Department for Education

Research Report DFE-RR215a

Evaluation of the City Challenge programme: appendices

Appendix A Targets set in each City Challenge area

Appendix B Research and Policy review

Appendix C Assessing the Impact of KTS interventions

Merryn Hutchings[1], Charley Greenwood[2], Sumi Hollingworth[1], Ayo Mansaray[1] and Anthea Rose[1], with Sarah Minty[1] and Katie Glass[1]

[1]Institute for Policy Studies in Education, London Metropolitan University

[2]Evaluation and Research Practice, Coffey International Development

This research report was commissioned before the new UK Government took office on 11 May 2010. As a result the content may not reflect current Government policy and may make reference to the Department for Children, Schools and Families (DCSF) which has now been replaced by the Department for Education (DfE).

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

These targets were modified over the course of the programme. In the Black Country, for example, the targets 'double' or 'cut by half' were changed to 'increase' or 'reduce'. Stakeholders recognised that with the change in Ofsted framework, the Ofsted-related targets originally set were now unrealistic. Similarly, floor targets changed during the course of the programme. In all areas, particularly in the final year when funding was reduced, schools became the main focus, and the targets relating to Higher Education and employment assumed less importance.

Sources:

DCSF (2008) *Vision for London 2008-11: London education on the way to world class*, Nottingham, DCSF http://www.dcsf.gov.uk/citychallenge/downloads/00481-2008DOM-EN[1].pdf, accessed September 2010.

DCSF (2008) *Raising the bar, closing the gap: a vision and operational strategy for the Greater Manchester Challenge*, Nottingham, DCSF http://www.dcsf.gov.uk/citychallenge/downloads/Manchester_Challenge.pdf, accessed September 2010.

DCSF (2008) *We're up for the Black Country Challenge: Are you?* http://www.theblackcountrychallenge.co.uk/, accessed September 2010.

APPENDIX B Research and Policy Review

1 Introduction

This review introduces the research bases that are relevant to this evaluation (school effectiveness and school improvement), and identifies the key processes and dimensions of school improvement such as capacity building and accountability, leadership, networking and collaboration, system leadership and the challenges faced by schools in disadvantaged areas. In addition, it reviews the ways in which policy has focused on school improvement; and, drawing on evaluations of school improvement initiatives, identifies some of the issues that were taken into account in designing an appropriate methodology for the evaluation of City Challenge.

2 Research

This evaluation will engage with the two key research bases which underpin current thinking about how to conceptualise and identify the processes and mechanisms which contribute to schools success and improvement. These two strands are; school effectiveness research and school improvement research. We would like to add that whilst proponents writing within these fields argue that 'schools do indeed make a difference' (Macbeath and Mortimore, 2001: 6), others take a more sceptical view of the same evidence. For example, Gorard (2010) has reviewed some of the evidence advanced in favour of the claim that schools make a measurable difference, and found that the most important determinant of variation in performance between schools remains pupil intake (2010: 747). The fields of school effectiveness and improvement are therefore not without their controversies.

2.1 School effectiveness

Since its inception in the 1960s in the USA, the school effectiveness movement has claimed that schools can and do make a difference to educational outcomes. An effective school is defined as one in which students progress further than might be expected from consideration of its intake: 'an effective school thus adds extra value to its students' outcomes, in comparison with other schools serving similar intakes' (Sammons, 2008: 13). According to Macbeath and Mortimore, this value resides at the school level, and refers to the 'value added by the organisation as a whole, through its ethos, culture, policy and planning' (2001: 11).

Whilst there are disagreements about the extent of the 'school effect', there is a 'broad consensus that it lies somewhere in the region of 5 to 15 per cent. That is, with all other factors held constant, there is a 5 to 15 per cent variance between more or less effective schools' (Macbeath and Mortimore, 2001: 6). Researchers in the field have identified the key factors which characterise effective schools: Macbeath and Mortimore (2001: 7) reproduce a list of 11 such characteristics, and a slightly different summary list is provided in Sammon's recent review of the literature (2008: 23). These characteristics have been grouped to illustrate the similarities in the two lists: both include leadership, staff development, a focus on teaching and learning, high expectations, a learning culture or environment, monitoring and parental involvement (Table C1).

Table C1: Key characteristics of effective schools as identified in the literature

McBeath and Mortimore 2001	Sammons 2008
Professional leadership Shared vision and goals	Professional leadership
A learning organisation	Practice-oriented staff development
Purposeful teaching Concentration on learning and teaching Positive reinforcement	Effective instructional arrangements Focus on central learning skills
High expectations	High expectations
A learning environment	Productive climate and culture
Monitoring progress	Appropriate monitoring
Home-school partnership	Parental involvement
Pupil rights and responsibilities	

Whilst in general, effective schools tend to be good for all pupils, there are significant exceptions to this rule. Some schools are more or less effective for different kinds of pupils, boys, girls, particular social and ethnic groups etc. The research also indicates that, even within effective schools, there are likely to be differences in the level of effectiveness across subjects. This is part of what is referred to as ‘differential effectiveness’. Increasingly the school effectiveness movement is paying closer attention to the impact of ‘social mix’ on a school’s effectiveness. Equally, the literature has traditionally focused on highly effective schools, and the dynamics of ineffective schools have not been extensively explored. Ineffective schools are not mirror images of effectiveness and have their own dynamics and complexities.

2.2 School improvement

Although we can characterise an effective school, this does not indicate what is needed to transform an unsuccessful school into a successful one. A widely used definition of school improvement is that it is, a ‘strategy for educational change that enhances student outcomes as well as strengthening the school’s capacity for managing change’ (Hopkins quoted in Macbeath and Mortimore, 2001: 17). According to Harris, ‘what distinguishes the school improvement movement from other school reform efforts is the understanding that it is necessary to focus upon student outcomes in academic performance as the success criteria, rather than teacher perceptions of the innovation’ (Harris quoted in Gray, 2001: 10-11). Similarly, Creemers and Reezight (2005: 359) write: ‘school effectiveness is more directed to finding out ‘‘what works’’ in education and ‘‘why’’; school improvement is practice and policy oriented and intended to change education in the desired direction’. Or, as Sammons observes: ‘one is concerned with identifying and learning more about the features and relationships amongst school and teacher effectiveness factors, while the other seeks to study and generate changes in practice and organisation’ (Sammons, 2009: 128). Historically in the US the school effectiveness movement has informed by, and closely related to, school improvement practice, whereas in the UK, the two have grown in parallel utilising different methodological approaches (Reynolds and Stoll, 1993). For this research, the focus is on school improvement.

There are different conceptualisations of school improvement and what are held to be the underlying processes and dimensions, and how they should be leveraged. In Table C2, we compare the processes of school improvement identified by Brighouse and Woods (1999: 10) with a list of ten key findings from Harris’s (2002) review of school improvement literature, and with processes of school improvement identified in Sammons’ (2008: 35) recent review of the school improvement literature. As with the school effectiveness lists, we have used these simply to illustrate ideas in the field.

Table C2: Key dimensions of school improvement as found in the literature

Brighouse and Woods 1999	Harris 2002	Sammons 2008
The exercise of leadership The practice of management and organisation	Focus on leadership	Clear leadership Redefining structures, frameworks, roles and responsibilities Monitoring, problem-solving and evaluation Using an evolutionary development planning process
The practice of staff development	Focus on teacher development	Staff development and teacher learning
The involvement of parents and the community		Involving pupils, parents and community
The practice of teaching and learning and assessment		Emphasis on teaching and learning
	Focus on student outcomes and performance	
		Celebration of success Developing a shared vision and goal
The creation of an environment most fit for learning		
	Focus on understanding and working with school culture	
The exercise of collective review and self-evaluation		
		External support, networking and partnership
	There is no single blueprint for action and no single approach, different schools different strategies	

Some factors / processes are found in all three lists: for example, leadership and staff development. Others appear in two lists: for example, involvement of parents and the community, and a focus on teaching and learning. All three lists include some factors that appear to be to do with ethos or school culture. Only the most recent (Sammons, 2008) includes external support and networking; this is a key aspect of City Challenge. Reviews by other researchers (Hopkins, 2007; Muijs *et al.*, 2004; Ylimaki *et al.*, 2007), also list additional factors such as hiring staff aligned to the school's mission, creating an information rich environment, accountability and scrutiny, a focus on consistency, coherence and continuity in expectations, curriculum, values and staff. An examination of the Journal *School Effectiveness and School Improvement* illustrates the ways in which thinking about school improvement has developed over the years. Recent issues have focused on a number of issues, such as career and professional development for school leaders, classroom factors, and the impact of pupil demographic change on school performance.

A number of key dimensions and processes underlying school improvement have been identified in the literature, such as leadership, capacity and accountability, school collaboration and networking all of which we discuss later. First, however, we will consider the specific challenges of school improvement in difficult circumstances.

The City Challenge programme aimed to raise standards in some of the most challenging schools in London, Greater Manchester and the Black Country, and break the link between deprivation and underachievement. We therefore examine the literature to investigate the link between deprivation and school improvement.

Levin (2006: 402) points out that 'the improvement process will often be more difficult in schools in challenging circumstances because such schools often face pressure on resources, higher staff turnover, high pupil turnover, lower levels of overt parent support and a history of lack of success.' He points out that where improvement has taken place it is often linked to changes in external conditions such as increased employment (e.g. by Harris *et al*., 2006), and points to US evidence (Bracey, 2004) showing that sustained improvement over time in high poverty schools is highly unusual; research has often focused on case studies of schools that succeed against the odds, rather than the problems that prevent others from bringing about improvement.

One reason for this, according to Levin, is that studies of school improvement have found highly variable levels of commitment to, and implementation of, research-based school improvement strategies, and that a factor in this appears to be that teachers in schools in areas of poverty often see the problem of poor outcomes as residing in the students and their families (e.g. Corbett *et al*., 2002; Dyson, 2004). It has also been suggested that efforts to improve educational outcomes in high-poverty communities should be targeted at non-school issues, such as nutrition, housing, education for parents, etc. (e.g. Rothstein, 2002; Levin, 2006; Earl *et al*., 2003). Research cited by Levin in the US indicates that the majority of schools which sustain improvement tend to be in more affluent areas, and schools in deprived localities are less likely to show consistent improvement over a two year period (2006: 402).

However, the research evidence clearly indicates that the factors which facilitate school improvement for schools in general are equally relevant for schools in challenging circumstances. For example, various studies highlight the importance of leadership, professional development for staff, focus on teaching and learning, collaboration, the use of data, capacity-building and parental involvement (Muijs *et al.*, 2004; Sebring *et al.*, 2006; Ylimaki *et al.*, 2007). Where there is a difference, it is mainly in the emphasis on discipline, safety and order, and a concern that strategies of improvement are designed to be responsive to the contexts of the school and its students (Ainscow *et al.*, 2005). As Reynolds and colleagues observe, 'schools facing challenging circumstances must look for an approach to intervention that ensures a fit between the cultural state of the school and the developmental strategies employed', (2006: 437). The literature suggests that there are common elements in most successful approaches and it these which we turn to next.

2.2.1 Leadership

Leadership is recognised as pivotal in school improvement and effectiveness. It is one of the most powerful mediating factors in school improvement and effectiveness.

> As the key intermediary between the classroom, the individual school and the education system as a whole, effective school leadership is essential to improve the efficiency and equity of schooling. Within each individual school, leadership can contribute to improve student learning by shaping the conditions and climate in which teaching and learning occur. Beyond the school borders, school leaders can connect and adapt schools to changing external environments. And at the school-systems interface, school leadership provides a bridge between internal school improvement processes and externally initiated reform. (Pont *et al.*, 2008b: 16)

Barber and colleagues, in their McKinsey report on school leadership, cite quantitative studies which suggest that 'nearly 60% of a school's impact on student achievement is attributed to principal and teacher effectiveness' (2010: 5). A wide ranging study by Day and colleagues (2009), looked at the impact of school leadership on pupil outcome in schools where that had been identified as having significantly raised pupil attainment levels over a three year period (2003-2005). They found that headteachers educational values and strategic actions improve school conditions which lead to improvements in pupil outcomes. Quantifying the direct impact of school leaders on students' achievement has been more difficult. According to some researchers, 'school leaders have a measurable, mostly indirect influence on learning outcomes,' which tends to be mediated by other people, events, and organisational factors such as school ethos, classroom practices, instructional coherence, etc. (Pont *et al.*, 2008b: 33). Research by Leithwood and Jantzi (2006: 223) found that leadership can have an important influence on teachers' classroom practice, through stimulating and encouraging change which in turn can translate and lead to greater pupil learning. Given this rather indirect impact of school leaders on attainment, a number of researchers have suggested that using attainment as the sole measure potentially narrows our understanding of leadership, and neglects the fact that 'results are strongly conditioned by student background' (Barker, 2007: 38).

For Elmore (2008), 'leadership practice is what connects policy to performance in schools', and it is not the personal attribute of an individual but 'a collection of patterned actions, based on a body of knowledge, This practice is embedded in particular institutional structures, settings and incentive structures, which are difficult to generalise from, so for example, schools cannot simply import leadership models from the private sector.

2.2.2 *Capacity and Accountability*

At its most basic capacity is a resource within schools that refers to the skills and dispositions of staff that can be developed. The concept of 'capacity' has increasingly been used within the literature to capture the dynamic nature of school change and the processes which facilitate or hinder schools' efforts towards improvement. Capacity refers to the 'power to engage in and sustain continuous learning of teachers and the school itself for the purpose of enhancing pupil learning' (Stoll, quoted in Stoll *et al.*, 2001: 171). For Harris, capacity building within schools 'essentially … implies that people take the opportunity to do things differently, to learn new skills and to generate more effective practice' (Harris, 2011: 627). Elmore (2008), working within a system leadership perspective (which is discussed in detail below), further expands on the concept of capacity and internal accountability in school improvement processes. He suggests that capacity and internal accountability are empirically and theoretically linked to improvements in student outcomes:

> The first, most obvious, finding from our research is that schools and school systems respond differently, depending on their capacity and their internal accountability. In simple terms…capacity is the fund of skill and knowledge that the organization can bring to bear in responding to external pressure, and internal accountability is the degree of coherence in the organization around norms, values, expectations, and processes for getting the work done. (Elmore, 2008: 44)

According to Elmore, accountability is about institutional response to, rather than compliance with, a specific measure or audit such as Ofsted. Schools already have accountability and instead of asking how schools should implement accountability measures, 'we should ask what their responses are to the panoply of incentives they face, what the determinants of these responses are, and how they adjust to alterations of these incentives over time' (2008: 43). For Elmore, successful school leaders should 'use the accountability system to position themselves and their organisations in a favourable place to gain resources and capacity' (2008: 44). Elmore found that schools that have high capacity and internal accountability, do well on external performance tests irrespective of whether their orientation is aligned with those tests (2008: 44). There is also some evidence that in national systems where data use is linked to external measures of accountability such as standardised tests, this can help build capacity by providing comparable benchmarks which identify failure (Sun *et al.*, 2007: 96). Whilst clearly important, capacity building and internal accountability are made possible through clear and strong leadership, a factor which we turn to next.

2.2.3 School collaboration and networking

There is an increasing emphasis on collaboration and networking by schools as reflected in a special issue of the journal *School Effectiveness and School Improvement* (21,1). Muijs and colleagues define networking as 'at least two organisations working together for a common purpose for at least some of the time' (2010: 6). They add that, 'the move towards networks can also be seen as part of a more general realignment of the relationship between central government and both local government and the market, characterised by increasing decentralisation, privatisation, and collaboration between government agencies, and between government and the private sector' (2010: 6-7). They identify the three most common goals that networks are established to support as being: school improvement; broadening opportunities (including networking with non-school agencies such as social services or business); and resource sharing.

According to Muijs and colleagues (2010), networking can impact on schools in a number of ways: by increasing capacity and creating opportunities for learning and knowledge creation rather than buy-in or externally imposed programmes; changing staff's perceptions of disadvantaged learners and reducing polarisation between schools involved in the networks; and by improving outcomes for vulnerable learners through collaborative networks with health and social care organisations.

A report by the National Audit Office (2009) examined the extent of 'partnering' activities and programmes directly funded by the then Department for Schools Children and Families (DCSF) and its impact in English secondary schools (in terms of student outcomes for 11-14 year olds and behaviour). They defined partnering as 'a school working with one or more other schools or organisations towards agreed objectives' (2009: 4). Their study produced a number of important findings about networking across schools. They found that:

- partnering to improve attainment was more common than partnering to improve behaviour and developed largely in response to locally defined needs rather than externally mandated alliances;
- better performing schools were less likely to involved in partnering, and tended to view partnering as something for schools with 'problems';
- issues of governance, objectives, and accountability were often unclear in partnering;
- the most effective partnership was often one which covered the transition of pupils between primary and secondary (a period where student progress is known to be vulnerable); and
- schools were unclear about the financial costs of partnering and could not quantify any savings made in virtue of being involved in a partnering relationship (2009: 4-8).

The report also used statistical regression to estimate the impact of partnering on schools. Better attainment outcomes were statistically associated with the period of time a school was in an attainment or behaviour partnership or sharing resources with other schools. The report acknowledged that 'it is difficult to demonstrate a direct, quantifiable impact of partnering on attainment and behaviour across schools nationally because other factors are likely to have substantial effects' (2009: 7). They noted that often the costs of partnering are 'concealed' in overall costs, particularly salaries, and recommended more local evaluation and analysis of costs.

Research by Chapman and colleagues (2010) on schools involved in formal federations found that improvement processes were strongest when 'ownership of the change is locally owned and resides with teachers … [and] addressed their own internally driven priorities in their own way' (2010: 69). They suggest that federations worked well where schools were coming together for their own purposes, rather than when they were compelled by external pressure to forge alliances and solve problems. Others observe that there is a need for 'key people who serve as boundary spanners and have considerable influence in the school' to leverage the power of networking to influence classroom practice (Katz and Earl, 2010: 39). Overall, the evidence indicates that the benefits of collaboration between schools results in the sharing of effective practice; enhanced opportunities for career development and support; wider curriculum choice and raised student expectations (and in some cases attainment); improved local strategic leadership capacity; facilitating access to additional funding and resources and economies of scale (Higham *et al.*, 2009: 5). The factors which contribute to successful collaboration are summarised from the work of West (2010) and the National Audit Office (2009) in Table C3. Both lists stress the importance of local authority support and involvement; credible external input, evaluation and support; and commitment to change from all staff.

Table C3: key dimensions of successful collaboration between schools as found in the literature

West (2010)	National Audit Office (2009)
A willingness and desire amongst local authority staff to support and engage with the collaborative process, exploring and developing new roles and relationships	Local authority support and, where there is a clear role to play, direct involvement
External help from credible consultants/advisers (from the local authority or elsewhere) who also have the disposition and confidence to learn alongside their school-based partners	Regular evaluation with independent input
The development of a sense of collective responsibility for improvement in all partner organisations	Recognition that all partner schools have something to contribute, and willingness to share success
The identification of common improvement priorities that are seen to be relevant to a wide range of stakeholders; Headteachers and other senior staff in schools who are willing and able to drive collaboration forward	Inclusive of all those who have the skills and knowledge to usefully contribute, whatever their role
Incentives to encourage stakeholders to participate and be interested	Trust, goodwill and commitment among members
	Good alignment with local context
	Simple governance with periodic review to assess whether the partnership is meeting its full potential and should continue
	Clear and consensual objectives

2.2.4 Other factors in school improvement

Other factors identified as being important within the research literature include data use and strategy, and 'instructional coherence'. Research by Campbell and Levin (2007), in Toronto, shows that data use strategy, embedded in assessment and learning practices can drive school improvement. The use and understanding of data can help schools to gauge the extent to which they are improving and becoming more effective. They argue that for data use to successfully drive school improvement, data rules have to be clear and comparable; systems have to be consistent and intelligible to each other. They cite research which shows that assessment for learning can be a powerful strategy to support teachers' pedagogy. This can inform curriculum and learning goals which can increase student performance (Campbell and Levin, 2007: 48). They add that 'effective use of assessment and data to support positive outcomes for educators and students requires careful attention to building capacity to access, understand and apply data' and this is not something that is always present within schools (2007: 49). A recent study (Kelly and Downey, 2011) of secondary school teachers' attitudes towards use of pupil performance data, suggests that data needs to be more democratic. At the moment, data is managed and perceived as the responsibility of specialist staff or as tool of senior management. The study found a statistical association between having high CVA scores and the high usage of data. The authors speculate that this:

> ... may be indicative of a developed culture in these schools around the use of data at both pupil and school level. School-level CVA score is dependent on the degree of progress made by every pupil in a cohort (whereas threshold measures simply count the proportion of pupils who manage to get ''over the bar''), and, unlike threshold measures, improving school-level CVA is not easily achieved by targeting borderline students.' (Kelly and Downey, 2011: 432).

Evidence from urban school reform in Chicago during the nineties (Newmann *et al.*, 2001: 299) suggests that student attainment improves more in schools where improvement efforts 'are guided by a common framework for curriculum, instruction, assessment, and learning climate and are pursued over a sustained period.' The authors refer to this as 'instructional coherence'. Other researchers have also noted the importance of tapping into student motivation. All too often school researchers have viewed all pupils as the same, and ignored the differential effectiveness of intervention for certain groups of students (Jamieson and Wikeley, 2000). Research has also identified the importance of tapping into student motivation and voice, parents' voices and participation in school improvement (McCall *et al.*, 2001; Rudduck *et al.*, 1996). In the next section, we consider one of the most influential approaches to school improvement, system leadership.

2.3 System leadership

One of the key trends in school improvement scholarship in the past decade has been the focus on what is termed 'system leadership'. It is an increasingly influential approach to school improvement and effectiveness across the globe and underlies a number of English government education policies such as Extended Schools, Every Child Matters (ECM), Leading Edge Schools (LES), Networked Learning Communities (NLCs) and City Challenge (Huber *et al.*, 2008; Pont *et al.*, 2008a). System leadership is an approach to school improvement which, according to O'Leary and Craig, 'has a consistent set of principles that can and should be applied within and between institutions' (2007: 3). These principles are based on what is known as 'systems theory'. A central insight of this theory is the notion of a system, composed of interdependent parts, which are necessary to its functioning. The power of the system resides at the level of interactions and connections between its parts.

By extension, schools are viewed as both systems in themselves and nested within larger interdependent social systems. Parts of the nested system include: the educational system (relationships between schools); systems of public services (relationships between schools and other public services); and local communities (relationships between schools in a local area, including parents and carers). The solutions to the problems schools face therefore reside in harnessing the power of those connective relationships and functional interdependence. According to O'Leary and Craig, a system leader is someone who recognises this fundamental interdependence and relational aspect of school leadership, sees and acts 'on the system as a whole' rather than on isolated parts (2007: 4). They suggest that system leaders can drive school improvement by drawing on the differing perspectives of individuals (within a system) to focus on outcomes which create a shared a vision; supporting the autonomy of those in the system by setting a few simple rules while maintaining high minimum standards; connecting individuals to one another to encourage individual autonomy and help them to create meaningful relationships which support problem solving and shared learning; supporting learning and continuous improvement, by creating feedback loops – this can be done via data gathering and sharing or the exercise of pupil/parental voice – to help those within the system to improve its overall performance; and acknowledging the importance of an open learning organisational culture (O'Leary and Craig, 2007: 5).

According to Hopkins (2008) system leaders have a number of characteristics: a moral purpose, which is expressed through measuring success in terms of improving student learning and achievement and striving to narrow achievement gaps; a commitment to teaching, learning and curriculum and personalisation; the development of schools as personal and learning communities, with relationships built across and beyond schools; and a striving for equity, inclusion and empowerment (2008: 22). Hopkins defines system leaders as: 'those head teachers who are willing to shoulder system leadership roles, who care about and work for the success of other schools as well as their own' (2008: 22). Hopkins adds, 'system transformation depends on excellent practice being developed, shared, demonstrated and adopted across and between schools' (2008: 30).

> Advocates argue that system leadership can distribute responsibilities in novel ways, enabling leaders to influence student outcomes in a school. Examples of innovative system leadership roles are: executive headship, consultant heads, leaders of extended schools, co-leaders, those responsible for network relationships in federations and consortia, or change agents such as National Leaders of Education (NLEs) and Local Leaders of Education (LLEs), School Improvement Officers and Partners (SIOs and SIPs). System leadership roles enable individuals to

> take responsibility for wider processes and interconnections between various systems in their own and other institutions.

Until recently, LAs were much seen as important partners in driving systems based school improvement policies, partly through the legal powers to act to compel what are deemed as underperforming schools to take necessary steps towards improvement. Some of these powers include compelling schools to work with an external partner organisation for the purpose of school improvement, and take actions such as appointing additional governors, replacing the entire governing body, and taking back the school's delegated budget. Research findings from the NFER (Keating *et al.*, 2009) indicated that LAs were reluctant to use their statutory powers of compulsion, and did so as a last resort, preferring voluntary collaborative arrangements where possible, to facilitate school improvement. Some LAs felt that warning notices worsen school-LA relations and provoked resistance from some schools. In the next section, we look in detail at the variety of system leader roles in practice.

In a recent development of the system leadership approach, Hargreaves has advanced the notion of a 'self-improving school system' (SISS). The idea is that system leaders should work together in tight localised clusters of schools to harness the 'cumulative logic' of shared capacity (Hargreaves, 2010). Hargreaves also suggests that leadership should be distributed at all levels within, and between, schools. In essence, schools should take on the responsibility of improvement themselves. He adds:

> In a self-improving school system, more control and responsibility passes to the local level in a spirit of mutual aid between school leaders and their colleagues, who are morally committed to imaginative and sustainable ways of achieving more ambitious and better outcomes. (Hargreaves, 2010: 23)

Hargreaves' concept of the SISS appears therefore to resonate very much with aspects of the City Challenge programme, such as the Families of Schools programme.

2.3.1 System leadership roles

In a study of a system leader role, Earley and Weindling (2006) looked at the role of 'consultant heads' as part of an evaluation of the leadership strategy strand of the London Challenge. They found that consultant heads had two key roles; that of consultancy (mentoring, coaching and facilitating) and brokerage (providing access to external expertise and training); and that this strategy seems to have had a positive reception among the heads and the receiving schools in the evaluation (2006: 51).

NLEs grew out of the leadership strategy strand of the London Challenge. According to Higham and colleagues, 'together with their schools [NLEs] epitomise system leadership in action' (2009: 106). They describe the NLE role as going beyond the mentoring and coaching role provided by consultant heads. They provide leadership and oversight of school improvement measures and strategies ensuring actions are implemented. NLEs increase and enhance leadership capacity and in so doing secure greater performance for schools. The authors report evidence showing that NLEs share 'several compelling and largely common characteristics' such as demonstrated a consistency in their approach, purpose and motivation; they were highly-driven, optimistic, resilient, and had high expectations (2009: 116). Higham and colleagues present evidence that 'NLE involvement is associated with improvement in the client school being supported and with continuing improvement in the NSSs (National Support Schools)' (2009: 124).

The recent report by the NFER on the leadership strand of City Challenge affirms the value of system leaders in providing, and brokering school-to-school support and facilitating improvement (Rudd *et al.*, 2011). The authors found that the NLE/LLE role was viewed as highly successful by schools, with schools valuing their breadth of expertise and commitment. The leadership strategies were seen as representing good value for money and useful building leadership capacity of schools. The NLE/LLE roles have also been cited as exemplars by Barber and colleagues in their McKinsey review of effective leadership development and support (Barber *et al.*, 2010: 21).

In addition, Ainscow (2010a) has looked at the role and characteristics of school improvement officers (as supporting arms of LAs). His analysis suggests that they often had deep knowledge of their specific schools and contexts, which enabled them to mobilise resources where it was needed and 'which made their interventions authentic' (2010a: 79). Ainscow found that LA staff had an important role in 'supporting and challenging schools in relation to the agreed goals of collaborative activities, whilst headteachers shared responsibility for the overall management of improvement efforts within their schools' (Ainscow, 2010a: 90). He adds that LA staff should not be leading or managing change but rather, working in partnership with senior leadership in schools to develop and strengthen capacity. He writes:

> In such contexts, they can ensure that specific challenges which derive from their knowledge of the bigger picture across the authority are addressed and also contribute to the clarity of purpose and practical working arrangements, as well as playing an important role in the monitoring and evaluation of progress. At the same time, they can help to broker the sharing of resources and expertise. However, the changes in attitude and practice that this implies are likely to be challenging to the existing thinking of many experienced local authority staff. (Ainscow, 2010a: 90)

Another aspect of external system leaders' role is often described as being that of a 'critical friend'. Critical friends are supposed to offer schools a supportive and challenging relationship and are part of the 'high challenge, high support' strategy used in the English approach to system leadership. Research on external change agents has explored this notion of 'critical friendship'. According to Swaffield:

> [A] critical friend is generally portrayed as a detached outsider who assists through questioning, reflecting back and providing another viewpoint. The critical friend prompts the other person towards honest reflection and re-appraisal, a seeing anew that may be challenging and uncomfortable, yet enhancing. The relationship is neither cosy nor collusive, but rather one that cultivates constructive critique. (2007: 205-6)

Swaffield examined critical friendship in the context of a specific 'light touch critical friendship' between universities and schools as part of the Learning How to Learn initiative (LHL). The LHL was an innovative collaboration between schools and universities to share knowledge and practice within networks established as part of the Economic and Social Research Council's Teaching and Learning Research Programme (TLRP). Swaffield found that critical friends supported school improvement through the roles they played as actors within the networks in which schools were embedded. These roles included: liaison, advocacy and engagement; presentation and training; supporting reflection; questioning; supporting development; spreading and critiquing practice; and feeding back data (Swaffield, 2007: 2010-15). He suggests that the factors which contribute to effective working relationships between 'critical friends' and schools are: role clarity; willing to engage; continuity, and ability to cope with changes in personnel; the ability to negotiate competing priorities and maintain momentum; trust, reputation and credibility; initial meeting and site visits by 'critical friends' (2007: 205-10).

2.3.2 System leadership in an international context

Recent studies conducted for the OECD on the theory and practice of system leadership in different countries highlight the global interest and influence of the approach and its perceived cross-cultural application (Pont *et al.*, 2008a; Pont *et al.*, 2008b). One of the reports identifies and discusses several case study regions which have made system leadership the core of their school improvement efforts, such as Flanders (Belgium), England, Finland, Victoria (Australia) and Austria (Pont *et al.*, 2008a).

The OECD reports demonstrate the diversity across countries and how national contexts influence system thinking and implementation. This makes generalisations about which strategies work best difficult. In Finland for example, system leadership is of an 'organic' kind that 'is embodied…in the way policy making and collaboration play out,' (Pont *et al.*, 2008a: 261). There is widespread consultation at all levels of the educational system and a sense of ownership over change. Teachers and school leaders have a high degree of professional autonomy and trust, 'teamed with a high degree of decentralisation and a national consensus on the value of learning'. The whole system cooperates for school improvement' (Pont *et al.*, 2008a: 261). This contrasts with the English approach of 'high challenge, high support'. High challenge because it has sought to prompt schools to take responsibility for improvement through, largely, external accountability measures which create pressures for schools to reflect, develop and change practices; and high support, because it has been felt that many schools may not be able to or have the capacity to meet their responsibilities unless they work with partners (Huber *et al.*, 2008: 115; Sun *et al.*, 2007). A recent international study of the key drivers of continuous school improvement in England, Scotland, Australia, Finland and Canada, suggested that national educational systems that have 'a cohesive alignment of policies and practices that describe, develop and evaluate teacher, school leader and school capacity' have the greatest impact on student learning and outcomes (McIntyre, 2011: 4).

In a summary of the lessons learned from the international cases studies of system leadership in different regions, Pont *et al.* point to a number of common themes in the strategies and focus of efforts (2008a: 263-265). These include across all regions, a focus on the development and strengthening of leadership capacity; seeking more efficient modes of resource allocation; increased co-operation with other schools and partners; use of distributed leadership at various levels of the educational system; and a clear emphasis on improving school outcomes. In the final section of our discussion of the literature we look at how schools can maintain improvement.

2.4 New types of school

So far we have discussed approaches to school improvement which focus on changing various aspects of a school or the system in which it is imbedded. However, more radical approaches to school improvement which attempt to change the whole system or implement structural solutions are increasingly popular. The free schools movement, which started in Sweden in the 1990s, charter schools in the US, and Academies in England, can all be seen as part of this trend. The underlying rationale for these approaches, is that introducing new types of schools (often replaced failing ones with new leadership and sometimes in conjunction with pupil vouchers), that are more autonomous, free of local authority or direct government control, will enable greater parental choice. This, it is claimed, will inject market competition into the education system and thus drive up standards/school performance for pupils not in free schools as well as those within. This is because these new types of schools will promote educational innovation and expanded opportunities for teacher-led learning and involvement in the governance of schools (McKinsey and Company, 2007). Despite the enthusiastic political support for these reforms, the evidence from the Swedish and US example remains mixed (Bunar, 2010). Böhlmark and Lindahl's (2007) econometric study found 'evidence that the competitive forces unleashed by the 1992 school reform in Sweden induced higher pupil achievement, but also higher costs and greater segregation' (2007: 42). In their second study Böhlmark and Lindahl's (2008) revealed that an increase in the number of free schools in a municipality moderately improved grade point averages at 15-16 yrs, but found no medium or long-term impact on educational outcomes such as high school exit test results, university attainment or years of schooling. They speculate that this might be because 'the entry of new private[1] [free] schools not has been followed by the closing down of public schools. Hence, it might be that increasing shares of school budgets have been devoted to maintain operating public schools of poor quality' (2008: 23). In view of the econometric evidence, Allen (2010), observed that free schools have tended to be most prevalent in 'urban, affluent and gentrifying areas and in those places with second-generation immigrant communities … so the

[1] Unlike their English counterparts, Swedish free schools are run as for-profit private enterprises.

overall system' has a small but noticeable stratifying effect (2010:4). Allen also reports evidence which indicates that the biggest beneficiaries of the reform 'are children from highly educated families … [and that] the impact on low educated families and immigrants is close to zero' (2010:5). In summary Allen (2010:8) notes that the: 'evidence on the impact of the reforms suggests that, so far, Swedish pupils do not appear to be harmed by the competition from private schools, but the new schools have not yet transformed educational attainment in Sweden'.

In relation to charter schools in the US, the evidence is similarly mixed. One study of charter schools in Florida showed a modest raise in Maths attainment for pupils in neighbouring non-charter schools, but no significant differences in overall performance between charter and publicly funded schools (Sass, 2006). Other studies of charter schools in North Carolina echo these findings (Bifulco, 2006). A comprehensive and robust econometric study was conducted by the Center for Research on Education Outcomes (CREDO, 2009) using longitudinal student-level analysis of charter school impacts in 16 States. The researchers drew on a sample of more than 70 per cent of all the students in charter schools in the US. Their analysis found significantly wide variation in performance across States. Overall, they concluded that:

> ...a decent fraction of charter schools, 17 percent, provide superior education opportunities for their students. Nearly half of the charter schools nationwide have results that are no different from the local public school options and over a third, 37 percent, deliver learning results that are significantly worse than their student would have realized had they remained in traditional public schools. (2009: 1)

Within this mixed picture, the authors observed that students in poverty and those who are classified as being English Language Learners (ELL) do 'experience larger learning gains [in Reading and Maths] in charter schools' (2009:44). This is interesting given the evidence from Sweden which seems to suggest that it is the more advantaged students who benefit most. Further research conducted by RAND (Zimmer et al., 2009) examined the performance of charter schools in eight States, with a focus on the long-term impact on educational outcomes and found on the positive side that: charter school students had a higher probability of graduating and attending college than public school students; they do not increase racial stratification and skim the higher-achieving students from local state schools, nor do they harm overall student achievement in those schools. However charter schools only produce achievement gains in middle and high schools that are equivalent to those in traditional public schools (Zimmer et al., 2009).

2.5 Sustainability

We conclude this review of the literature by considering issues of sustainability. Once schools begin to improve, the issue of how to sustain improvement is raised. Less is known about how schools sustain improvement. Whilst schools often make initially impressive gains in terms of student attainment it is acknowledged that 'school improvement is a slow process because it is about maturation' (Macbeath and Mortimore, 2001: 17). Research discussed by Gray (2001) indicates that few schools manage to secure year on year improvement in results, especially past the third year, and there is a plateau, which indicates that it is much more difficult for schools to change their effectiveness. In fact, he argues that 'to succeed in sustaining improvement is the exception rather than the rule' (Gray, 2001: 26). This insight that improvement is not just movement in a linear direction, and that it has an effectiveness dimension has led to a rapprochement between the two research agendas (Stoll, 1996). Researchers now think of the effectiveness dimension of improvement. Therefore, an improving school is not simply one in which results are consistently increasing (in terms of measured student performance), but they are doing so with whatever constitutes successively 'similar cohorts' of pupils; in other words, they are becoming more effective.

Elmore (2008) points out that high quality and high performing schools are not the end state of school improvement but rather a phase in school development. School improvement, according to

Elmore, is a developmental process. He suggests that what is happening as schools are improving is that changes are occurring in the structure, processes and normative dimensions around which the work of staff and students are organised (2008: 49). Thus school transformation implies a school is getting better at its core functions; changing the way learners and staff think about their role in the process of teaching and learning; and increasing internal accountability by managing the organisation in progressively more coherent ways. Elmore suggests that quality (which might be recast as effectiveness) and performance are not the same thing, and that 'many high-performing schools are stunningly mediocre in their practice, and produce most of what they do with social capital' (Elmore, 2008: 54).

Gray and colleagues distinguish three levels at which improvement efforts can be approached; the tactical, strategic and capacity-building. In their research, most schools operated at the tactical level, where improvement involved focusing teachers' energies and effort into improving student performance (possibly for the first time), and on more 'obvious' interventions that could provide immediate results. Schools at the strategic level were in a minority, and delved more into student learning, focussed more systematically on their weaknesses, reviewed approaches across the school to raising achievement and were more aware of the limitations of tactical levers, and had begun to forge links between classroom practice and pupils' learning. Few schools had engaged in capacity-building. As we indicated earlier, capacity-building is an important factor in school transformation. In Gray's study it was a mature stage of improvement in which schools demonstrated a sophisticated grasp of the problems they faced, knew how to undertake change, and showed 'a willingness to go beyond merely incremental approaches' (Gray in Gray, 2001: 31).

In recent publications documenting a decade of school improvement efforts in a specific geographic region, Ainscow and colleagues identify a number of factors which provide the basis for sustained school improvement (Ainscow, 2010a, 2010b; Muijs *et al.*, 2010). These include:

- realising untapped potential, often implicit, which schools can use to drive change and improvement;
- the use of evidence as a catalyst in the identification of issues that need attention and the mobilisation of resources to support improvement efforts in relation to these issues;
- school to school collaboration as a powerful means of facilitating improvement particular for schools in challenging contexts;
- cross-border collaboration;
- system leadership;
- local authority roles, brokering collaboration, support for leadership development, identifying priorities etc.

Ainscow and colleagues' list can be seen therefore to reinforce the messages from the overall evidence base reviewed, particularly those from the system leadership approach, which emphasise the importance of schools developing collaborative relationships with external partners and the development and utilisation of system leadership roles. In the next section, we review government policies that have aimed at enhancing students outcomes through school improvement.

3 Policy

Over the years, the successive governments have aimed to raise standards by influencing what goes on in schools. Every aspect of the work of schools has been targeted: school leaders (e.g. through the creation of the National College for School Leadership and the establishment of the National Professional Qualification for Headship), teachers (e.g. through the introduction of Professional Standards for Teachers), support staff (e.g. workforce remodelling, establishment of the HLTA role), curriculum (e.g. the National Curriculum), teaching methods (e.g. National Strategies), use of data, and so on. Some of these changes were driven by concerns that internationally, English children were lagging behind their peers, and that there were lessons to be learnt from high-performing education systems such as those in Finland, Singapore, Alberta etc. (McKinsey and Company, 2007).

Other broad approaches to school improvement have included Ofsted inspections, league tables and floor targets. Changing the inspection framework or the floor targets tends to change what goes on in schools, as each school aims to maximise its own performance.

As well as strategies to improve all schools, governments have focused a number of initiatives and strategies on improving schools in areas of poverty and deprivation, usually in inner-urban areas. This group includes Excellence in Cities, Education Action Zones, the Extra Mile Project, London Challenge, and subsequently City Challenge. While many initiatives have been clearly focused in schools, others have been designed to address educational disadvantage as part of a wider programme tackling other aspects of disadvantage (e.g. New Deal for Communities).

Governments have also tried to bring about improvement by closing failing schools. Fresh start schools were schools that were underachieving or in Ofsted categories and were closed and then reopened on the same site. In 2000, the Academies programme was announced introducing a new type of school to replace seriously failing schools. Academies were to be built and managed by partnerships between government and other sponsors, and would have the freedom to introduce innovative forms of management, governance, teaching and curriculum (Blunkett, 2000). By April 2011, when City Challenge ended, 271 sponsored academies had opened.

The 2010 White Paper, *The Importance of Teaching*, signals a rather different approach, aiming to create a school system which is self-improving, and schools are responsible for their own improvement:

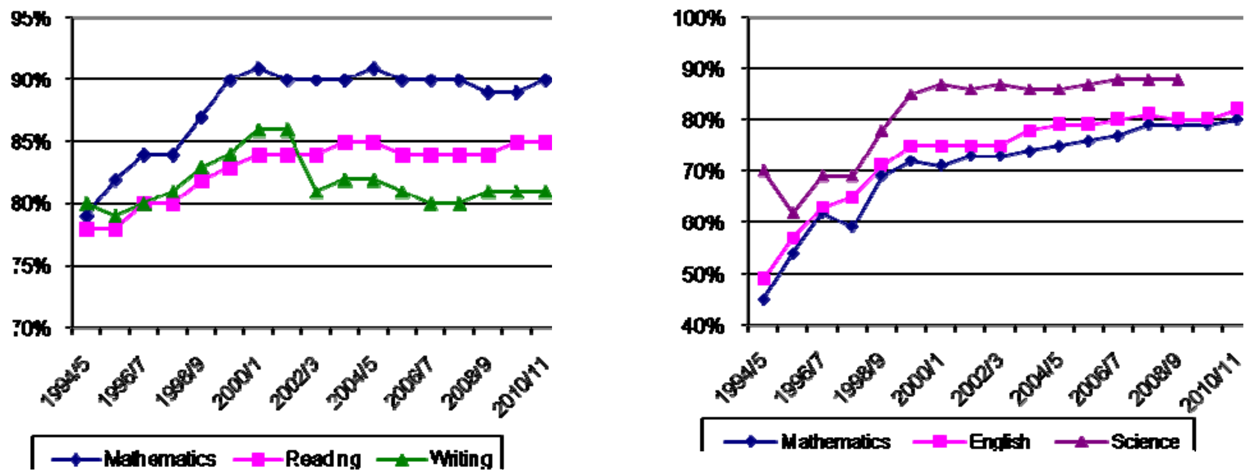
> We will expect schools to set their own improvement priorities. As long as schools provide a good education, we will not mandate specific approaches. Schools will determine what targets to set for themselves, choose what forms of external support they want and determine how to evaluate themselves. (2010: para 7.6)

3.1 School standards

In this section we consider overall trends in attainment. Figures C1 and C2 show changes in attainment from 1996/6 to 2010/11.

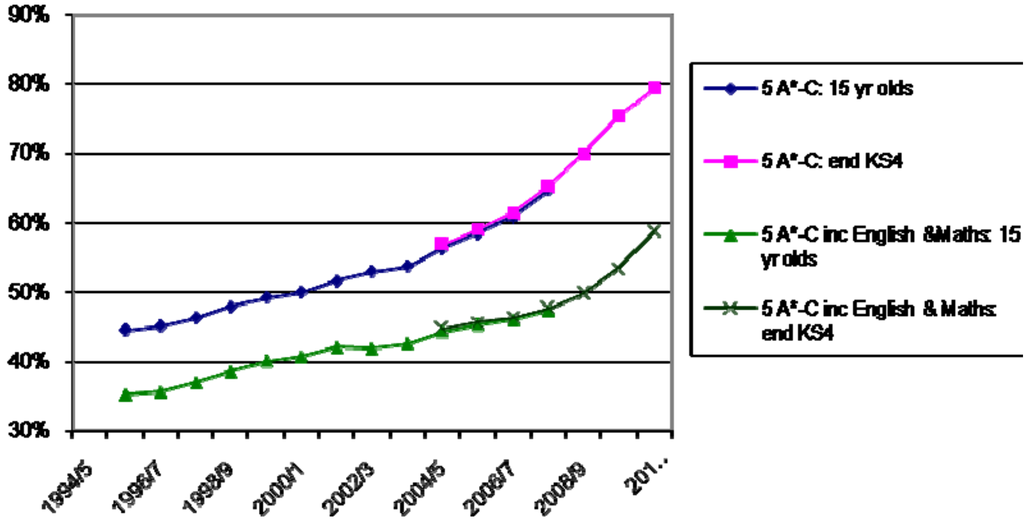
Figure 1: Percentage of primary pupils achieving expected level or above in the national key stage tests, all Schools, England, 1995/6 to 2010/11

Key Stage 1 tests, percentage achieving Level 2 *Key Stage 2 tests, percentage achieving Level 4*



Note: Since 2005, KS1 results have been based on teacher assessment. Key Stage 2 science tests were replaced by teacher assessment from 2009/10.
Source: DfES, 2006; DfE, 2011a, 2011b

Figure C2: Percentage of secondary pupils achieving expected level, maintained schools, England, 1995/6 to 2009/10



Note: Over this period the expected level changed from 5A*-C to 5A*-C including English and mathematics. Hence both are shown on this graph. The group whose attainment is recorded changed from 15-year olds to pupils at the end of Key Stage 4.
Source: DfE, 2012

It is quite clear that, over the last fifteen years, secondary school standards, as measured by GCSE results, have improved. However, in primary schools, after a rapid increase in the late nineties, improvement slowed down or stagnated. A particular effort has been put into reducing the number of secondary schools with low attainment; while in 2005, over 900 secondary schools had fewer than 30% of pupils achieved 5+ A*-C including English and Mathematics. In 2011, this was the case for only 45 schools.

3.2 The impact of specific initiatives

A number of evaluations have attempted to identify the impact of specific innovations. However, in some cases, these show relatively limited impact. For example, the evaluation of Excellence in Cities (EiC) (Kendall *et al*., 2005) reports that that Key Stage 3 pupils attending EiC schools had higher levels of attainment in mathematics than otherwise similar pupils attending schools that were not part of EiC, but that there was no substantial evidence to show that EiC had an impact on levels of attainment in English or science at the end of KS3. Similarly, there was little evidence to suggest that pupils in EiC areas were making more progress during Key Stage 4 than similar pupils in non-EiC areas.

The analysis of the impact of the New Deal for Communities (NDC) on educational attainment aimed to assess whether educational attainment had improved over and above any improvements that might have been expected in the absence of the NDC Programme, and did this by making comparisons with the attainment of similar children living in similarly deprived areas. It concluded that there was little evidence of a programme-wide improvement in attainment outcomes (Wilkinson and McLennan, 2010).

The London Challenge has been reviewed very positively. In 2006, Ofsted reported that standards at KS3 and KS4 are rising faster in London than nationally, and goes on to say ‘The investment in London Challenge … has helped schools and local authorities to improve’. It also concludes that ‘financial support and high quality advice, together with good decisions at ministerial level, have helped the performance of the five target LAs facing the greatest difficulties’. Similarly, the 2010 Ofsted report on the London Challenge states that both primary and secondary attainment had risen faster than in schools nationally, and attributed this to the London Challenge. Factors leading to this success were identified as ‘successful collaboration between London school leaders and teachers across schools' (2010, p. 6), and the strategic deployment of support to schools involving accurate audit of needs by Challenge advisors, and deployment of resources to meet those needs.

The development of leadership roles such as NLEs as part of City Challenge, has according to Higham and colleagues (2009), impacted positively on school performance. They suggest that the deployment of NLEs has been pivotal 'in pulling a growing number of schools out of Ofsted categories… [and are the] most likely agents to raise the standards of National Challenge Schools above floor targets' (2009: 126).

Evaluations of the Academies programme by PriceWaterhouseCoopers (2008) and the National Audit Office (2010) both report that on average, attainment and pupil attendance in academies had improved faster than in maintained schools with similar intakes. A recent robust analysis of outcomes in academy schools with a comparator control group of state-maintained schools that go on to become academies indicates that academy conversion generates an improvement in the quality of the pupil intake and a significant improvement in pupil performance, and that these effects are strongest for the schools that have been academies for longest (Machin and Vernoit, 2011). However, all these reports highlight that improvement has not been uniform across all academies. Moreover, it is impossible to disentangle which aspects of academies have led to the improvement. The start-up funding, and in many cases new buildings which are more appropriate for modern teaching, may have impacted on attainment. Outstanding leaders may have been attracted by the higher pay offered. Thus it is not clear to what extent the improved attainment relates to innovative forms of management, governance, teaching and curriculum.

3.2.1 The difficulty of establishing causality

The Wilkinson and McLennan (2010) report makes some useful observations about the evaluation of area-based initiatives. The authors state that 'although rigorous statistical analysis gives reliable and accurate information about change, it does not inform us about causality' (p.5). They stress the need to have sufficient information about interventions to be able to attribute changed outcomes to the interventions. This was a factor in the evaluation of the Black Children's Achievement Programme conducted by IPSE (Maylor *et al*., 2009); it was clear that schools did not always target all Black children, and that some schools added children who were not Black to the target group. Therefore any statistical analysis of attainment was of limited value since it was unclear which children had been targeted. Wilkinson and McLennan (2010) also note the importance of understanding what a specific initiative such as NDC is adding to mainstream policies to bring about school improvement, and what the overlaps are. The more complex the initiative, the more difficult it becomes to disentangle what aspects of it have brought about change.

References

Ainscow, M. (2010a), 'Achieving excellence and equity: Reflections on the development of practices in one local district over 10 years'. *School Effectiveness and School Improvement,* 2` (1)*,* 75-92.

Ainscow, M. (2010b), Moving knowledge around: Some lessons about how to improve schools from the Greater Manchester Challenge.

Ainscow, M., West, M. and Stanford, J. (2005), 'Sustaining improvement in schools in challenging circumstances: A study of successful practice'. *School Leadership & Management,* 25 (1)*,* 77-93.

Allen, R. (2010), *Replicating Swedish 'free school' reforms in England*. Bristol: Bulletin of the Centre for Market and Public Organisation, Issue 10, Summer 2010, p.4-7.

Barber, M., Whelan, F. and Clark, M. (2010), *The leadership premium: How the world's top school systems are building leadership capacity for the future*: McKinsey & Company.,.

Barker, B. (2007), 'The leadership paradox: Can school leaders transform student outcomes?'. *School Effectiveness and School Improvement,* 18 (1)*,* 21-43.

Bifulco, R. (2006), 'The impacts of charter schools on student achievement: Evidence from north carolina'. *Education Finance and Policy Winter,* 1 (1), 50-90.

Blunkett, D. (2000), Transforming Secondary education. Speech to the Social Market Foundation, March 15th, (available at: http://www.dfes.gov.uk/speeches/media/documents/transforming_secondary_education.doc)

Böhlmark, A. and Lindahl, M. (2007), '*The Impact of School Choice on Pupil Achievement, Segregation and Costs: Swedish Evidence'*, IZA Discussion Paper No. 2786.

Böhlmark, A. and Lindahl, M. (2008), '*Does School Privatization Improve Educational Achievement? Evidence from Sweden's Voucher Reform'*, IZA Discussion Paper No. 3691.

Bracey, G. (2004), The trouble with research, part 2, *Phi Delta Kappen*, 85, 8, 635-636.

Brighouse, T. and Woods D. (1999), *How to improve your school*, Routledge.

Bunar, N. (2010), 'Choosing for quality or inequality: Current perspectives on the implementation of school choice policy in sweden'. *Journal of Education Policy,* 25 (1)*,* 1-18.

Campbell, C. and Levin, B. (2007), 'Using data to support to educational improvement'. *Educational Assessment, Evaluation and Accountability,* 21*,* 47-65.

Chapman, C., Lindsay, G., Muijs, D., Harris, B., Arweck, E. and Goodall, J. (2010), 'Governance, leadership, and management in federation of schools'. *School Effectiveness and School Improvement*, 21 (1), 53-74.

Corbett, D, Wilson, B. and Williams, B. (2002), *Effort and excellence in urban classrooms*. New York: Teachers College Press.

CREDO (2009), Multiple Choice: Charter school performance in 16 states. Center for Research on Education Outcomes (CREDO), Stanford University. Stanford: California, (available at: http://credo.stanford.edu).

Creemers, B., P and Reezight, G., J. (2005), 'Linking school effectiveness and school improvement: The background and outline of the project'. *School Effectiveness and School Improvement,* 16 (4)*,* 359-371.

Day, C., Sammons, P., Harris, A., Leithwood, K., Gu, Q., Brown, E., Ahtaridou, E. and Kington. (2009), *The impact of school leadership on pupil outcomes: Final report*. Nottingham: DfES.

DfE (2011a), National Curriculum Assessments at Key Stage 1 in England, 2011 http://www.education.gov.uk/rsgateway/DB/SFR/s001022/index.shtml.

DfE (2011b), National Curriculum Assessments at Key Stage 2 in England 2010/2011 (revised) http://www.education.gov.uk/rsgateway/DB/SFR/s001047/index.shtml.

DfE (2012), GCSE and Equivalent Results in England, 2010/11 (Revised) http://www.education.gov.uk/rsgateway/DB/SFR/s001056/index.shtml.

DfES (2006), National Curriculum Key Stage 1. Percentage of pupils achieving level 2 or above in England 1995 to 2007, http://www.education.gov.uk/rsgateway/DB/TIM/m002011/index.shtml.

DfE (2010), *The Importance of Teaching. The Schools White Paper 2010. Cm7980*. Nottingham: DfE.

Dyson, L. (2004), *The effect of neighbourhood poverty and low income on the school context*. Paper presented to the Canadian Society for the Study of Education.

Earl, L., Watson, N., Levin, B., Leithwood, K., Fullan, M., Torrance, N., Jantzi, D., Blair, M. and Volante, L., (2003), *Watching and learning 3*: Final report of the OISE/UT evaluation of the implementation of England's National Literacy and Numeracy Strategies. OISE/UT.

Earley, P. and Weindling, D. (2006), 'Consultant leadership - a new role for head teachers?'. *School Leadership & Management,* 26 (1)*,* 37-53.

Elmore, R., F. (2008), 'Leadership as a the practice of improvement. In B. Pont, D. Nusche and D. Hopkins (eds), *Case studies on system leadership* (Vol. 2, pp. 37-65). Paris: OECD.

Gorard, S. (2010), 'Serious doubts about effectiveness'. *British Educational Research Journal,* 36 (5)*,* 745.

Gray, J. (2001), 'Introduction: Building for improvement and sustaining change in schools serving disadvantaged communities. In M. Maden (ed.), *Success against the odds - five years on: Revisiting effective schools in disadvantaged areas* (pp. 1-39). London: RoutledgeFalmer.

Hargreaves, D., H. (2010), *Creating a self-improving school system*. Nottingham: National College of Leadership of School and Children's Services.

Harris, A. (2002), *School Improvement: What's in it for Schools?* London: Routledge Falmer.

Harris, A. (2011), 'System improvement through collective capacity building'. *Journal of Educational Administration,* 49 (6)*,* 624 - 636.

Harris, A., Chapman, C., Muijs, D., Russ, J., and Stoll, L. (2006), Improving schools in challenging contexts: exploring the possible. *School Effectiveness and School Improvement, 17*(4), 409-424.

Higham, R., Hopkins, D. and Mathews, P. (2009), *System leadership in practice*. Maidenhead: Open University Press.

Hopkins, D. (2007), *Every school a great school: Realizing the potential of system leadership*. Maidenhead: Open University Press - McGraw Hill Education.

Hopkins, D. (2008), 'Realising the potential of system leadership In B. Pont, D. Nusche and D. Hopkins (eds), *Case studies on system leadership* (Vol. 2, pp. 21-36). Paris: OECD.

Huber, S., Moorman, H. and Pont, B. (2008), 'The english approach to system leadership. In B. Pont, D. Nusche and D. Hopkins (eds), *Case studies on system leadership* (Vol. 2, pp. 111-152). Paris: OECD.

Jamieson, I. and Wikeley, F. (2000), 'Is consistency a necessary characteristic for effective schools?'. *School Effectiveness and School Improvement,* 11 (4)*,* 435-452.

Katz, S. and Earl, L. (2010), 'Learning about networked learning communities'. *School Effectiveness and School Improvement,* 21 (1)*,* 27-51.

Keating, A., Marshall, H. and Rudd, P. (2009), *Local authorities and school improvement: The use of statutory powers*. Slough: National Foundation for Educational Research.

Kelly, A. and Downey, C. (2011), 'Professional attitudes to the use of pupil performance data in english secondary schools'. *School Effectiveness and School Improvement,* 22 (4)*,* 415-437.

Kendall, L., O'Donnell, L., Golden, S., Ridley, K., Machin, S., Rutt, S., McNally, S., Schagen, I., Meghir, C., Stoney, S., Morris, M., West, A. and Noden, P. (2005) *Excellence in Cities: The National Evaluation of a policy to raise standards in urban schools 2000-2003*. Nottingham: DfES.

Leithwood, K. and Jantzi, D. (2006), 'Transformational school leadership for large-scale reform: Effects on students, teachers, and their classroom practices'. *School Effectiveness and School Improvement,* 17 (2)*,* 201-227.

Levin, B. (2006), 'Schools in challenging circumstances: A reflection on what we know and what we need to know'. *School Effectiveness and School Improvement,* 17 (4)*,* 399-407.

Macbeath, J. and Mortimore, P. (2001), 'School effectiveness and improvement: The story so far. In J. Macbeath and P. Mortimore (eds), *Improving school effectiveness* (pp. 1-21). Buckingham: Open University Press.

Machin, S., & Vernoit, J. (2011), *Changing School Autonomy: Academy Schools and their Introduction to England's Education*. London: LSE, Centre for the Economics of Education.

Maylor, U., Smart, S., Kuyok, K. A. and Ross, A. (2009), *Black Children's Achievement Programme Evaluation*. Nottingham: DCSF.

McCall, J., Stoll, L., Thomas, S., Sammons, P., Smees, R., Macbeath, J., Boyd, B. and MacGilchrist, B. (2001), 'Views of pupils, parents and teachesr: Vital indicators of effectiveness and for improvement. In J. Macbeath and P. Mortimore (eds), *Improving school effectiveness* (pp. 74-101). Buckingham: Open University Press.

McIntyre, A. (2011), *Continuous school improvement - what matters most*. Canberra: The Winston Churchill Memorial Trust.

McKinsey and Company (2007), *How the world's most improved school systems keep getting better*. McKinsey and Company.

Muijs, D., Chapman, C., Stoll, L. and Russ, J. (2004), 'Improving schools in socioeconomically disadvantaged areas - a review of research evidence'. *School Effectiveness and School Improvement,* 15 (2)*,* 149-175.

Muijs, D., West, M. and Ainscow, M. (2010), 'Why network? Theoretical perspectives on networking'. *School Effectiveness and School Improvement,* 21 (1)*,* 5-26.

National Audit Office. (2009), *Partnering for school improvement*. London: HMSO.

National Audit Office (2010), *The Academies Programme*. London: HMSO.

Newmann, F., M, Smith, B., Allensworth, E. and Bryk, A., S. (2001), 'Instructional program coherence: What it is and why it should guide school improvement'. *Educational Evaluation and Policy Analysis,* 23 (4)*,* 297-321.

Ofsted (2006), *Improvements in London schools 2000-06*. Ofsted Publications (available at: http://www.ofsted.gov.uk/Ofsted-home/Publications-and-research/Browse-all-by/Documents-by-type/Thematic-reports/Improvements-in-London-schools-2000-06.

Ofsted (2010), *London Challenge*. Ofsted Publications. Reference 100192 (available at: http://www.ofsted.gov.uk/resources/london-challenge.

O'Leary, D. and Craig, J. (2007), *System leadership: Lessons from the literature*. Coventry: National College for School Leadership.

Pont, B., Nusche, D. and Hopkins, D. (eds) (2008a), *Case studies on system leadership*. (Vol. 2): OECD.

Pont, B., Nusche, D. and Moorman, H. (2008b), *Improving school leadership: Policy and practice*. (Vol. 1): OECD.

PriceWaterhouseCoopers (2008) *Academies Evaluation Fifth Annual Report*, Nottingham: DCSF

Reynolds, D., Harris, A., Clarke, P., Harris, B. and James, S. (2006), 'Challenging the challenged: Developing an improvement programme for schools facing exceptionally challenging circumstances'. *School Effectiveness and School Improvement,* 17 (4)*,* 425-439.

Reynolds, D. and Stoll, L. (1993), ‘Linking school effectiveness knowledge and school improvement practice: Towards a synergy’. *School Effectiveness and School Improvement,* 4 (1), 37-58.

Rothstein, R. (2002), *Out of balance: our understanding of how schools affect society and how society affects schools*, Chicago: Spencer Foundation.

Rudd, P., Poet, H., Featherstone, G., Lamont, E., Durbin, B., Bergeron, C., Bramley, G., Kettlewell, K. and Hart, R. (2011), *Evaluation of city challenge leadership strategies: Overview report*. Slough: NFER.

Rudduck, J., Chaplain, R. and Wallace, G. (eds) (1996), *School improvement: What can pupils tell us?* London: David Fulton Publishers.

Sammons, P. (2008), *School effectiveness and equity: Making connexions. A review of school effectiveness and improvement research its implications for practitioners and policy makers*. Reading: CfBT Education Trust.

Sammons, P. (2009), ‘The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools’. *School Effectiveness and School Improvement,* 20 (1), 123-129.

Sass, T., R. (2006), ‘Charter schools and student achievement in florida’. *Education Finance and Policy Winter,* 1 (1), 91-122.

Sebring, P., B, Allensworth, E., Bryk, A., S, Easton, J., Q and Luppescu. (2006), *The essential supports for school improvement*. Chicago: Consortium on Chicago School Research at the University of Chicago.

Stoll, L. (1996), ‘Linking school effectiveness and school improvement: Issues and possibilities. In J. Gray, D. Reynolds, Fitz-Gibbon and D. Jessson (eds), *Merging traditions: The future of research on school effectiveness and school improvement* (pp. 51-73). London: Cassell.

Stoll, L., Macbeath, J., Smith, I. and Robertson, P. (2001), ‘The change equation: Capacity for improvement. In J. Macbeath and P. Mortimore (eds), *Improving school effectiveness* (pp. 169-190). Buckingham: Open University Press.

Sun, H., Creemers, B., P and De Jong, R. (2007), ‘'Contextual factors and effective school improvement', school effectiveness and school improvement’. *School Effectiveness and School Improvement,* 18 (1), 93-122.

Swaffield, S. (2007), ‘Light touch critical friendship’. *Improving Schools,* 10 (3), 205-219.

West, M. (2010), ‘School-to-school cooperation as a strategy for improving student outcomes in challenging contexts’. *School Effectiveness and School Improvement,* 21 (1), 93-112.

Wilkinson, K. and McLennan, D. (2010), Narrowing the gap? Analysing the impact of the New Deal for Communities Programme on educational attainment. London: Communities and Local Government.

Ylimaki, R., M, Jacobson, S., L and Drysdale, L. (2007), ‘Making a difference in challenging, high-poverty schools: Successful principals in the USA, england, and australia’. *School Effectiveness and School Improvement,* 361-381.

Zimmer, Ron, Gill, B, Booker, K, Lavertu, S, Sass, T, R, Witte, J. 2009). Charter Schools in Eight States: Effects on Achievement, Attainment, Integration, and Competition. Santa Monica, CA: RAND Corporation. (available : http://www.rand.org/pubs/monographs/MG869. Also available in print form.

APPENDIX C Assessing the Impact of KTS interventions

1 Summary

This report summarises work carried out by Evaluation and Research Practice, Coffey International, on behalf of London Metropolitan University. It develops and applies an approach to identifying the potential impacts of Keys to Success (KTS) interventions (including Pathways to Achievement in the Black Country) on Secondary Schools during the period 2003-2011, and for Primary Schools during the period 2008-2011. As outcomes of interest for secondary schools we use the percentage of pupils attaining 5 or more GCSEs at Key Stage 4, and for Primary schools, the average percentage of pupils (across Maths and English) attaining Level 4 in at Key stage 2.

After appraisal of other potential approaches and testing their appropriateness to the data we model the within-year treatment effects for these outcomes, using observations of scores across all relevant years. The treatment effects are estimated relative to the results a school would be predicted to have given their preceding year's results. This accounts for differential levels of regression to the mean across the distribution and also allows for effects of other factors affecting performance such as local deprivation. We test for robustness in a number of areas and compare regions.

Applying this approach, we find evidence for a positive impact for KTS intervention in both primary and secondary contexts.

2 Methodology

2.1 Data

Since KTS is a school level intervention, one approach to evaluating impact is to model effects at the school level. Reasonably complete data exists for English Secondary schools on performance on the KS4 indicator of percentage of KS4 pupils attaining 5 or more GCSES at A-C, and for Primary Schools on the KS2 indicator of percentage of KS2 pupils attaining Level 4 at English and Maths. Pupil-level data is available, and where this can be linked with data on resource based interventions which respond to pupil characteristics it may be advantageous to model outcomes at the level of pupils (Machin et al 2004). However, the focus of KTS is on schools, as part of a wider area based initiative, rather than on a particular type of pupil. Since schools may vary considerably in terms of their 'average pupil' we concentrate on analysis of school level effects. Pupils in many KTS target areas also tend to be mobile and a proportion may move between treated and untreated school which would complicate analysis. Analysing at the level of schools allows us to use time series information by making use of annually published results across the period.

2.2 Selection of Approach

The most basic way to attempt to identify treatment effects for schools would be to compare scores for treated and untreated establishments after the treatment has been underway for some time. However, since the KTS intervention was not randomly assigned but targeted purposefully to schools at least partly on the basis of results it follows that treated schools can be expected to start from a lower base than an average untreated school. Hence the effect of treatment could not be inferred from a simple post hoc comparison.

Difference-In-Difference

One way to address this issue is to compare the amount of 'change' in scores between treated and untreated schools over a relevant time period. The appropriate time period would be between pre-treatment and post-treatment for treated schools and an equivalent period of time for untreated schools (ideally the same historical time period). This can be captured in a regression framework with a specification that interacts treatment group membership with time period applied to data containing a pre and post measurement for treated and untreated schools (Card and Krueger, 1994).

$$\mathbf{Y = \boldsymbol{B_0}X_0 + \boldsymbol{B_1}X_1 + \boldsymbol{B_2}X_2 + \boldsymbol{B_3}X_1{*}X_2 + \boldsymbol{\varepsilon}}$$

where
X_1 = 1 for treated schools
X_2 = 1 for all schools post-treatment
$X_1{*}X_2$ = 1 for treated schools post-treatment

This specification controls for prior differences between the two groups, captured by B_1, while the average change between the two time periods is expressed by B_2. The coefficient for the interaction term B_3 captures the differential change that the treatment group made above and beyond the untreated. If B_3 is positive then the treated schools made greater progress than untreated. Comparing the amount of *change* over this period 'removes' the effect of prior differences. This is known as a difference-in-difference approach (Wooldridge, 2010).

Restrictions

A difference-in-difference approach is not suitable if data indicates differences in growth *rates* prior to treatment, or where there is a theoretical reason to assume such difference. Differences in prior growth rates imply differences in the expected rate of change in the absence of treatment. In other words they mean that the untreated group is not an appropriate counterfactual for treatment if we are considering only the 'difference-in-difference'. A diagnostic test for the appropriateness of using a difference-in-difference approach is therefore to examine the prior rate of change for the two groups.

In terms of empirical evidence that the pre-treatment trajectories of KTS schools are not the same as those for untreated schools one possible source is a comparison of scores over time. Figure 1 shows that the average pre-treatment scores for KTS schools and for untreated schools are not equivalent either in terms of level or rate of change over time. This strongly implies that there are systematic differences in the two groups in terms of prior trajectories, and if the two groups were consistent in terms of membership over time this would be sufficient to rule out difference-in-difference. However, although the untreated group is consistent the group of pre-treatment schools is changing across the period as schools enter treatment. Hence the differences in the average trends may be due to changes in the composition of the group.

A more appropriate measure of change is the year-in-year change for *individual schools.* The average year-in-year change for a pre-treatment KTS school in the percentage of pupils achieving 5+ GCSEs is 2.51%. This compares with an average annual year-in-year change of 1% for untreated (non KTS) schools. KTS schools have more positive annual change, even aside from compositional change in the group, and (as Figure 1 shows) they start from a lower base. These two facts confirm that untreated (non-KTS schools) as a whole are not a suitable contrast group for analysis based on a difference-in-difference approach.

There are some a priori reasons to anticipate systematic differences in how much change we can expect between KTS and other schools. If we divide the year-to-year change in school scores into systematic and random fluctuation each of these may be linked to the schools current location in the distribution. If a school is performing badly then it is plausible that its systematic long-term growth may be slower than other schools. However, it is also plausible that schools which do badly in a given year through random fluctuation (chance fluctuations in pupil cohorts, teaching inputs, or exam success) may have a relatively positive expectation of year-in-year change for the following year as their fortunes recover. These two effects may operate in different directions if they do occur. However, there is no reason to assume that they will tend to cancel one another out exactly.

The importance of these considerations for establishing a treatment effect is that if we do not select counterfactuals from the same part of the results distribution, we will not have the same *expectation* of change, even if other characteristics are held equal, or controlled for by a ‘differencing’ on prior scores. This could be a particular problem for evaluation where treatment is concentrated in the lower part of the distribution where expected year-on-year change due to random fluctuation is likely to have a positive average, mimicking the growth we are hoping to identify from the intervention.

Matching

Matching is an approach which may help in dealing with differences between treated and untreated groups in the prior trajectories of outcomes. It may be possible to match pre-treatment schools with a sub-set of untreated schools on their current location in the distribution. We may not be able to obtain exact matches on these scores, but if the scores are in the same general part of the distribution our expectations of year-in-year change may be similar. By combining matching to identify treatment and contrast groups in the same general part of the distribution, with analysis of difference, it can be possible to use difference-in-difference analysis to recover the difference in change over time between treated and counterfactuals (Heckman et al 1997; Machin et al 2004).

In some contexts it may be possible to make use of a naturally occurring distinguishing characteristic to identify an appropriate matched sample. Machin and Vernoit (2011) assess the effects of conversion of schools to Academy status by treating schools which had applied for but not yet converted to Academies as counterfactuals. This assumes that some policy change causes schools that would otherwise have become Academies to continue instead on a counterfactual path. Even aside from the assumptions required about uptake varying across the life of a programme, the approach is not available for the current analysis because, for KTS, there is no group of ‘candidate’ schools.

Matching purely on prior scores is problematic, because as noted above, we are matching schools towards the bottom of the distribution. When we find an untreated school which has a similar score to a treated school in the relevant time period, there is always a chance that the score is unrepresentative of that school’s ‘typical’ trend. Unfortunately, for the matches we find for KTS, the chances of this random error being in the “upwards” direction (relative to school trend) are smaller than that they lie in the “downwards” direction, because there are fewer schools lower down the distribution than the treated schools. In other words, there is a bias in the direction of ‘random’ fluctuations. Hence, in trying to match for schools near the bottom of the distribution (i.e. badly performing schools) we are more likely to tend to select untreated schools which happen to have had an untypically (for them) low score that year. This means that the normal trajectories of matched untreated schools can be different (typically higher) than those of the treated schools we match them to, even though the matching is even-handed. Having a large number of schools available to match from, does not help here as it simply provides more schools which have had an abnormally bad year by chance to select from.

If this happens the observations on which the matched sample are selected are not representative of their general trajectory, and the following year their scores may revert back to their trend (which will on average be in an upwards direction). This has implications for any analysis based on the assumption that the matched sample is a fair counterfactual for the treated sample with only trivial (and essentially unbiased) differences. This upwards reversion in subsequent years will not be seen in the treated schools (if they are treated for genuine, persistent reasons rather than for very short term dips in performance). Since the expected (counterfactual) trajectories are not the same, the comparison would not be equitable, even using a difference-in-difference method on matched schools.

Testing matching methods against data

Matching was attempted on single-year pre-treatment scores for the KTS datasets. Examination of scores predating this point and after showed that selected untreated comparison schools often had unusually low scores in the year of comparison and somewhat higher scores prior to that. The treatment schools did not have these earlier higher scores. This suggests that the comparison schools selected through single-year matching were unsuitable counterfactuals.

In an attempt to overcome the effect of this random fluctuation matching on longer term trajectories - a two year or three year trend - was also attempted. This approach is problematic in terms of data availability as it requires an ample run-in period of data. Even matching on a multiple-year prior trajectory shows some evidence of a ‘dip’ among untreated schools during the years of matching, which suggests they may still provide an unfair comparator group. Figure 2 shows an example of the average pre and post treatment trajectories (on percentage Level 4 at KS2 Maths and English) for KTS schools starting treatment in 2009/2010. Alongside these are plotted the trajectories of untreated schools matched on the three year trend of scores between treatment-minus-3-years (T-3) and treatment-minus-1-year (T-1). With data for many hundreds of untreated schools it is possible to find matches which approximate these pre-treatment KTS trajectories closely as Figure 2 shows in the closeness of lines at these points. However, the trajectory for the untreated group changes direction markedly between T=-1 and T=0, even though these schools have not been exposed to an intervention. What has happened is that we have simply chosen a number of schools which happen to have a poor run of years which are (on average) unrepresentative of their ongoing trajectory. Unfortunately having a large sample of untreated schools is not necessarily helpful in this sense as it simply means there are likely to be a larger number of schools to select which by chance have an 'unusual' deviation in trajectory. Whatever the cause, the schools clearly show a sharp change in direction after this point, and this casts doubt on their use as a plausible counterfactual.

Another approach to matching is to broaden the scope beyond prior scores and match on more characteristics with the aim of identifying other 'markers' of treatability, and thus picking up counterfactuals who are more like the treated schools in critical respects rather than just happening to have similar scores at the time. The hope is that the broader set of criteria adopted would yield a set of schools which are not simply exhibiting temporarily deviant trajectories but which are more generally appropriate as counterfactuals and which thus also may exhibit plausible counterfactual trajectories.

Matching on multiple characteristics can be achieved through an approach called propensity score matching in which as many factors as possible which predict treatment are considered in a model of selection into treatment. The propensity score is the predicted probability of selection (which can be modelled using logistic regression). Usually, having estimated the expected score (probability of selection) for every school, we will find an overlap in the distribution of the score between treated and untreated samples. We can then select data points from the overlapping area of the sample and discard cases with scores with no comparable propensity score matches. If the discarded portions do not render the matched sample unrepresentative of the original population of interest, the matched records can then be treated in analysis as counterfactuals (Heckman et al, 1997; Rosenbaum, & Rubin, 1983).

Such a 'likelihood of treatment' model was applied to create propensity scores for KTS participation. This was defined as the predicted probability of treatment estimated by a model in which onset of treatment was modelled against a number of factors, such as prior scores, local area deprivation, and (for KS4 schools) most recent Ofsted score (specifically the 'problematic' ratings of a school's performance). This yielded results in which treated and untreated schools could be found dispersed across a range of predicted probabilities, which is a good indicator of successful matching. Taking the example of Key Stage 4, of 142 KTS secondary schools 136 were matched on propensity scores to untreated schools. In some instances, no exact match was found, but two untreated schools were found with slightly higher and lower propensity scores than the treatment school and these were both included and weighted down accordingly. The range of scores on matched schools ran from 0.001 to 0.299. Six schools with propensity scores 0.269, 0,280, 0.308, 0.496, 0.573, 0.692 could not be matched. Figure 3 illustrates the distribution of record matching. Although the proportion matched is high, Figure 3 shows that the unmatched cases are concentrated at one end of the distribution, namely that were predicted probability of selection was high. This is a warning that the unmatched sample does not contain schools which were counterfactuals for the KTS schools with the strongest selection probabilities.

As recommended for propensity score matching, the pre-treatment outcome trajectories of treated and untreated schools were compared. Unfortunately the results from checking these prior trajectories did not support the use of this type of matching. Taking the example of Key Stage 4 schools starting treatment in 2003; matched untreated schools had markedly lower scores beforehand and (ignoring the probably coincidental similarity of annual change in the short period 2001-2) had a generally quite strongly positive trajectory over the next few years (Figure 4). Although a difference-in-difference can in principle adjust for differences in prior levels it appears very unlikely that the counterfactual trajectory for treated schools would resemble that plotted for the matched untreated group. Overall, the combination of differences in levels and trajectories are sufficient to raise strong doubts about the matched cases' suitability as counterfactuals.

The question can be raised as to why matching does not yield a plausible counterfactual. There are two potential answers to this; firstly that the matching is not finding appropriate schools because we are not able to control for important unobserved factors such as leadership qualities or school ethos, which determine counterfactual trajectories and selection. Certainly it makes sense to consider that there are a range of factors which influence selection probabilities and that not all these are available in the data upon which we match. In principle this issue could be addressed by extension of the observable factors. A second answer is that that the matching approach was not 'able' to find schools with similar levels on other (measurable) 'relevant' issues because the intervention has selected the majority of the relevant schools (by targeting the most 'relevant' schools in the most 'relevant' areas). This seems plausible given that the schools with propensity scores at the top end of the distribution were the ones that were not matched by untreated schools. The nature of propensity score matching is such that a general shortage of untreated schools with other 'relevant' factors would be compensated for in the propensity scoring by identification of schools with even lower scores on the outcome measure (since lower scores here would balance out higher scores in the 'other' measures in the estimation of the propensity score). This would have the effect of intensifying the problem of selecting comparator schools which 'happened' to be doing badly in a given year. This appears to be the best explanation of why the propensity score matching has yielded untreated schools which have substantially *lower* pre-treatment scores, and is sufficient to raise concerns about the suitability of the post-treatment period trajectories of the untreated schools as counterfactuals for those of the treated schools.

Propensity score matching is most clearly recommended when sample sizes are large (Shadish, Cook, & Campbell, 2002), which would suit analysis of interventions that exhibit measurable variation at the pupil level (see Machin et al 2004), but is less so when variations in inputs can only be identified at the school level. As discussed in preceding sections, the approach appears to be less appropriate where measures are subject to random error and regression to the mean as they are with year-on-year school-level scores, and here the large number of untreated schools may actually lead to misleading matches unless careful checking of prior trajectories is undertaken. For the empirical and theoretical reasons outlined above we do not present the results of analysis based on matching.

Lagged dependent variable model.

In response to the issues discussed above, an approach is required which does not rely on matched samples and which does take account of the differential expectations of change across the results distribution. The approach adopted is to estimate a school's predicted score Y_t conditionally on where the school was in the previous year's distribution of scores Y_{t-1}. A model in which previous values of the outcome variable is treated as a predictor is known as a lagged dependent variable (LDV) model. In a model where Y is the outcome and X_2 is the independent variable of interest, but where Y is also strongly influenced by past values of itself, the model specification is of the form shown in Equation 1.

$$\mathbf{Y_t = \boldsymbol{B_0}X_0 + \boldsymbol{B_1}Y_{t-1} + \boldsymbol{B_2}\boldsymbol{X_2} + \varepsilon} \qquad \text{Equation 1}$$

where

Y_t : Current year's outcome score
X_0 : the intercept.
Y_{t-1} : Previous year's outcome score
X_2 : variable of interest.

The lagged dependent variable approach is appropriate to the estimation of treatment effects in this context, when scores are strongly related to preceding years' scores, and where pre-treatment scores for the treatment group are located below the average. The LDV model allows for the consideration of factors affecting current scores (such as treatment) whilst controlling for the effect of the preceding year's score. If the scores are standardised within-year this means that we are controlling for the location in the distribution in the preceding year. If we expect more positive change at the bottom of the distribution this will therefore be controlled for, which means the estimated effect of treatment is net of differential expectations of change due to location in the distribution. This addresses the confounding effect of regression to the mean. Conversely, it may be the case that positive change is hard to achieve among treated schools due to persistent contextual factors like local area deprivation, which might make the treatment appear less effective than regular (non-intervention based) efforts by schools to improve their performance. The effect of these persistent local effects, whether observed or unobserved, can also be controlled for through the inclusion of the lagged dependent variable as the effect of these persistent factors is likely to already be embodied in previous years outcomes. The lagged dependent variable has some of the benefits of a fixed effect model in that it controls for unchanging school-level factors, but in modelling year-on-year change it also captures the differential effect of location in the distribution which is important in this context.

As noted above, the LDV specification adopted allows us to model treatment effects conditional on location in the distribution, and through this to control for the influence of contextual factors as well as adjust for the differential effect of regression to the mean. Multiple observations are available for each school and in some analytical contexts, simply modelling treatment-status of the current observation while restricting the analysis to the treatment group would directly identify the treatment effect. However, to more reliably estimate the confounding effect of contextual factors (by having a bigger, more representative sample) it might be beneficial to extend the analysis to include untreated units. In this case the treatment status variable would need to be accompanied by a treatment-group membership identifier (1 for all observations for the treatment group, zero otherwise) in order to control for the pre-treatment differential between treated and untreated.

For these reasons, the analysis presented includes untreated cases, to allow us to more accurately model the average effect of location in the distribution and (indirectly) of contextual factors. Hence in order to recover the treatment effect we need to include a term for treatment group membership.

Both a treatment-group membership term and treatment status term are therefore included in the specification adopted. Using these two predictors, and contextual controls, it would be possible, if expected annual change was consistent between treated and untreated groups, to estimate the treatment effect relative to an outcome defined as annual change in the score (Y_t minus Y_{Tt-1}). However, this approach would not deal with the confounding effect of location in the distribution. If year-to-year change in scores is highest when scores are lowest (which is suggested by Figure 1 and differences reported earlier), scores will (on average) be improving less quickly after treatment start simply because, by that time, they are nearer the mean score (regardless of whether this is due to regression to the mean or the effect of treatment). This means that the effect of treatment-status on the rate of change may be masked by the simple change in location in the distribution that may occur afterwards.

It is for these reasons that the LDV specification is adopted, rather than matching, in order to control for the effect of location in the distribution. The LDV considers what the current expectation of a school's scores is given their previous year's location in the distribution. The differences in the strength of tendency towards the mean according to location in the distribution is controlled for, which means that the differences remaining between schools in their year-to-year expectation can be ascribed directly to their status as pre-treated, post-treated or untreated.

Implementation

Models with LDVs can be estimated using OLS (Ordinary Least Squares), making them relatively straightforward to implement. In general the model can produce consistent estimators for B_3 when sample sizes are large and there is no autocorrelation (Keele and Kelly, 2006). Autocorrelation occurs when the residual (error) between successive observations in data is correlated. With annual school level data this means observations from successive years. Under these circumstances the model with LDV may no longer be an unbiased estimator (Davidson and MacKinnon, 1993)[2]. However, Keele and Kelly (2006) show that when a model is demonstrably dynamic, i.e. when the absolute value of the lagged coefficient nears 1, the effect of autocorrelation on bias is not as strong as is sometimes suggested in the literature. Specifically, using sampling simulations they show that with B_1 (the LDV coefficient) at 0.75 and a modest sample size (n=250), the autocorrelation can rise to 0.20 with only a fairly small bias arising in the estimate for B_3 (-5.19% in relative terms). For an evaluation model this would bias estimates of the treatment effects downwards (i.e. it would be conservative). Since the present study offers large samples, which will help reduce bias through convergence, and since the observed random movement from year-to-year suggest that year-to-year autocorrelation in errors may not be excessive, we adopt an LDV model, and report the levels of autocorrelation.

Another issue concerns the reporting of standard errors. Since the scope of the analysis is as a retrospective evaluation of a specific intervention (KTS) rather than projection of equivalent effects for other contexts we do not propose to make inferences to other populations. In terms of the post hoc assessment of KTS effects, data is available on the entire school population and so we are not seeking to support inference from a sample to a wider population, or to make predictions. Parameter estimates are therefore a primary concern rather than standard errors. For this reason

[2] Although it does remain consistent, which means that as sample size increases the bias diminishes.

we present results from models specified using standard OLS estimators, with attention paid to issues of bias in estimators of effect sizes, but not to inflation or otherwise of estimated standard errors. Hence we present the estimated standard errors only as a guide to the reliability of the results.

Interpretation

Scores are standardised within each year so that the annual scores are de-trended. This means that we are modelling only change in the location of a school in the distribution. Unless the estimated value of the lagged dependent variable coefficient (B_1) is equal to 1 we will expect this year's scores to be different from last year's. Since we are using standardised scores we have removed the possibility that the distribution is stretching due to polarisation or contracting due to bunching. Instead we are only concerned with changes in location in the distribution. A reasonable expectation for B_1 is therefore that it will be less than one, due to regression towards the mean. In other words our expectation for scores below the mean will be that they revert upwards (to the mean) and for scores above the mean that they will revert downward (towards the mean).

The value for B_1 forms the basis for conditional expectations in the model. As noted earlier, it also is likely to control for some of the effects of other contextual variables, such as deprivation, since the effect of these factors will also be manifested in previous year's results.

Using previous years' scores as our basis for expectations of current scores we can then identify the underlying differential associated with being a member of the treatment group. The differential on scores (raw or standardised) is likely to be negative, given the selection criteria. However, in the current specification we are speaking of the differential conditional on the prior year's location in the distribution. If intervention was carried out on schools which happened to be doing badly in a given year, but which were just as good (pre-treatment) as other schools at reverting towards average scores then this conditional group differential will be zero. Arguably schools in this situation are not really candidates for selective intervention, since they will do just as well as untreated schools in the absence of treatment. A negative coefficient for treatment group differential (B_2) will signify that the treatment group were less good than untreated schools at bringing poor results up towards the mean. In some respects therefore the estimated group differential in this specification will be a useful measure of 'appropriateness' of treatment selection, although we will not comment on this at length.

On top of the treatment group differential, using the year-by-year information about the onset of treatment, we can identify the treatment effect (B_3). To achieve this we model the effects of treatment across the multiple annual observations for each school (rather than using a single before-and-after measurement approach). Treatment is coded as "on" for all years following from the initial onset of intervention[3], since it is assumed that the lessons learned are not lost when the programme intervention is discontinued. The core specification for the analysis is therefore as shown in Equation 2.

$$\mathbf{Y_t = \mathit{B}_0X_0 + \mathit{B}_1Y_{t-1} + \mathit{B}_2X_2 + \mathit{B}_3X_3 + \varepsilon} \qquad \text{Equation 2}$$

where

Y_t : Current year's outcome
X_0 : Constant
Y_{t-1} : Previous year's outcome
X_2 : Being a treatment-group school (all observations)
X_3 : intervention (e.g. any observation from year of intervention onwards)

Positive estimates for coefficient B_3 can be interpreted as indicating a more positive year-on-year change in results after the onset of treatment compared with pre-treatment, adjusting for the 'typical' expectation given the (previous year's) location in the distribution. All scores are standardised within-year, so the onset of treatment (a one-unit increase in X_3) is associated with a B_3 standard deviations increase or decrease in scores relative to pre-treatment and adjusting for the schools location in the previous year's distribution of scores. As noted above, the outcome measure scores are standardised within year. This allows the models to handle slight inconsistencies in the scaling of data across years while still representing important distributional properties of outcomes, and also de-trends the data which simplifies interpretation and reduces the possibility of confounding if periods of faster growth in average scores happen to occur in years when more treatment occurs. Standardisation also allows some (informal) comparisons between model specifications and between KS2 and KS4. The estimated effect sizes are presented in tables in this standardised metric. To facilitate interpretation the relationship between the standardised score and the original outcome scale is presented for KS4 in Table 1 and for KS2 in Table 8, where the scaling is shown for each year. At key places in the text, results are reported in both standard deviations and the original metrics.

3 Results

3.1 Key stage 4

a) Confirming the effect of location in the distribution

[3] The first year of 'intensive' or 'unspecified' intervention

Analysis of key stage 4 data from all years 2001-2011 for all secondary schools is carried out to confirm that outcome scores are strongly predicted by the previous year's outcome (Equation 3) - one of the requirements of using an LDV model.

$$\mathbf{Y_t = \mathit{B}_0X_0 + \mathit{B}_1Y_{t-1} + \varepsilon} \qquad \text{Equation 3}$$

where

Y_t : Current year's outcome - percentage of pupils attaining 5+ GCSEs (standardised)
X_0 : Constant
Y_{t-1} : Previous year's outcome

Last year's score is a strong positive predictor of the current year's score (Table 1, B_1). Predicted scores are slightly closer to average scores than the previous year's, suggesting some regression to the mean (Table 2). Figure 5 plots the expected value of this years score against last years score.

b) Treatment Effect

Having confirmed the predictive power of the previous year's score, and thus the dynamic nature of the model, we model the group differential in outcomes associated with being in the KTS program and the additional effect of the onset of treatment (Equation 4).

$$\mathbf{Y_t = \mathit{B}_0X_0 + \mathit{B}_1Y_{t-1} + \mathit{B}_2X_2 + \mathit{B}_3X_3 + \varepsilon} \qquad \text{Equation 4}$$

where

Y_t : Current year's outcome - percentage of pupils attaining 5+ GCSEs (standardised)
X_0 : Constant
Y_{t-1} : Previous year's outcome
X_2 : Being a KTS school (all observations)
X_3 : KTS intervention (e.g. any observation from year of intervention onwards)

Members of the set of schools exposed to KTS intervention have a lower expected result relative to their location in the previous year's distribution than untreated schools (Table 3, columns 1-2, B_2). The coefficient for treatment onset is positive (Table 3, columns 1-2, B_3). This means that the scores for KTS schools having started treatment are more positive than they were before treatment on average, controlling for their location at all times in the distribution. The difference is around a tenth of a standard deviation. The standard deviation for percentage attaining 5+ GCSEs in its original metric is around 20% (Table 1), across the period 2003-2011, which implies a per-year treatment effect of around +2% in the school-level percentage attaining 5+ GCSEs, relative to location in the distribution and previous progress. This is the key result for Key Stage 4. Figure 6 illustrates this effect across values of previous year's results with the dashed black line representing the expected score given location in the distribution (B_1*Y_{t-1}), the blue line representing the effect of adding the (negative) treatment group differential ($B_1*Y_{t-1}+B_2$) and the red line being the net expected value for the treated group once treatment has started ($B_1*Y_{t-1}+B_2+B_3$). The net effect of group differential and treatment effect is such that after treatment commences predicted values for treated schools are comparable to untreated schools at a similar location in the previous year's distribution. Interpreted causally, this does not mean that treatment immediately brings schools up to have average scores, but rather that it brings their ability to progress towards the average score up to a more typical level for a school. This finding is consistent with how we might expect a school improvement intervention to operate if successful.

This key finding for KS4 of a positive coefficient for B_3 is consistent with the idea of a positive effect for treatment, controlling for prior location in the distribution. We now apply tests of robustness to the results, testing for bias due to problems with error structure, functional form, omitted variables and issues arising from heterogeneity of outcomes, with the latter including some analysis of effects for subgroups and regions.

c) Bias assessment

- Error Structure

As noted earlier, autocorrelation may lead to a bias in the estimates of coefficients. Keele and Kelly suggest this bias will be modest (around 5% of the coefficient size) if certain conditions are met. For the current model, autocorrelations are a little higher than Keele and Kelly's test value of 0.20 ($|\boldsymbol{\rho}|$ = 0.248), but the magnitude of the coefficient for the lagged dependent variable suggests that the model is dynamic (B_1 = 0.867) and the large sample size ($\mathbf{N_{obs}}$ = 27,430) should mitigate bias in estimates of coefficients for treatment effects and make it unlikely that the finding is qualitatively misleading about the nature of the treatment effect.

- Functional Form

The specification adopted assumes that the treatment group differential and treatment effect are both uniform across the range of prior values. This is a simplifying assumption in order to identify a single-number treatment effect. However, it is possible that although the average treatment effect is positive the effect varies across the distribution of prior results. If so it might yield more positive effects for higher performing schools and less positive effects for the poorer performing schools which arguably need it most. This would be qualitatively important in our interpretation of results. To assess this possibility we apply a specification in which the relationship between treatment status and results can vary across prior results (by interacting treatment status with prior results). We also relax the assumption of linearity in all the terms (by including a quadratic for each term). Figure 7 illustrates the effect of the re-specification by plotting predicted outcome scores against the relevant conditions[4]. The plot suggests that it is likely that in the regions where we would expect the intervention to be most important from a policy perspective, with more poorly performing schools, the substantive key finding holds good, namely that the net effect of onset of treatment is positive for the treated group. In addition the secondary finding that net expectations when treatment commences are broadly similar to those for untreated groups is also maintained.

- Omitted Variables

Our primary specification does not explicitly control for contextual factors which might confound the effects of treatment. In part this is because the specification controls for pre-treatment KTS group differentials and there is only likely to be a modest change in these contextual factors (whether observed or unobserved) within these schools over the period of the analysis. In addition, the lagged dependent variable controls for confounding effects which influence outcomes up until the preceding year's measurement. However, it is worth testing the assumptions about how the model is functioning by carrying out a sensitivity test of the results to the inclusion of controls for (observable) contextual variables.

Further controls were thus added to the existing specification and the results were as follows.

Deprivation tends to be higher in inner city schools, which means we can expect levels to be high in treated schools. Deprivation is measured through the Index of Deprivation Affecting Children (IDACI) which is defined for our analysis through the post-code of the school. This is a proxy for the area-level deprivation affecting performance of pupils (which may vary from year to year with intake) and school. Controlling for IDACI does not markedly affect the treatment group differential or treatment effect (Table 3, columns 3-4, B_2 & B_3).

School type is another potential confounder of the effect of treatment if treatment is concentrated in specific school types that, for reasons unconnected with KTS interventions, are improving faster or slower than average. Information about school type varies in form between primary and secondary school, but broadly relates to whether the school is a community school, has selective intake policies and gender-selection and the type of funding status. Adding school type dummies to adjust for

[4] $\mathbf{Y_t = \mathit{B}_0X_0 + \mathit{B}_1Y_{t-1} + \mathit{B}_2X_2 + \mathit{B}_3X_3 + \mathit{B}_4X_2{*}Y_{t-1} + \mathit{B}_5X_3{*}Y_{t-1} + \mathit{B}_8{Y^2}_{t-1} + \mathit{B}_6X_2{*}{Y^2}_{t-1} + \mathit{B}_7X_3{*}{Y^2}_{t-1} + \varepsilon}$

differences between school types again made little difference to the estimated effect of treatment (Table 3, columns 5-6, B_3).

School change. Another potential driver of change in results is a change in school identity. Struggling schools have in some cases been relaunched with a new identity or premises, or been relaunched as academies. These processes overlap with the Keys to Success intervention and may thus confound (positively or negatively) the treatment effects of KTS. However, if we exclude from the analysis all schools where such institutional change happens we might discard some of the potentially positive effect of KTS where this operates through reformed schools. For this reason we have not excluded schools which changed in these ways, but instead treated institutions with an equivalent role to predecessor school in the local school market as having a continuous identity in our analysis, while trying to test for any confounding effects of the other changes.

We denote a consistent identity for educational institutional units by using Local Authority Establishment as the identifier across time. In this sense we are testing the effect of KTS on localised educational outputs rather than specifically on a school name. This approach is consistent with the spirit of City Challenge as an intervention aimed at local areas rather than at specific named institutions. Nonetheless, we can test whether a change in school identity does affect our estimate of the impact of treatment by controlling for this type of change in the analysis. This is defined by whether the school changes its identity during the measurement period (not necessarily during the treatment period), either in terms of name or address. This picks up changes to academy status or other rebranding or relaunching along with less radical shifts in location. Table 3 (columns 7-8, B_3) shows that there is virtually no change in the estimated treatment effect when school change is controlled for. This implies that the treatment effects we identify in the analysis are not *driven* by confounding changes (such as conversion to academy). In other words, if interpreted causally, KTS intervention is associated with positive change net of the effect of accompanying school re-organisation.

In order to further confirm the lack of a confounding effect of omitted variables. Table 4 shows the results of modelling treatment *within* each relevant sub-population of secondary schools. As before, estimated effect of the onset of treatment among treated schools is given by the B_3 coefficient (Equation 3). However the expected result based on previous results is estimated purely within the subgroup in question, so that we are estimating treatment effects within a subpopulation of schools. The estimated treatment effects are relatively similar in all six subgroups, with no marked variations from the pattern for the whole sample (Table 4, columns 1-6, B_3). This is consistent with the idea that the KTS treatment effects are not the result of (driven by) systematic differences in other covariates such as a school's intake, type or institutional reorganisation.

- Heterogeneity of Outcomes

The percentage attaining 5+GCSEs is the primary KS4 measure examined in this analysis as it is a reliable and consistent measure across the years of data. The preceding analysis suggests that KTS was linked with an improvement in the outcome. A concern with using this outcome measure to assess impact would be that improvements in educational practice would be concentrated on work undertaken with pupils at the borderline (threshold) for the indicator in order to drive up the school-level results on the target indicator to the maximum degree possible relative to effort. If this were the case, the results for all children (including those not close to the threshold) may be smaller than the results for the target indicator suggest, which again would have policy-relevant implications. Alternative measures can provide some insight into whether this is the case.

Average scores. Table 5 (columns 1-2) shows the results for a model in which the dependent variable (and lagged dependent variable) are mean GCSE points score. This information is not available as consistently across all years as percentage attaining 5+GCSEs, as in some years the points score is capped, but standardising within-year reduces the impact of this change. As with preceding analysis the previous year's score is a strong predictor of current score, with a modest degree of regression to the mean (B_1). KTS schools have a less positive underlying level of year-on year change (B_2). However, the KTS treatment effect is once again positive (B_3). The results are not sensitive to the addition of controls for school level characteristics (Table 5, columns 3-4).
The magnitude of the per-year treatment effect for school-level average is comparable with that for school-level percentage at around a tenth of a standard deviation[5]. If the treatment effect were concentrated solely among children just below the threshold (due to a focus on targets by the school) there would have to be a very sizable effect among this group to yield an average effect of the magnitude estimated here for the whole school. Such a disparity would be readily detected in descriptive analysis of change in results. It is therefore more plausible that the impact is not concentrated solely on pupils drawn from a narrow ability-range.

Value-added scores. Analysing value-added scores represent another way to test the heterogeneity of effects. If KTS led to schools focusing efforts on children who could cross the relevant target

[5] A standard deviation is around 75 points in the original metric (centred around 400) between 2005 and 2011, so this implies a per-year treatment effect of around +10 points, relative to location in the distribution and previous progress.

threshold, the easiest approach would be concentrating on children who were relatively able, but underperforming. If this strategy was adopted the effect on value added-scores might turn out to be smaller than the effect for absolute results.

As noted in the examination of omitted variable bias, there is no evidence that the treatment effects are attributable to changes in the school identity. One of the ways that school transformation, rebranding or a change in reputation associated with KTS intervention might have an effect on outcomes is by encouraging a rapid intake of more able pupils from surrounding areas. This would drive up results in a way which would inflate the treatment effects unfairly. Analysing with respect to value-added scores helps controls for this effect as the prior attainment of the new recruits is already fixed[6].

Value-added scores are not available consistently across the entire period for Key Stage 4. They are not available before 2002. For 2002-3 they are available relative to Key Stage 3, and from 2004 onwards relative to Key Stage 2. Once again standardisation within-year can ameliorate some of the impact of this discontinuity by translating raw scores on differing scales into distributional ones.

Table 5 (columns 5-6) shows the effects for analysis with standardised value-added scores as the outcome measure. As with other outcomes, the KTS group differential (B_2) is negative, conditional on previous year's results. However, as before, the KTS treatment effect (B_3) is positive. The magnitude of the per-year treatment effect for value added is around two tenths of a standard deviation[7]. There is therefore no evidence to suggest that the treatment effect for value-added outcomes is smaller than for absolute school-level scores or percentages as might be expected if the schools were focusing efforts on pupils who were closest to threshold targets. Once again the results are not sensitive to the addition of controls for school and area level characteristics (columns 7-8).

Region. The analysis is repeated with treatment schools restricted firstly to those in London (Table 6, columns 1-4), secondly The Black Country (columns 5-8) and thirdly Greater Manchester (columns 9-12). Untreated schools continue to be drawn from all English schools. The results (with and without controls for area and school factors) suggest that there may be some variation by location in group membership differential (B_2) and the treatment effect (B_3), conditional on the location in the distribution, although this is not a straightforward comparison since the average number of years of treatment over which we measure is different in the three areas. A specification which interacts

[6] The new recruits may bring a higher quota of value-added attainment from their previous intermediate schooling location but it is less plausible that large numbers of pupils already benefitting from higher value added at their current location would switch to a recently struggling school.

[7] A standard deviation is around 20 points in the original metric (centred around 1000) between 2005 and 2011, so this implies a per-year treatment effect of around +4 points, relative to location in the distribution and previous progress.

treatment region with pre-treatment group differential and treatment effect[8] suggests that Greater Manchester has a less negative pre-treatment differential than London, but no other 'significant' differences by region. Even this minimal finding should be treated with caution given the relatively small samples, varied treatment periods by region and the introduction to the LDV model of interaction terms which may have different properties in terms of bias arising from autocorrelation. For these reasons we do not report the results of these specifications. It should be noted that despite the regional variation, the regional treatment effects are generally such as to offset the regional KTS group differential and bring the year-on-year change approximately into line with the general average given the location in the distribution.

The intention of the present analysis is to quantify the impacts of treatments in their context, rather than to compare impacts across contexts. Although directly comparing the magnitude of impacts between regions may be inadvisable, it is possible to test the sensitivity of the findings about regional treatment effects to altered assumptions about the uniformity of the conditional expectations used in the model. Preceding specifications have made use of data about untreated schools from across England to estimate the average expectation for change relative to the previous year's location in the distribution. If the assumption of uniformity of this effect across regions is relaxed it offers a sensitivity test of the findings about treatment effects within each region. Table 7 shows the result of allowing the effect of prior location in the distribution to vary by region. This is achieved by running separate models in each area (with only local untreated schools included rather than a common pooled set of all English schools). The treatment effects (B_3) are relatively unaffected (as we would hope if the model is estimating impact), but the KTS pre-treatment differential and effect of location in the distribution alter, with the KTS differential (B_2) becoming larger. The pre-treatment differential is now too large to be offset by the treatment effect in each of the three regions. The pre-treatment difference between KTS schools and untreated schools in making progress appears larger in (negative) magnitude when the contrast is made only with local untreated schools than when made with a national average. In other words evaluation of the scale of the KTS impact relative to the scale of the problems experienced by the relevant schools may vary depending on whether a national or local perspective on the nature of the 'problem' is applied.

[8] $\mathbf{Y_t = \mathit{B}_0X_0 + \mathit{B}_1Y_{t-1} + \mathit{B}_2X_2 + \mathit{B}_3X_3 + \mathit{B}_4X_{reg2} + \mathit{B}_5X_{reg2}X_2 + \mathit{B}_6X_{reg2}X_3 + \mathit{B}_7X_{reg3} + \mathit{B}_8X_{reg3}X_2 + \mathit{B}_9X_{reg3}X_3 + \varepsilon}$

3.2 Key stage 2

For Key Stage 2, interventions are restricted to the second half of the decade, so data on attainment is used from 2005-2011. A composite average of the percentage attaining level 4 at KS2 English combined with percentage attaining level 4 at KS2 Maths is used, again standardised within-year[9]. Treating the subjects separately did not give not markedly different results (results not presented).

a) Confirming the effect of location in the distribution

As with Key stage 4, the predicted value of results relative to previous years' results (Equation 5) showed some regression to the mean (Table 9, Figure 8).

$$\mathbf{Y_t = \mathit{B}_0X_0 + \mathit{B}_1Y_{t-1} + \varepsilon} \qquad \text{Equation 5}$$

where

Y_t : Current year's outcome - average (across English & Maths) of percentage of pupils attaining level 4 in KS2 (standardised)
X_0 : Constant
Y_{t-1} : Previous year's outcome

b) Treatment Effect

Having confirmed the dynamic nature of the model for KS2, Table 10 and Figure 9 show the results of modelling the effect for the onset of treatment (Equation 6).

$$\mathbf{Y_t = \mathit{B}_0X_0 + \mathit{B}_1Y_{t-1} + \mathit{B}_2X_2 + \mathit{B}_3X_3 + \varepsilon} \qquad \text{Equation 6}$$

where

Y_t : Current year's outcome - average (across English & Maths) of percentage of pupils attaining level 4 in KS2 (standardised)
X_0 : Constant
Y_{t-1} : Previous year's outcome
X_2 : Being a KTS school (all observations)
X_3 : KTS intervention (e.g. any observation from year of intervention onwards)

As with KS4 the results suggest that Primary schools had a negative group differential (B_2) as might be expected, indicating less capacity to raise their scores towards the average. Schools having started treatment show positive treatment effects of around four tenths of a standard deviation. The standard deviation for percentage attaining Level 4 at KS2 in its original metric is around 13%

[9] Each subject score is standardised within year then, the average is taken, and these averages standardised within year again.

(Table 8) across the period 2005-2011, which implies a per-year treatment effect of around +5% in percentage attaining level 4 in English or in Maths, relative to location in the distribution and previous progress. This is the key finding for Key stage 2. Since the average observed length of treatment is only a year or two for this intervention we might expect to see some diminution of the average per-year effect over the longer term. Figure 9 illustrates this effect across values of previous year's results with the dashed black line representing the expected score given location in the distribution (B_1*Y_{t-1}), the blue line representing the effect of adding the (negative) treatment group differential ($B_1*Y_{t-1}+B_2$) and the red line representing the net expected value for the treated group once treatment has started ($B_1*Y_{t-1}+B_2+B_3$). The net effect of group differential and treatment effect is such that after treatment commences predicted values for treated schools are comparable to untreated schools at a similar location in the previous year's distribution.

c) Bias assessment

- Error Structure

The value for autocorrelations ($|\boldsymbol{\rho}|$ = 0.204), is within the range assessed by Keele and Kelly, and although the coefficient for the lagged dependent variable (0.625) is lower than the test value of 0.75, sample size is large ($\mathbf{N_{obs}}$ = 73,880) which should mitigate (through convergence) against bias in estimates of coefficients for treatment effects. It is unclear whether the bias is theoretically larger or smaller than was the case for KS4 analysis, because the reduction in dynamic nature of the model may be offset by the increase in sample size. Again, there is no reason to think that the bias is substantial.

- Functional Form

Figure 10 illustrates the effect with a relaxation of the assumption of linearity[10]. As with KS4, the plot shows that in the regions where the intervention is likely to be considered most policy relevant the results hold good, namely that the net effect of onset of treatment is positive for the treated group. The secondary finding, that net expectations when treatment commences are broadly similar to those for untreated groups is also supported.

- Omitted variables

Columns 3-8 of Table 10 show the results of sensitivity tests involving adding control for deprivation, school type and change in school identity. As with Key stage 4 the controls for deprivation explain some of the pre-treatment difference between treated and untreated schools. However, as with KS4

[10] $\mathbf{Y_t = \mathit{B_0}X_0 + \mathit{B_1}Y_{t-1} + \mathit{B_2}X_2 + \mathit{B_3}X_3 + \mathit{B_4}X_2{*}Y_{t-1} + \mathit{B_5}X_3{*}Y_{t-1} + \mathit{B_8}Y^2_{t-1} + \mathit{B_6}X_2{*}Y^2_{t-1} + \mathit{B_7}X_3{*}Y^2_{t-1} + \varepsilon}$

again, the estimated effect of treatment onset is not markedly reduced by the addition of the controls.

Table 11 shows the results of modelling treatment effects within the relevant sub-populations of primary schools, high and non-high deprivation, community schools and other schools, schools which do not change identity across time (name and postcode) and those that do. There are no marked exceptions to the result described for the whole sample.

- Heterogeneity of Outcomes

The average of the percentage attaining level 4 at Key stage 2 in English and in Maths is the primary KS2 measure examined in this analysis and analysis suggests that KTS was linked with an improvement in the outcome. Alternative measures can provide some insight into whether this effect is driven by changes at the threshold of the target.

Average scores. Table 12 (columns 1-2) shows the results for a model in which the dependent variable (and lagged dependent variable) are the school-level mean KS2 points score (averaged for English and Maths). As with preceding analysis the previous year's score is a predictor of current score, with some regression to the mean. The magnitude of the treatment effect for school-level average is comparable with that for school-level percentage at around four tenths of a standard deviation[11]. There is thus no evidence to support the idea that schools have improved their inputs only for children close to the threshold. Results are not sensitive to controls for school and area-level factors.

Value-added scores. Analysing value-added scores represent another way to test the heterogeneity of effects and the confounding effects of changes in intake due to changes in school reputation.

Table 12 (columns 5-6) shows the effects for analysis with standardised value-added scores as the outcome measure. The direction of the differential for the underlying effect of being a KTS school and the treatment effect are consistent with earlier results (negative and positive respectively). The magnitude of the treatment effect for value-added score is around five tenths of a standard deviation[12]. The results do not support the idea that the impact is concentrated near the threshold. In addition, compared with other KS2 outcomes, the ratio of the treatment effect to group

[11] A standard deviation is around 1.66 points in the original metric (centred around 27.5) between 2005 and 2011, so this implies a per-year treatment effect of around +0.66 points, relative to location in the distribution and previous progress.

[12] A standard deviation is around 1 unit in the original metric (centred around 100) between 2005 and 2011, so this implies a per-year treatment effect of around +0.5 points, relative to location in the distribution and previous progress.

differential is larger, giving a more positive net effect, which may suggest that the intervention helped primary schools add value rather than simply attain threshold targets. Again the results are not sensitive to addition of controls (columns 7-8).

Region. The analysis is repeated separately with treatment schools restricted to London (Table 13, columns 1-4), The Black Country (columns 5-8) and Manchester (columns 9-12), but with all untreated schools as contrast in each case. The results suggest there may be some variation by region both in the pre-treatment group differential and the treatment effect, conditional on the location in the distribution. Specifications comparing explicitly between groups[13] suggest that Manchester has a less positive treatment effect than London and that the Black Country has a more negative pre-treatment group differential than either London or Manchester, although these findings should be treated with caution given the relatively small samples, short treatment periods and the introduction to the LDV model of interaction terms (results not presented).

The intention of the analysis is to quantify the impacts of treatments in context, rather than to compare impacts across contexts, and although comparing the magnitude of impacts between regions may be inadvisable it is possible to test the sensitivity of the findings about treatment effects to altered assumptions about the uniformity of the conditional expectations used in the model. The specifications have made use of untreated data to estimate the average expectation for change relative to the previous year's location in the distribution. If this assumption of uniformity across the regions is relaxed it provides a useful sensitivity test of the findings about treatment effects within each region.

In table 14 the effect of location in the distribution is allowed to vary by region. As was the case for KS4 the treatment effects (B_3) are relatively unaffected, but the KTS differential and effect of location in the distribution change somewhat, with the KTS differential appearing larger (more negative), such that the positive effect of the intervention is no longer broadly offsetting the KTS differential. As noted for analysis of KS4, the pre-treatment difference between KTS schools and untreated schools in progressing toward the mean appears greater when the benchmark is local schools rather than the national average. Evaluation of KTS impact relative to the problems faced by targeted schools may vary depending on whether a national or local perspective on the problems is applied.

[13] $\mathbf{Y_t = \mathit{B}_0X_0 + \mathit{B}_1Y_{t-1} + \mathit{B}_2X_2 + \mathit{B}_3X_3 + \mathit{B}_5X_{reg2}X_2 + \mathit{B}_6X_{reg2}X_3 + \mathit{B}_8X_{reg3}X_2 + \mathit{B}_9X_{reg3}X_3 + \varepsilon}$

4 Conclusions

Identification of treatment effects for KTS is not straightforward due to the non-random selection into treatment and the distributional location of treated schools which brings different 'expected' levels of change. Due to the nature of the distributional properties of the data and other theoretical considerations it has not been appropriate to apply difference-in-difference and / or matching approaches to compensate for these issues. However, analysis of change relative to previous results using a lagged dependent variable specification with appropriate robustness checks provides some evidence that treatment is associated with positive change above and beyond what would be expected in the absence of treatment, and also provides some information about how this compares with typical patterns of change among untreated schools. We estimate a positive impact of 2% per year in school-level achievement of KS4 targets (2003-2011) and 5% per year for KS2 targets (2008-2011), although the latter in particular are based on rather short-run observation periods. We find some evidence (applying a causal interpretation) that a typical effect at both KS2 and KS4 seems to be to raise year-in-year change towards the nationally 'typical' amount for untreated schools in the same part of the results distribution. There may be some indication that KS2 impacts are larger relative to the overall distribution of results than those for KS4, but this may alter as treatment periods lengthen for KS2 and may be conditioned by the smaller number of subjects embodied in the KS2 outcome measure assessed. We also find some evidence that impact varied by region, with a tentative finding of larger impact in London than Manchester at both KS2 and 4, and a similarly tentative finding that KTS may have been applied to primary schools with greater relative prior difficulties in attaining typical year-in-year growth in the Black Country than in other regions.

Tables

Table 1 - Secondary schools: All schools, all years

Dependent variable: school-level percentage of pupils attaining 5+ GCSEs (standardised)

Year	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*
1 Standard Deviation (SD) increase in Dependent Variable =											
	23.7	23.2	22.8	22.4	21.4	20.5	19.3	17.9	16.4	14.6	18.8
Percentage point Increase in *pupils attaining 5+ GCSEs*											

Table 2 - Secondary schools: All schools, all years

	Dependent variable: school-level percentage of pupils attaining 5+ GCSEs (standardised)	***B***	***Sig.***
	(Constant)	-.031	
Y_{t-1}	Previous year's GCSE score (standardised)	.871	***

R^2= .758; **N_{obs}** = 27,430; **$N_{schools}$** = 2,743; **$\rho_{autocorrelation}$** = -0.251; **$d_{DurbinWatson}$** = 2.408

Table 3 - Secondary schools: All schools, all years

	Dependent variable: school-level percentage of pupils attaining 5+ GCSEs (standardised)	Model		+ IDACI		+ type		+ change	
		B	***Sig.***	***B***	***Sig.***	***B***	***Sig.***	***B***	***Sig.***
	(Constant)	-.029		-.032		-.082		-.088	
Y_{t-1}	Previous year's GCSE score (standardised)	.867	***	.860	***	.825	***	.824	***
X_2	KTS school (all observations)	-.116	***	-.104	***	-.108	***	-.108	***
X_3	**KTS treatment (during treatment period)**	.109	***	.113	***	.108	***	.108	***
	IDACI ranking			✔		✔		✔	
	Dummies for school type					✔		✔	
	Dummies for school change							✔	
	R^2=	.758		.759		.763		.763	
	N_{obs}=	27,430		27,430		27,430		27,430	
	$\rho_{autocorrelation}$ =	- .248		- .242		- .216		- .216	
	$d_{DurbinWatson}$ =	2.403		2.393		2.356		2.356	

$N_{schools}$ = 2,743;

Table 4 - Secondary schools: Restricted samples, all years

	Dependent variable: school-level percentage of pupils attaining 5+ GCSEs (standardised)	Deprivation		School type		School change	
		high*	non-high	CYmix	Other	No	Yes
		B	***B***	***B***	***B***	***B***	***B***
	(Constant)	-.073	-.022	-.094	.012	-.028	-.030
Y_{t-1}	Previous year's GCSE score (standardised)	.834	.873	.795	.900	.869	.860
X_2	KTS school (all observations)	-.118	-.108	-.116	-.148	-.111	-.133
$\mathbf{X_3}$	**KTS treatment (during treatment period)**	.114	.107	.103	.128	.103	.123
	R^2=	.701	.762	.640	.818	.760	.752
	N_{obs}=	5,530	21,900	15,970	11,460	21,090	6,340
	$\rho_{autocorrelation}$ =	- .184	-.265	-.192	-.283	-.250	-.233
	$d_{DurbinWatson}$ =	2.292	2.438	2.306	2.465	2.329	2.119

Notes: all coefficients significant at 1% level (excluding constants).
* below 25th percentile of IDACi rankings.

Table 5 - Secondary schools: All schools, all years.

		Dependent variable: school-level mean GCSE points score (standardised within-year)*				*Dependent variable: school-level KS4 value-added score** (standardised within-year)*			
		Model		+ controls		Model		+ controls	
		B	***Sig.***	***B***	***Sig.***	***B***	***Sig.***	***B***	***Sig.***
	(Constant)	.011		.000		-.002		-.045	
Y_{t-1}	Previous year's outcome score (standardised)	.897	***	.884	***	.703	***	.694	***
X_2	KTS school (all observations)	-.099	***	-.099	***	-.152	***	-.139	***
$\mathbf{X_3}$	**KTS treatment (during treatment period)**	.131	***	.128	***	.213	***	.204	***
	IDACI ranking			✔				✔	
	Dummies for school type			✔				✔	
	Dummies for school change			✔				✔	
	R^2=	.784		.785		.495		.498	
	N_{obs}=	25,758		25,758		23,167		23,167	
	$\rho_{autocorrelation}$ =	-0.075		-0.065		-0.066		-0.062	
	$d_{DurbinWatson}$ =	2.137		2.120		2.110		2.104	

* mean points score available 2001-2011. Lagged model available 2002-2011.
** value-added figures available 2002-2003 (Key stage 3-4) and 2004-2011 (key stage 2-4). Lagged model available 2003-2011.

Table 6 - Secondary schools: All schools, all years.

		Treatment schools - London				Treatment schools - Black Country				Treatment schools - Greater Manchester			
		Model		+ controls		Model		+ controls		Model		+ controls	
		B	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*
	(Constant)	-.028		-.086		-.028		-.090		-.028		-.089	
Y_{t-1}	Previous year's GCSE score (standardised)	.870	***	.827	***	.872	***	.827	***	.872	***	.827	***
X_2	KTS school (all observations)	-.155	***	-.158	***	-.123	***	-.113	***	-.070	**	-.054	*
X_3	**KTS treatment (during treatment period)**	.147	***	.152	***	.128	*	.131	*	.075	†	.088	††
	IDACI ranking			✔				✔				✔	
	School type			✔				✔				✔	
	School change			✔				✔				✔	
	R^2=	.760		.765		.759		.765		.759		.764	
	N_{obs}=	26,810		26,810		26,290		26,290		26,430		26,430	
	$\rho_{autocorrelation}$ =	-0.250		-0.218		-0.256		-0.222		-0.254		-0.221	
	$d_{DurbinWatson}$ =	2.407		2.359		2.415		2.366		2.411		2.363	

† P = .109; †† P = .058;

Table 7 - Secondary schools: All schools, all years.

		Treatment schools & control schools - London				Treatment schools & control schools - Black Country				Treatment schools & control schools - Greater Manchester			
		Model		+ controls		Model		+ controls		Model		+ controls	
		B	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*
	(Constant)	.022		-.040		.005		-.234		-.010		-.092	
Y_{t-1}	Previous year's GCSE score (standardised)	.843	***	.809	***	.831	***	.738	***	.824	***	.761	***
X_2	KTS school (all observations)	-.233	***	-.230	***	-.195	***	-.218	***	-.135	***	-.130	***
X_3	**KTS treatment (during treatment period)**	.146	***	.150	***	.125	*	.124	*	.088	†	.109	*
	IDACI ranking			✔				✔				✔	
	IDACI ranking			✔				✔				✔	
	School type			✔				✔				✔	
	R^2=	.789		.792		.777		788		.730		.738	
	N_{obs}	3,550		3,550		780		780		1,510		1,510	
	$\rho_{autocorrelation}$ =	-0.214		-0.190		-0.242		-0.174		-0.236		-0.188	
	$d_{DurbinWatson}$ =	2.355		2.315		2.434		2.333		2.383		2.306	

† P = .069;

Table 8 - Primary schools: All schools, all years

Dependent variable: average (for English & Maths) of school-level percentage of pupils attaining level 4 at KS2 (standardised)

Year	2005	2006	2007	2008	2009	2010	2011
1 Standard Deviation (SD) increase in Dependent Variable =							
Percentage point Increase in level 4 attainment in KS2 **English**	13.8	13.6	13	12.2	12.8	12.5	12.8
Percentage point Increase in level 4 attainment in KS2 **Maths**	14.8	14.4	13.2	12.7	12.6	12.3	12.8

Table 9 - Primary schools: All schools, all years

	Dependent variable: school-level percentage of pupils attaining KS2(EM) level 4 (standardised)	***B***	***Sig.***
	(Constant)	.048	
Y_{t-1}	Previous year's KS2(EM) score (standardised)	.632	***

R^2 = 0.459; N_{obs} = 73,880 ; $N_{schools}$ = 12,324; $\rho_{autocorrelation}$ = -0.208; $d_{DurbinWatson}$ = 2.347

Table 10 - Primary schools: All schools, all years

	Dependent variable: school-level percentage of pupils attaining KS2(EM) level 4 (standardised)	Model		+ IDACI		++ type		+++ change	
		B	***Sig.***	***B***	***Sig.***	***B***	***Sig.***	***B***	***Sig.***
	(Constant)	.054		.059		.010		-.045	
Y_{t-1}	Previous year's KS2(EM) score (standardised)	.625	***	.561	***	.541	***	.540	***
X_2	KTS school (all observations)	-.440	***	-.365	***	-.354	***	-.354	***
X_3	**KTS treatment (during treatment period)**	.432	***	.428	***	.426	***	.426	***
	IDACI ranking			✔		✔		✔	
	Dummies for school type					✔		✔	
	Dummies for school change							✔	
	R^2=	*.462*		*.481*		*.487*		*.487*	
	N_{obs}=	*73,880*		*73,880*		*73,880*		*73,880*	
	$\rho_{autocorrelation}$ =	*- .204*		*- .160*		*- .147*		*- .147*	
	$d_{DurbinWatson}$ =	*2.341*		*2.263*		*2.244*		*2.244*	

$N_{schools}$ = 12,324;

Table 11 - Primary schools: Restricted samples, all years

Dependent variable: school-level percentage of pupils attaining KS2(EM) level 4 (standardised)		Deprivation		School type		School change	
		high*	non-high	CY	Other	No	Yes
		B	**B**	**B**	**B**	**B**	**B**
	(Constant)	-.110	.111	.006	.149	.056	.032
Y_{t-1}	Previous year's KS2(EM) score (standardised)	.589	.592	.629	.568	.623	.645
X_2	KTS school (all observations)	-.375	-.385	-.402	-.527	-.442	-.423
X_3	**KTS treatment (during treatment period)**	.479	.311	.437	.406	.426	.481
	R^2=	*.410*	*.398*	*.469*	*.374*	*.458*	*.490*
	N_{obs}=	*19,395*	*54,485*	*45,298*	*28,582*	*66,884*	*6,996*
	$\rho_{autocorrelation}$ =	*- .171*	*- .183*	*- .198*	*- .180*	*- .201*	*- .230*
	$d_{DurbinWatson}$ =	*2.274*	*2.306*	*2.332*	*2.303*	*2.336*	*2.387*

Notes: all coefficients significant at 1% level (excluding constants).
* below 25th percentile of IDACi rankings.

Table 12 - Primary schools: All schools, all years

		Dependent variable: school-level mean KS2 points score (standardised within-year)*				*Dependent variable: school-level KS2 value-added score** (standardised within-year)*			
		Model		+ controls		Model		+ controls	
		B	**Sig.**	**B**	**Sig.**	**B**	**Sig.**	**B**	**Sig.**
	(Constant)	.015		-.075		.009		-.004	
Y_{t-1}	Previous year's outcome score (standardised)	.716	***	.622	***	.551	***	.548	***
X_2	KTS school (all observations)	-.472	***	-.390	***	-.256	***	-.268	***
X_3	**KTS treatment (during treatment period)**	.443	***	.439	***	.500	***	.500	***
	IDACI ranking			✔				✔	
	Dummies for school type			✔				✔	
	Dummies for school change			✔				✔	
	R^2=	.528		.551		.308		.310	
	N_{obs}=	53,910		53,910		65,622		65,622	
	$\rho_{autocorrelation}$ =	-0.254		-0.191		-0.085		-0.086	
	$d_{DurbinWatson}$ =	2.381		2.286		2.128		2.130	

Table 13 - Primary schools: All schools, all years.

		Treatment schools - London				*Treatment schools - Black Country*				*Treatment schools - Greater Manchester*			
		Model		+ controls		Model		+ controls		Model		+ controls	
		B	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*
	(Constant)	.054		-.047		.054		-.048		.054		-.046	
Y_{t-1}	Previous year's KS2(EM) score (standardised)	.627	***	.542	***	.628	***	.543	***	.628	***	.543	***
X_2	KTS school (all observations)	-.425	***	-.317	***	-.547	***	-.481	***	-.406	***	-.353	***
X_3	**KTS treatment (during treatment period)**	.483	***	.481	***	.450	***	.458	***	.348	***	.330	***
	IDACI ranking			✔				✔				✔	
	School type			✔				✔				✔	
	School change			✔				✔				✔	
	R^2=	.457		.483		.457		.482		.456		.481	
	N_{obs}=	73,148		73,148		72,554		72,554		72,770		72,770	
	$\rho_{autocorrelation}$ =	-0.205		-0.147		-0.205		-0.147		-0.205		-0.147	
	$d_{DurbinWatson}$ =	2.342		2.244		2.341		2.244		2.341		2.244	

* mean points score available 2001-2011. Lagged model available 2002-2011.
** value-added figures available 2002-2003 (Key stage 3-4) and 2004-2011 (key stage 2-4). Lagged model available 2003-2011.

Table 14 - Primary schools: All schools, all years.

		Treatment schools & control schools - London				*Treatment schools & control schools - Black Country*				*Treatment schools & control schools - Greater Manchester*			
		Model		+ controls		Model		+ controls		Model		+ controls	
		B	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*	*B*	*Sig*
	(Constant)	.107		.119		.070		-.017		.118		.059	
Y_{t-1}	Previous year's KS2(EM) score (standardised)	.596	***	.524	***	.572	***	.467	***	.558	***	.456	***
X_2	KTS school (all observations)	-.508	***	-.486	***	-.642	***	-.644	***	-.536	***	-.507	***
X_3	**KTS treatment (during treatment period)**	.479	***	.479	***	.450	***	.455	***	.337	***	.318	***
	IDACI ranking			✔				✔				✔	
	School type			✔				✔				✔	
	School change			✔				✔				✔	
	R^2=	.487		.509		.518		548		.439		.473	
	N_{obs}=	9,276		9,276		2,097		2,097		4,792		4,792	
	$\rho_{autocorrelation}$ =	-0.202		-0.153		-0.183		-0.117		-0.177		-0.114	
	$d_{DurbinWatson}$ =	2.327		2.249		2.314		2.20		2.309		2.202	

* mean points score available 2001-2011. Lagged model available 2002-2011.
** value-added figures available 2002-2003 (Key stage 3-4) and 2004-2011 (key stage 2-4). Lagged model available 2003-2011.

Figures

Figure 1 - Proportion attaining 5+ GCSEs, by year, by KTS status

Secondary schools: All schools, all years

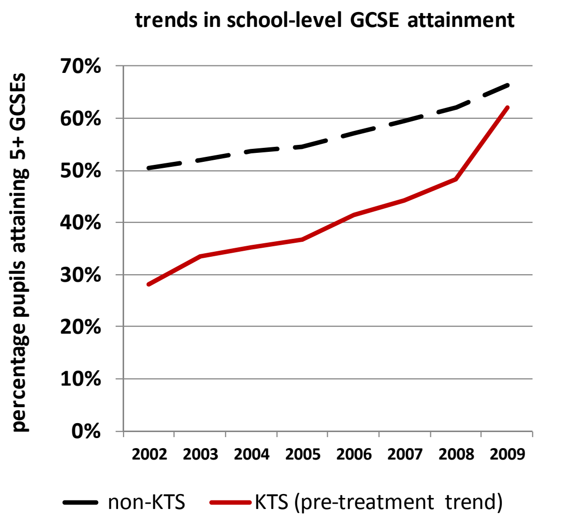


Figure 2 - Example matching issues. Matching on trajectory between years T-1 & T-3.

Primary schools: Treatment starting in 2009

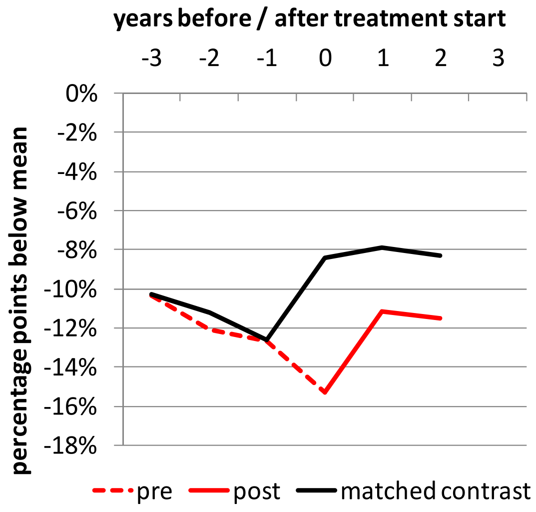


Figure 3 - Propensity score distribution by matching success.

Secondary schools: All KTS schools, all years

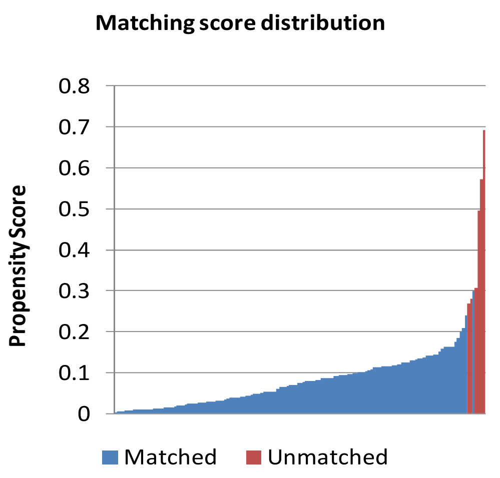


Figure 4 - Propensity score matching comparing trajectories for schools with treatment commencing in 2003 and matched schools

Secondary schools: All schools, all years

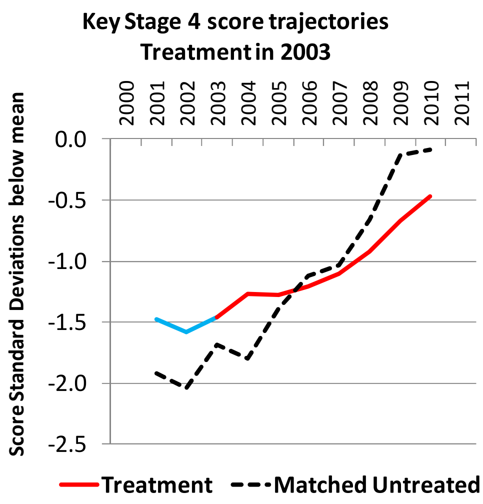


Figure 5 - Predicted proportion (standardised) attaining 5+ GCSEs, by preceding year's result.

Secondary schools: All schools, all years

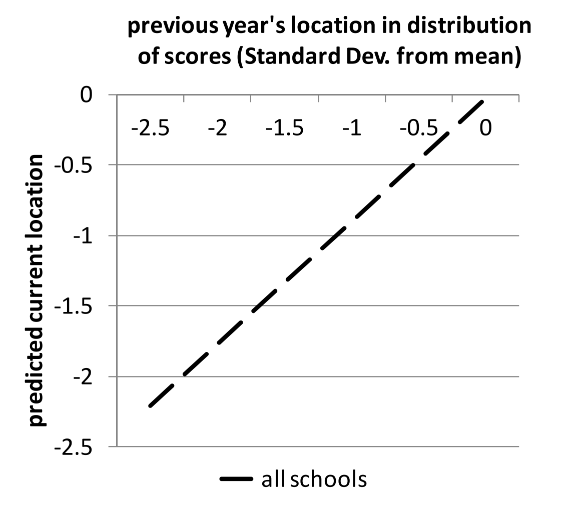


Figure 6 - As Figure 2 with KTS group differential (blue line) and net effect of treatment (red line).

Secondary schools: All schools, all years

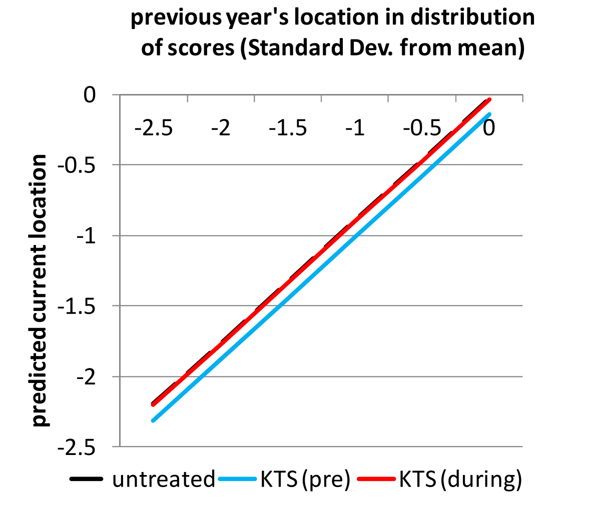


Figure 7 - As Figure 3 with relaxation of parallel and linear assumptions

Secondary schools: All schools, all years

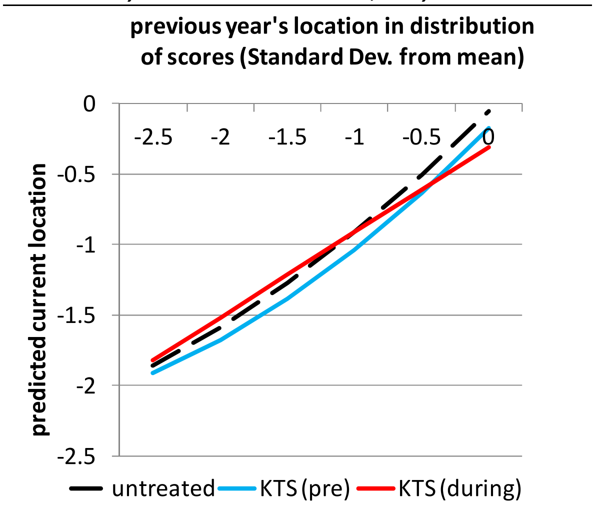
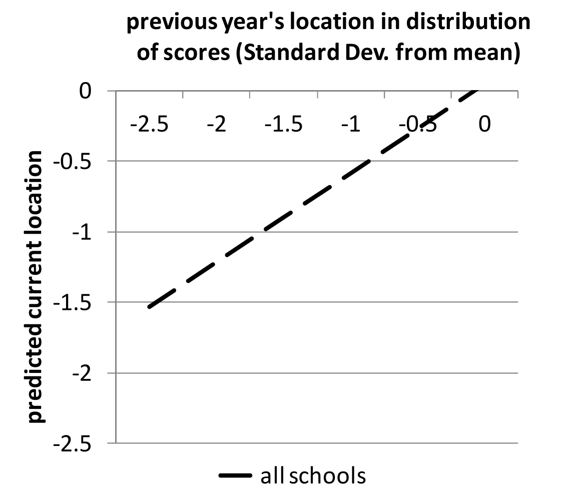


Figure 8 - Predicted proportion (standardised) attaining Level 4 at KS2 (average of English &

Figure 9 - As Figure 5 with KTS group differential (blue line) and nett effect of

Maths), by preceding year's result.

Primary schools: All schools, all years



treatment (red line).

Primary schools: All schools, all years

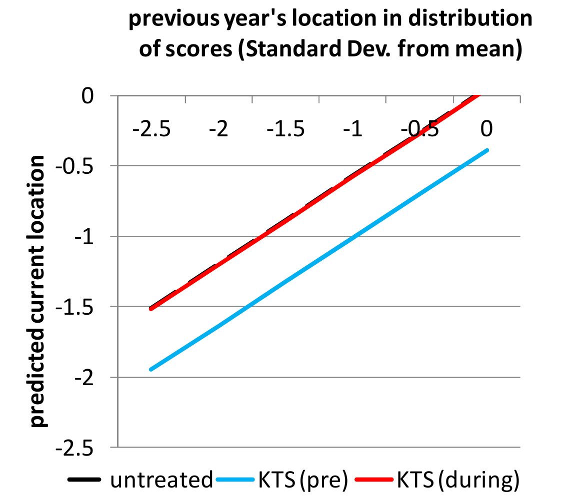
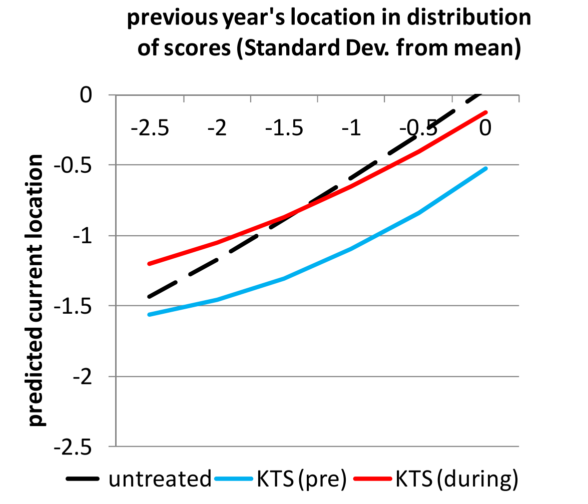


Figure 10 - As Figure 6 with relaxation of parallel and linear assumptions

Primary schools: All schools, all years



References

Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*. The American Economic Review, 84(4)*, 772-793.

Davidson, R. & MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. New York, Oxford University Press.

Heckman, J., Ichimura, H. & Todd, P. E. (1997) 'Matching as an Econometric Evaluation Estimator', *Review of Economic Studies, 65,* 261-294.

Keele, L. & Kelly, N. J., (2006). Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables. *Oxford Journal of Political Analysis, 14 (2),* 186-205. doi: 10.1093/pan/mpj006. First published online: November 23, 2005

Machin, S., McNally, S. & Meghir, C. (2004), 'Improving Pupil Performance in English Secondary Schools: Excellence in Cities', Journal of the European Economics Association, 2, 396-405.

Machin, S. & Vernoit, J. (2011). Changing School Autonomy: Academy Schools and their Introduction to England's Education. *CEE Discussion Paper CEEDP0123.* Centre for the Economics of Education, London School of Economics. London

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70 (1):* 41–55

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Wooldridge, Jeffrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press Books, The MIT Press.

Ref: DFE-RR215a

June 2012