of Chicago Press.

Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., and Hyde, J. S. (2012), 'Helping Parents to Motivate Adolescents in Mathematics and Science: An Experimental Test of a Utility-value Intervention', *Psychological Science*, 0956797611435530.

Hastings, J. S., and Weinstein, J. M. (2008), 'Information, School Choice, and Academic Achievement: Evidence from Two Experiments', *Quarterly Journal of Economics*, **123**(4), 1373–414.

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

— Stixrud, J., and Urzua, S. (2006), 'The Effects of Cognitive and Non-cognitive Abilities on Labor Market Outcomes and Social Behavior', *Journal of Labor Economics*, **24**(3), 411–82.

Heller, S., Pollack, H. A., Ander, R., and Ludwig, J. (2013), 'Preventing Youth Violence and Dropout: A Randomized Field Experiment', NBER Working Paper 19014.

Hoxby, C. M., and Murarka, S. (2009), 'Charter Schools in New York City: Who Enrolls and How They Affect their Students' Achievement', NBER Working Paper 14852.

— Turner, S. (2013), 'Expanding College Opportunities for High-achieving, Low Income Students', Stanford Institute for Economic Policy Research Discussion Paper 12–014.

Jacob, B. A. (2011), 'Do Principals Fire the Worst Teachers?', *Educational Evaluation and Policy Analysis*, **33**(4), 403–44.

Jensen, R. (2010), 'The (Perceived) Returns to Education and the Demand for Schooling', *Quarterly Journal of Economics*, **125**(2), 515–48.

Kirby, K. N., Winston, G. C., and Santiesteban, M. (2005), 'Impatience and Grades: Delay-discount Rates Correlate Negatively with College GPA', *Learning and Individual Differences*, **15**(3), 213–22.

— Godoy, R., Reyes-Garcia, V., Byron, E., Apaza, L., Leonard, W., Perez, E., Vadez, V., and Wilkie, D. (2002), 'Correlates of Delay-discount Rates: Evidence from Tsimane' Amerindians of the Bolivian Rain Forest', *Journal of Economic Psychology*, **23**(3), 291–316.

Kohn, A. (1999), *Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes*, Boston, MA, Houghton Mifflin.

Kraft, M. A., and Dougherty, S. M. (forthcoming), 'The Effect of Teacher–Family Communication on Student Engagement: Evidence from a Randomized Field Experiment', *Journal of Research on Education Effectiveness*.

— Rogers, T. (in preparation), 'Teacher-to-Parent Communication: Experimental Evidence from a Low-Cost Communication Policy'.

Kremer, M., Miguel, E., and Thornton, R. (2009), 'Incentives to Learn', *Review of Economics and Statistics*, **91**(3), 437–56.

Krueger, A. B. (1999), 'Experimental Estimates of Education Production Functions', *Quarterly Journal of Economics*.

Lankford, H., Loeb, S., and Wyckoff, J. (2002), 'Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis', *Educational Evaluation and Policy Analysis*, **24**(1), 37–62.

Lavecchia, A., Liu, H., and Oreopoulos, P. (2014), 'Behavioral Economics of Education: Progress and Possibilities', NBER Working Paper 20609.

Levitt, S. D., List, J. A., and Sadoff, S. (2010), 'The Effect of Performance-based Incentives on Educational Achievement: Evidence from a Randomized Experiment', Working Paper.

— Neckermann, S., and Sadoff, S. (2012), 'The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance', NBER Working Paper 18165.

Leuven, E., Oosterbeek, H., and Van der Klaauw, B. (2010), 'The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment', *Journal of the European Economic Association*, **8**(6), 1243–56.

— Sonnemans, J., and Van Der Klaauw, B. (2011), 'Incentives versus Sorting in Tournaments: Evidence from a Field Experiment', *Journal of Labor Economics*, **29**(3), 637–58.

List, J. A., Livingston, J. A., and Neckermann, S. (2012), 'Harnessing Complementarities in the Education Production Function', Working Paper, available at https://econresearch.uchicago.edu/sites/econresearch.uchicago.edu/files/Livingston_AWFE%20version.pdf

— Sadoff, S., and Wagner, M. (2011), 'So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design', *Experimental Economics*, **14**(4), 439–57.

MacDonald, H., Bernstein, L., and Price, C. (2009), 'Foundations for Success: Short-term Impacts Report', Report to the Canada Millennium Scholarship Foundation, Ottawa.

Metcalfe, R., Burgess, S., and Proud, S. (2011), 'Student Effort and Educational Attainment: Using the England Football Team to Identify the Education Production Function', Centre for Market and Public Organization Working Paper.

Mischel, W., Shoda, Y., and Rodriguez, M. L. (1989), 'Delay of Gratification in Children', *Science*, **244**(4907), 933–8.

Muralidharan, K., and Sundararaman, V. (2011), 'Teacher Performance Pay: Experimental Evidence from India', *Journal of Political Economy*, **119**(1).

Neal, D. (2011), 'The Design of Performance Pay in Education', in E. A. Hanushek, S. J. Machin, and L. Woessmann (eds), *Handbook of the Economics of Education*, Elsevier.

Nguyen, T. (2008), 'Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar', Working Paper, available at http://www.povertyactionlab.org/doc/information-role-models-and-perceived-returns-education

Oreopoulos, P., and Salvanes, K. G. (2011), 'Priceless: The Nonpecuniary Benefits of Schooling', *Journal of Economic Perspectives*, **25**(1), 159–84.

Patel, R., Richburg-Hayes, L., de la Campa, E., and Rudd, T. (2013), 'Performance-based Scholarships: What Have We Learned? Interim Findings from the PBS Demonstration', New York, MDRC.

Rockoff, J. E., Jacob, B. A.,Kane, T. J., and Staiger, D. O. (2011), 'Can You Recognize an Effective Teacher when You Recruit One?', *Education Finance and Policy*, **6**(1), 43–74.

Rodriguez-Planas, N. (2012), 'Longer-term Impacts of Mentoring, Educational Services, and Learning Incentives: Evidence from a Randomized Trial in the United States', *American Economic Journal: Applied Economics*, **4**(4), 121–39.

Sacerdote, B. (2011), 'Peer Effects in Education: How Might They Work, How Big Are They and How Much do We Know thus Far?', *Handbook of the Economics of Education*, **3**, 249–77.

Sanders, M., and Chande, R. (2015), 'Are there Friction Costs in the UK University Market? Evidence from a Large National Field Experiment', Centre for Market and Public Organisation Working Paper.

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

Sharma, D. (2010), 'The Impact of Financial Incentives on Academic Achievement and Household Behavior: Evidence from a Randomized Trial', manuscript, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1681186

Sirin, S. R. (2005), 'Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research', *Review of Educational Research*, **75**(3), 417–53.

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J. R., McCaffrey, D. F., Pepper, M., and Stecher, B. M. (2010), 'Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT)', Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Steinberg, L., Graham, S., O'Brien, L., Woolard, J., Cauffman, E., and Banich, M. (2009), 'Age Differences in Future Orientation and Delay Discounting', *Child Development*, **80**(1), 28–44.

Thaler, R. H., and Sunstein, C. R. (2008), *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New Haven, CT, Yale University Press.

Wolf, P. J., Kisida, B., Gutmann, B., Puma, M., Eissa, N., and Rizzo, L. (2013), 'School Vouchers and Student Outcomes: Experimental Evidence from Washington, DC', *Journal of Policy Analysis and Management*, **32**(2), 246–70.