UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Southampton Education School

Volume 1 of 1

Utilising a school effectiveness approach to measuring non-cognitive outcomes for the Social and Emotional Aspects of Learning (SEAL) programme

by

Christopher John Downey

Thesis for the degree of Doctor of Philosophy

September 2013

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

Education

Thesis for the degree of Doctor of Philosophy

Utilising a school effectiveness approach to measuring non-cognitive outcomes for the Social and Emotional Aspects of Learning (SEAL) programme

Christopher John Downey

This study applies a school effectiveness approach to determining the nature and the extent of effects of non-cognitive outcomes of schooling, specifically in relation to the SEAL programme. The focus is on the potential to evaluate gains made through universal SEAL provision, as well as via intensive Family SEAL intervention for groups of children and their parents/carers. The research was undertaken within primary and secondary schools located in one English Local Authority region.

An extensive literature review of measures of student and school level academic progress in the cognitive domain shows how such measures have built on the insights of school effectiveness research and this leads to a consideration of studies of similar design focusing on non-cognitive outcomes of schooling, and the size of school and class level effects related to them.

Data from around 2000 students was used to determine a measurement model for a survey of SEAL related outcomes, and across a period of approximately 18 months, over 8500 students from 55 schools were eventually surveyed. Multilevel modelling of data on 7 non-cognitive dimensions derived from the survey was used to determine the proportion of the variance located at the student, class and school levels for each of the dimensions. Associations between each dimension and a range of student and school level factors were also investigated.

Significant associations were found to occur between non-cognitive outcomes and factors such as students' age, gender and whether they had experience with bullying behaviours. There was some limited evidence of associations for peer-effects of bullying and the mean socioeconomic status for some of the non-cognitive dimensions. Class and especially school level effects were found to be appreciably smaller than those observed for cognitive outcomes of schooling which was broadly in line with the findings of previous research in this area.

A small scale evaluation of the impact of Family SEAL provided tentative evidence that gains from engaging in this intervention are most likely to be made by students causing concern in their social and emotional development, and that these gains are most likely to be observed at school rather than at home.

The limitations of each element of the study were considered and taken into account in making a number of recommendations for practice in schools and local and national level policy making.

Contents

List of tables

List of figures

DECLARATION OF AUTHORSHIP

I, Christopher John Downey

declare that the thesis entitled

Utilising a school effectiveness approach to measuring non-cognitive outcomes for the Social and Emotional Aspects of Learning (SEAL) programme

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission, or [delete as appropriate] parts of this work have been published as: [please list references]

Signed:...

Date:..

Acknowledgements

I would like to express my sincere gratitude first and foremost to Tony Kelly in his role as supervisor and colleague. I appreciate that I have stretched his patience far too often in the first of these capacities, hopefully much less so in the latter. Thank you, Tony, for the balance of insight, support and challenge that you always provided as supervisor. I regret that I didn’t avail myself of as much advice as you were willing to provide, especially during the stages of writing up.

My gratitude also goes to the colleagues working for the school support service of the local authority in which this research was based. Especially to those who worked as National Strategies Behaviour and Attendance consultants during the period of data collection; for their encouragement, their support and their interest in the study.

I am grateful to many colleagues in the Southampton Education School for their ongoing support and interest in the progress of my thesis (or lack thereof), and especially to those who have encouraged me to “get the thing done”.

I have very much appreciated the opportunity to get to know and become part of the ICSEI community of researchers, practitioners and policy makers and the insights this has afforded into the allied fields of school effectiveness and school improvement.

Finally I am deeply grateful to my lovely wife and wonderful children for their unwavering patience and support during the course of the PhD process. When I started the children were young and naive enough to think that if their Dad ever became a doctor he could treat all their ailments. As the years have progressed the naivety has gone and been replaced by scepticism about whether the PhD would ever see completion. I am very sorry that I have not set a good example of an effective postgraduate research student. If, in spite of Dad’s experience, you elect to become one yourself, remember…

“Don’t do as I do…”

Chapter 1: Introduction

This chapter provides background to the development and nature of the Social and Emotional Aspects of Learning (SEAL) programme and the nature of SEAL, both as a universal curriculum and as a targeted intervention to develop the social and emotional competence of students. It sets out the challenge faced by schools to evaluate the impact of their implementation of SEAL, to evidence that impact to other stakeholders, and, in the absence of SEAL specific measures of student progress, illustrates the tendency to turn to more distal measures of cognitive outcomes for such evidence of impact. It then makes the case for a different approach to evidencing the impact of SEAL and sets out the work to be considered in the remainder of the thesis.

1.1 A brief introduction to SEAL and the context of this study

> *Social, emotional and behavioural skills underlie almost every aspect of school, home and community life, including effective learning and getting on with other people. They are fundamental to school improvement.* (DfES 2005: 7)

This is the broad and bold claim made in the guidance for the Primary Social and Emotional Aspects of Learning (SEAL) programme published by the then Department for Children, Schools and Families (DCSF), now the Department for Education (DfE).

The Primary SEAL programme (DfES 2005) was developed as part of the Primary Behaviour and Attendance Strategy Pilot that was conducted during 2003-2005. The then Department for Education and Skills (DfES) commissioned a report on the pilot that was conducted by researchers from the Institute of Education (Hallam *et al.* 2006). The background rationale for the programme is well described in a DfES report (Weare & Gray 2003). In essence the SEAL programme focuses on development of *social and emotional skills* of students, and other members of the school community, linked to five learning related *aspects of learning*, namely, self-awareness, managing feelings, empathy, motivation and social skills. These five aspects are derived from the work of Goleman on emotional intelligence (Goleman 1996) who in turn draws to a limited extent on the theoretical perspectives of a select set of researchers in the field of emotional competence (Gardner 1983; Salovey & Mayer 1990). Thus the focus is on developing affect, specifically social and emotional competence of young people, with a view to improving the learning environment. The Secondary SEAL Programme (DfES 2007) was launched nationally from September 2007 after completion of a two year pilot in a number of authorities from 2005. The SEAL programme was not a statutory element of the Primary and Secondary National Strategy, and so schools subscribed to the programme voluntarily.

The research that makes up the study that is the focus of this thesis took place in one local authority region in England. Many of the schools in the region are located in urban centres but much of region is semi-rural or rural. The region is not very diverse in terms of the proportion of students with English as an additional language, and mean levels of deprivation in the region are well below national means based on the proportion of students eligible for free school meals. In terms of school performance mean academic attainment at the end of primary and secondary schooling was consistently above national means, based on headline attainment statistics reported in national school performance tables during the period leading up to the start of the study (2003-2006), and this was maintained through to the period of data collection (2006-2008) and after, up to 2010 by which time the relative rise in performance by national means has closed this attainment gap.

Notwithstanding the non-statutory status of SEAL, the programme was eventually implemented around 70% of the Primary Schools in the region. A small group of schools, 7 in total, made up the first tranche to take up the programme in 2004, with two further tranches of schools taking up the programme in 2005 (20 schools) and 2006 (46 schools). A further 42 schools came on stream from 2007. There were initial plans for just over half the secondary in the region to take up the Secondary SEAL Programme from September 2007, together with all the designated special schools across the region, as well as each of the centres offering educational provisions for children with behavioural issues that led to extended periods of fixed-term exclusion from school.

1.2 Evaluating the impact of SEAL: a complex challenge requiring a considered approach

Evaluating a programme such as SEAL presents an enormous challenge due to the sheer scope and breadth of its aims, rationale and modus operandi. One of the core outcomes that Local Authorities are looking for is the impact that SEAL has on the capacity for learning in schools and the metrics of the moment focused particularly on the impact on cognitive outcomes of schooling in the form of academic performance in national tests and examinations. The Government ministry for Education (DCSF) was not averse to this, indeed, on its standards website it presented details of a SEAL case study school, Birchwood Junior School in Lincoln (DCSF 2007a). The case study focused on what was known as the intensifying support program (ISP) as well as on SEAL, and it is clear that SEAL formed a central part of the school's efforts to raise pupil attainment,

> *Central to the success of the programme, and ISP, was the introduction of the DfES materials for developing social, emotional and behavioural skills - the SEAL materials.*

(op cit.)

The background to the school states:

> *Birchwood Junior School is a three-form entry school in an area of high social deprivation. In September 2004 the school took part in a one year, pre-ISP programme. This helped the school to establish rigorous whole-school improvement systems to raise standards. Central to the success of the programme, and ISP, was the introduction of the DfES materials for developing social, emotional and behavioural skills - the SEAL materials.*

A key element of the case study lies in evidence of impact. Under a section headed "What has been the impact of SEAL within ISP?" the school reports, "SEAL has been central to the success of ISP in our school. The trend in data is upwards." and data for cognitive outcomes are given (Figure 1, the data which was presented in the form of a table, has been converted here to a line graph).

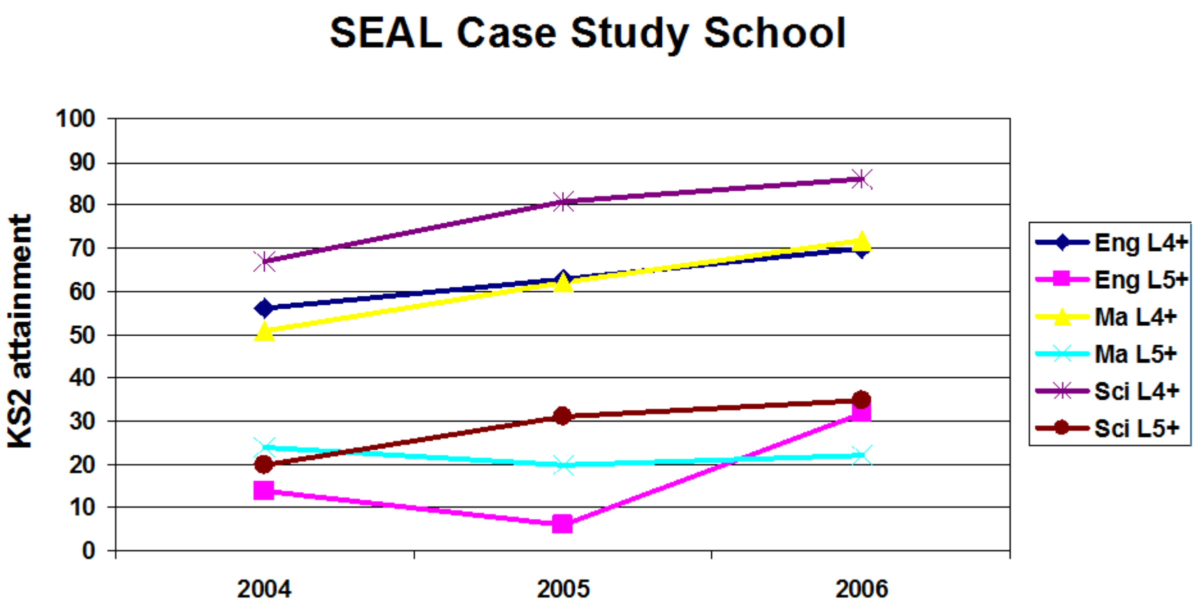


Figure 1: KS2 attainment data for SEAL case study school 2004-2006

Source: (DCSF 2007a)

The graph shows that standards were generally rising steadily during the period in which SEAL was implemented in the school, especially in terms of the percentage of pupils attaining the national expected standard for KS2 of level 4 and above. The improvement in the percentage KS2 passes at level 5+ (well above the national expectation) is not so marked although the three year trend for science and possibly also English shows an increase and the percentage attaining level 5+ in maths (the highest of the three core subjects in 2004) has remained broadly stable. Overall the picture is clearly positive and the implication was that SEAL was largely responsible for this picture of improvement.

The 'Raising Attainment Plan' (RAP – often referred to as the School Development Plan or School Improvement Plan) for the school was also included as an attachment to the case study webpage (DCSF 2007b). A cursory glance at the RAP that was drawn up prior to the implantation of SEAL makes it abundantly clear that a veritable raft of school improvement initiatives were planned for implementation in the school during the same period as the implementation of SEAL. Such initiatives include:

- in depth analysis of data from the pupil attainment tracker (PAT – the forerunner to RAISEonline) used to inform curriculum planning with a three week review cycle by subject leaders,

- tracking of pupil progress against targets for all pupils combine with more intensive monitoring and support for targeted groups such as those with SEN and those identified as Gifted and Talented students,
- the implementation of Assessment for Learning practices (AfL) into teaching which has a key focus on formative, peer and self-assessment,
- the development of teaching to accommodate varied learning styles using the VAK (visual, auditory and kinaesthetic) approaches,
- peer coaching by teachers
- enhanced home school relationships.

(*op cit*)

In the school's RAP all of these interventions, along with those drawn from the SEAL programme, were linked to expected improvements in academic outcomes of the students. How then can one make a robust claim that SEAL has had a significant impact on the outcomes of students when schools are clearly highly complex environments simultaneously engaging in a plethora of school improvement projects. What is the unique contribution that SEAL is making? Is it possible to make such claims?

No data had been provided in the case study to give a sense of the pre-SEAL baseline academic outcomes. Also, at the time of publication, the latest KS2 outcomes (representing learning across the ages of 7-11 years,) for assessments in spring/summer 2007 (DCSF, 2008), had not been added to the case study as an update. These data paint a different picture of the attainment of students in the school (Figure 2). There was a downturn in the percentages attaining level 4+ and level 5+ in all subjects except maths at level 5+. The level 4+ percentages were still higher than the figure in 2004 but have dropped below the pre-SEAL baseline attainment in 2003. For attainment at level 5+ the percentages had returned to their approximate pre-SEAL baseline values.

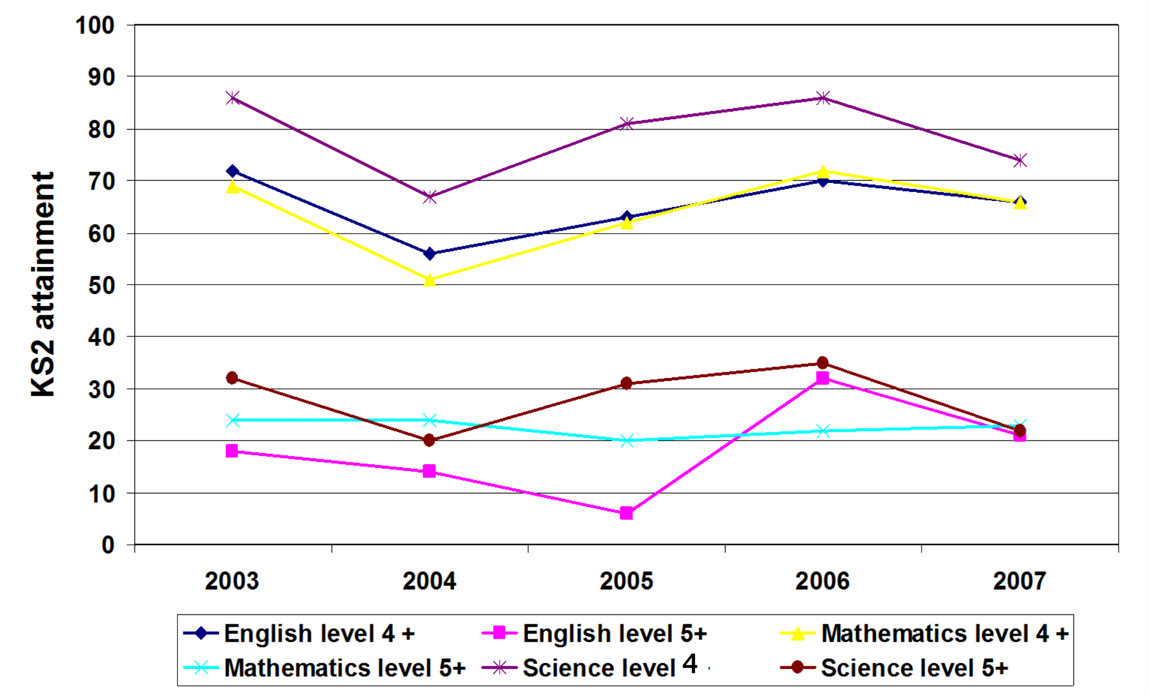


Figure 2: KS2 attainment data for SEAL case study school 2003-2007

Source: (DCSF 2007a)

The gradients for the trendlines (generated by ordinary least squares) for each core subject for the periods 2004-06 and 2004-07 (i.e. since the start of SEAL) are as follows:

Subject	**04-06 gradient**	**04-07 gradient**	**Subject**	**04-06 gradient**	**04-07 gradient**
English level 4+	9.5% yr^{-1}	3.7% yr^{-1}	English level 5+	9.0% yr^{-1}	4.7% yr^{-1}
Maths level 4+	10.5% yr^{-1}	5.5% yr^{-1}	Maths level 5+	-1.0% yr^{-1}	-0.1% yr^{-1}
Science level 4+	7.0% yr^{-1}	2.6% yr^{-1}	Science level 5+	7.5% yr^{-1}	1.0% yr^{-1}

Table 1: Trends in KS2 attainment data for SEAL case study school

Such data would, at first glance, suggest that the gains accrued during the 2-3 year implementation of SEAL had not been easy to sustain but, in the same way that one would find it hard to argue robustly that SEAL was predominantly responsible for the gains observed from 2004-2006 it would also be just as simplistic an analysis to equate the fall in standards with SEAL based on the same reasoning, namely, that a raft of other school improvement initiatives were planned for implementation over this time.

The Lincoln Birchwood school case study makes use only of the raw cognitive outcomes of schooling, unadjusted for prior attainment. As will be discussed in this thesis, more sophisticated measures of cognitive *progress*, adjusting for the prior attainment of students, had been developed from the work of early school effectiveness researchers who adopted a more sophisticated approach to evaluating the effects of schooling. At the time SEAL was implemented these value-added methodologies, as they came to be known, were being widely used and reported. It seems at odd therefore that a more simplistic approach would be adopted to publishing evidence of the impact of a programme like SEAL.

This extended quote from Mortimore (1998) an early influential school effectiveness researcher from England puts it eloquently:

> *Studies of variations between schools exist in both simple and more sophisticated forms. The simpler studies take little of no account of differences in the characteristics of students entering or attending the schools. They also tend to focus on only one outcome measure student scholastic achievement. The difficulties of the simple approach, as experienced teachers will recognise, is that schools do not receive uniform intakes of students. Some take high proportions of relatively advantaged students likely to do well in examinations; others (on the whole) receive high proportions of disadvantage students who, all things considered, are less likely to do well. To compare the results of scholastic achievement tests or examinations without taking into account these differences in the students when they enter the school, and attribute good results to the influence of the school may, therefore, be quite misleading.*
>
> (ibid., 114-115).

The published value-added scores (Table 2) for the school, measuring the progress of students between national assessments at Key Stage 1 (aged 6-7) and Key Stage 2 (aged 10-11), account for prior-attainment and provide a mean measure of the progress of all children in each cohort rather than simply the percentage crossing a certain threshold. This measure provides a different picture of the potential impact of SEAL alongside the other school improvement initiatives.

Year	**2003**	**2004**	**2005**	**2006**
Published VA score	98.7	97.7	98.4	98.5

Table 2: Value-added scores for SEAL case study school

Thus the observable gains in cognitive outcomes, in the form of raw attainment, made between 2005 and 2006, which were similar to the gains in attainment made between 2004 and 2005, did not reflect a corresponding rise in the value-added measure of progress as the cohort. This suggests at least some of the gains were as a result of the higher prior attainment of students on entry from other first schools. Interestingly the published VA score for 2003 of 98.7 was higher than any of the years that form part of the case study period. The attainment measures for 2003 show that percentages of students attaining level 4+ and level 5+ in the three core subjects were almost identical (+/- 3%) to the 'peak' results of 2006 with the exception of English at level 5+ which was some 14% lower than the percentage attaining level 5+ in 2006.

Contextual value-added (CVA) scores were published for all primary schools with students in Year 6 (10-11 years) for the first time in 2007. As will be discussed in more detail, CVA takes into account a range of background factors at both the level of the student and the school. These additional factors include measures of socio economic status and the formal Special Educational needs stage of children in the school. These were two key context factors highlighted in the case study background showing the school admits a higher proportion of children with higher levels of special education need and also from deprived neighbourhoods. The 2007 CVA score for Lincoln Birchwood Junior School was 98.3, with a 95% confidence interval of 97.9 to 98.8 signifying a CVA score significantly below the national mean of 100.0. In 2008 the CVA score was 98.1, and in 2009 it has risen to 99.2, but both also significantly below the national mean. The advantage of adjusting for prior attainment and other contextualising variables that may be out of the school's control such as gender, ethnicity, and socioeconomic status is that the school effect accounts for a larger proportion of the unexplained variance. If then even complex value added measures which seek to bring us closer to the *effects* that the school adds, fail to provide consistent evidence for the impact of SEAL on academic outcomes, does this mean that SEAL isn't having impact at the school? Even sophisticated highly contextualised measures of academic progress such as CVA are summary measures, drawing together the various strands of effectiveness operating in the school to present a measure of the overall school effect. They are, by nature restricted to a narrow 'input-output' model biased entirely towards the cognitive outcomes of schooling, or as Mortimore (1998) refers to it, 'scholastic achievement'.

Despite the mixed picture of the effectiveness of the case study school painted by the attainment and value-added progress measures discussed here, a later Ofsted report for the school (Ofsted 2008) suggested SEAL is indeed having an impact of the school:

> *…the social and emotional aspects of learning programme is having a positive impact on learning and behaviour.* (*ibid*: 6)

Close reading of the report reveals some other significant school level changes. At the time of the 2008 inspection the school had just appointed a new Headteacher and had experienced particularly high turnover of teachers with up to 50% of the staff changing over the few months prior to inspection, and replaced mainly by newly qualified teachers which the report suggests had placed additional burdens on the new school leadership.

Further investigation of the attainment profile of the school in the period 2000-2003 showed a 3 year rise in attainment, based on aggregated percentages of students attaining level 4+ in the three core subjects (DCSF 2004), peaking in 2003. The 2003 Ofsted inspection report revealed that during the period of inspection the Deputy Head had been acting as Headteacher, awaiting the arrival of a new Headtacher in Easter of that year (Ofsted 2003). While the report praised the work of the Acting Head and Acting Deputy, especially in appointing and inducting 5 new members of staff that year, the inspectors identified a lack of long term development planning (*ibid*: 21). This account provides striking parallels with the 2008 inspection report discussed above. One might therefore suggest that the ‘roller coaster’ like picture of pupil attainment at the school from 2000-2007 (a rise from 2000-2003 then a fall in 2004 followed by a rise through to 2006 which was then followed by another fall in 2007) reflects cycles of uncertainty and change in leadership, management and staffing over and above the impact of broad curriculum initiatives such as SEAL, or many of the other elements of the school’s Raising Attainment Plan.

Clearly school level attainment and even value-added progress measures only have a partial role to play in telling the story of SEAL. This is reflected in the fact that other evidence that was included with the case study (DCSF 2007a). Some of this evidence is quantitative in nature:

- Improved attendance: 92.7% in 2004, 94.5% in 2005 (further investigation revealed that the overall percentage attendance for 2005 was 94.0% and for 2007 was 94.1%)
- Fixed term exclusions down 50% with no permanent exclusions in 2005
- Reduction in the number of serious whole-school incidents recorded.

There is also a range of qualitative measures of the impact of the SEAL programme:

- Monitoring shows that children are much more able to sustain independent learning

- Improvements self-esteem, resilience, understanding of others' points of view and self-control
- Whole-school language established for children and adults to talk about emotions and behaviour

Unfortunately, no further evidence is provided for these key claims. Ofsted reports are unlikely to provide an independent source of such qualitative evidence of the impact of SEAL as, since inspection reports have become much shorter with fewer lessons being inspected (contrast 44 pages with only 11 for the two reports of the Lincoln Birchwood Junior School in 2003 and 2008 respectively).

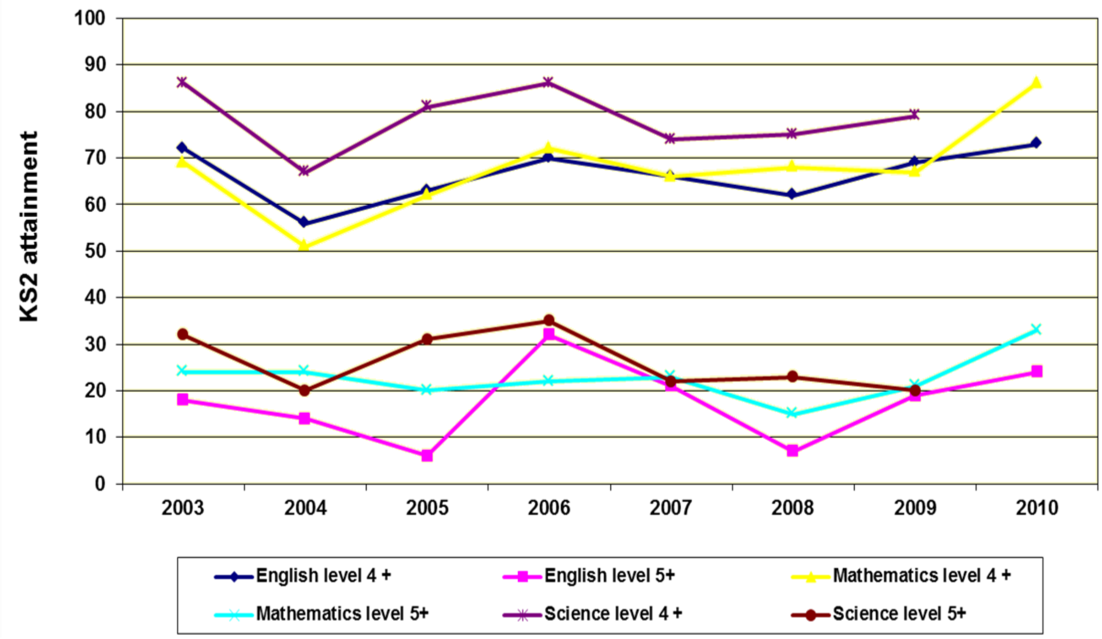


Figure 3: Extended KS2 attainment data for SEAL case study school

Figure 3 above shows that since 2008, the performance of the school in terms of student outcomes has improved. The Headteacher appointed just before the 2008 inspection was still in post by the date of the 2011 inspection (Ofsted, 2011) and the same Chair of Governors was also in post at the time of the 2011 inspection. In this report the school was described as "rapidly improving" with many of the issues raised in the previous inspection having been well addressed. The report makes clear that stability in key leadership and governance positions, together with the distribution of leadership responsibility among of a more stable teaching staff had certainly contributed to the improving trend between 2008 and 2010. This notwithstanding, despite the brevity of the 2011 report, there are indications that the SEAL heritage may have been playing a key role in the integration and care of students in the school.

Though SEAL is not mentioned explicitly, the following extensive quotation from the report indicates that a number of the key areas of Primary SEAL (that will be discussed in more detail in the next chapter) have contributed to the improved quality of provision for students:

> *Good care, guidance and support are a strength of the school. Pupils are highly valued as individuals, and the school's emphasis on raising their self esteem and self belief allows pupils to thrive. Every adult in the school regularly takes a 'chatter group' in which they converse with 10-12 pupils on matters of shared concern and interest. Having their views taken seriously in this way makes the pupils feel valued and increases their confidence. Teachers ensure that children settle in well when they first join the school, and older pupils told inspectors how well they feel the school prepares them to move on to secondary school. The learning support centre makes an outstanding contribution to the care and support of pupils whose circumstances put them at risk of failure. In its calm and purposeful environment expectations for work and behaviour are made very clear. It has had a marked impact on behaviour across the whole school.*
>
> (*ibid*: 7)

1.3 Determining the impact of SEAL: rising to the challenge in a different way

One of the issues apparent in discussion of the case study above is that the expectation that developing non-cognitive or affective skills will have a direct impact on the cognitive outcomes of schooling, and will do so in a relatively short term. Although the aims of SEAL state that the programme is focused on improving the overall learning environment in schools the case study illustrates that it would be one of a number of initiatives and policy factors at work at the school which could influence cognitive outcomes. The field of School Effectiveness Research (SER) has shown over recent decades (Sammons, 2007) that the factors which are associated with highly effective schools are many and the way their effects are mediated can be both varied and complex. The temptation to turn straight to the cognitive outcomes data for evidence of impact of a programme like SEAL is strong, and also understandable, as the students' results in national tests and assessments are widely seen as one of the principal outcomes of schooling. The current school accountability agenda, both in terms of published school league tables and Ofsted inspection, place high stakes for schools for such outcomes.

Humphrey et al (2010), in a national scale evaluation of the impact of Secondary SEAL commissioned by the DCSF, refer to what they term the *proximal* and *distal* outcomes of an initiative such as SEAL. They list the development of social and emotional skills, what we might term non-cognitive (Gray, 2004) outcomes of SEAL, as *proximal* benefits, while changes in attendance (another non-cognitive outcome) and attainment (in the cognitive domain) are *distal* benefits. This study seeks to determine whether there is scope to employ the measurement methodologies developed for the cognitive outcomes of schooling by researchers working in the school effectiveness research field, to these more proximal benefits of an initiative like SEAL that lay in the domain of non-cognitive outcomes of schooling. It will draw on the previous but limited body of work on non-cognitive outcomes conducted within a school effectiveness framework and apply the insights gleaned there to developing novel metrics of SEAL related, non-cognitive outcomes. The emphasis will be on utilising pre-existing measurement instruments which already command have high face validity with teachers, but applying those within new analytical frameworks with the aim of developing a robust approach to measuring the non-cognitive outcomes of SEAL. Using the methodology of SER, the research will seek to determine whether there is potentially a measurable school effect that exists for non-cognitive outcomes related to SEAL, based on data collected from students across the whole range of ages, phases of schooling and backgrounds represented in the participating region. The research will also consider how traditional quasi-experimental approaches might be utilised to evaluate the impact of short-term, intensive, SEAL-based interventions with targeted groups of students considered to have the greatest level of need to develop SEAL related skills.

This thesis will begin with a chapter giving critical overview of the SEAL programme, in terms of its development and the distinctive nature of the programme, and how it is applied in schools. It will consider the universal approach to SEAL, intended to develop the social and emotional competence of all students (and others in the school community), as well as a more targeted initiative, known as Family SEAL, which aims to involve parents in the development of these skills.

This will be followed by a chapter reviewing of the application of methods within school effectiveness research to the development of school effectiveness measures in the English school system. It will chart some of the advances in the measurement and modelling of a wide range of factors that help to explain the variation in the cognitive outcomes of schooling, and also to identify the size and nature of school level effects on these outcomes. It will also consider how this development has led to measures that are published to help inform schools in their self-evaluation and school improvement planning, and, more controversially, to hold schools publicly to account for their performance.

The next chapter will consider how the school effectiveness research framework has been used to study the non-cognitive outcomes of schooling. It will compare and contrast the challenges experienced and insights gleaned by researchers in determining the size and nature of the school effect for non-cognitive outcomes and compare these with the findings described in the previous chapter.

The methodology of the research study undertaken for this PhD will then be described followed by a presentation of the results and a critical discussion of the findings, before drawing conclusions and making recommendations for practice.

The results of this study will be of interest to a wide audience including practitioners working in schools and local authority school support services, especially those seeking to develop and evaluate the impact of programmes aimed at improving a wider set of outcomes, beyond the purely academic outcomes which tend to be pre-eminent in our understanding of what it means to be an effective school (Gray, 2004). It will also be of interest to researchers, especially those in the allied fields of school effectiveness and school improvement, for its contribution to our understanding of an under researched set of outcomes schooling, and the potential for identifying the presence of school effects for such outcomes.

Chapter 2: The Social and Emotional Aspects of Learning (SEAL) Programme

Chapter 2 provides a detailed overview of the nature and aims of SEAL, both as a universal element of curriculum provision for all children, and as a targeted intervention for groups of students, especially in the form of Family SEAL. It compares the structure of the initial version of SEAL, developed for use in primary schools, with the version that was eventually developed for secondary schools. The chapter concludes by considering the rationale for SEAL in more critical detail including some of the voices that have expressed concern with the implementation of SEAL as a universal curriculum initiative.

2.1 The nature and structure of universal SEAL

The United Nations Children's Fund (UNICEF) produced a report on the well-being of children in rich countries (UNICEF 2007) which sought to provide a measure for well-being across a range of categories including:

- Material well-being
- Health and safety
- Educational well-being
- Family and peer relationships
- Behaviours and risks
- Subjective well-being

Of the 21 OECD countries surveyed the United Kingdom was ranked lowest overall in terms of its average ranking based on a simple mean the nations' rankings for each of the six dimensions above. The UK ranked in the bottom five nations for all except *health and safety* (12th) and was ranked lowest for *family and peer relationships* and *behaviours and risks*. Each dimension was scored using a compound measure based on a diverse range of indicators (40 in total), with each indicator being equally weighted in its associated compound measure. The report recognises the simplistic nature of this approach and describes itself as "a significant advance" (ibid: 2) on previous attempts by UNICEF to measure child well-being but also "a work in progress, in need of improved definitions and better data" (ibid: 39). A similar picture is painted by an index of child well-being for 25 nations of the European Union (Bradshaw *et al.* 2007) in which England is ranked 21st above the Slovak Republic and the three Baltic coast nations of Latvia, Estonia and Lithuania.

It is into this background that the Primary and Secondary SEAL programmes were pitched. The most important source documents are the primary guidance booklet (DfES 2005) and it's secondary equivalent (DfES 2007a).

> *The resource aims to provide schools and settings with an explicit, structured whole-curriculum framework for developing all children's social, emotional and behavioural skills.* (DfES 2005: 5)

The materials divide social and emotional skills into five aspects based on five categorisations proposed by Goleman (1996). Goleman's 5-part model of emotional intelligence (Figure 4) might be considered to be widely known, through the sales of his popular books on the subject, but his approach has nonetheless received criticism from both advocates and opponents of the concept. The choice of Goleman's model to form the foundation of SEAL flows, at least in part, from the findings of a DfES commissioned review into effective practice in developing social and emotional competence (Weare & Gray 2003).

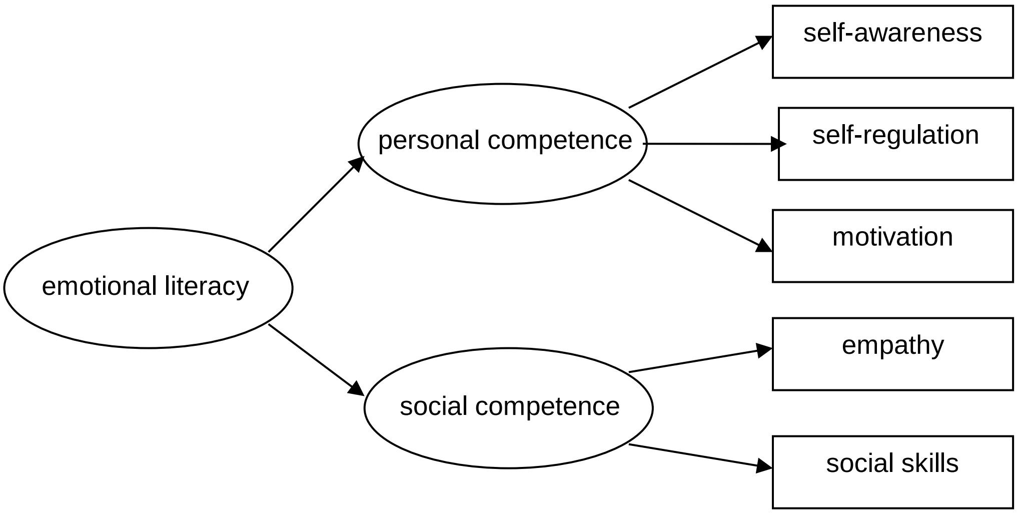


Figure 4: The five dimensional model of emotional intelligence (Faupel 2003)

The initiative was managed under the Behaviour and Attendance (B&A) strand of both the primary and secondary National Strategies. There was therefore a clear link made between the social and emotion and skills of students and their behaviour. What SEAL does is to take this firmly into the domain of the classroom and suggest that the development of such skills can have a direct influence on the learning of students and that by improving the social and emotional competence of children improved learning will follow.

> *Secondary SEAL is a comprehensive approach to promoting the social and emotional skills that underpin effective learning, positive behaviour, regular attendance, staff effectiveness and the emotional health and well-being of all who learn and work in schools. It proposes that the skills will be most effectively developed by pupils, and at the same time enhance the skills of staff, through:*
>
> - *using a whole-school approach to create the climate and conditions that implicitly promote the skills and allow these to be practised and consolidated;*
> - *direct and focused learning opportunities for whole classes (during tutor time, across the curriculum and outside formal lessons) and as part of focus group work;*
> - *using learning and teaching approaches that support pupils to learn social and emotional skills and consolidate those already learnt;*
> - *continuing professional development for the whole staff of a school.*
>
> (DfES 2007: 4)

The secondary guidance makes it clear that SEAL goes beyond developing the social and emotional skills of *students*, to consider the development of the skills of *staff* working at the school. Both primary and secondary SEAL resources contain a set of staff development activities and a perusal of the range of materials shows that these are designed not just for members of the teaching staff but for all staff in the school. Combine this with other materials which focus on involving the parents and carers of the children at the school and it soon becomes clear that SEAL is incredibly broad in its scope, seeking to involve the whole school community and beyond.

The structure of Primary and Secondary SEAL

Primary SEAL has been made concrete for schools by linking the broader SEAL objectives to a series of seven themes that are repeated each year providing and explicit spiral curriculum model designed to revisit and deepen the skills acquired in previous learning.

Seven Themes of Primary SEAL:

Theme 1: New beginnings

Theme 2: Getting on and falling out

Theme 3: Say no to bullying

Theme 4: Going for goals!
Theme 5: Good to be me
Theme 6: Relationships
Theme 7: Changes

> *The resource offers a spiral curriculum which revisits each theme (and the skills associated with that theme) offering new ideas yearly, so that a child entering the school in the Foundation Stage and leaving at the end of Year 6 will have experienced each theme (at an appropriate level) each year. Children can therefore demonstrate progress in the key social, emotional and behavioural skills as they progress through the school.*
>
> (DfES 2005: 12)

Awareness of each theme is designed to be raised through a whole-school assembly and then each year group works on the theme in different ways reflecting their age, maturity and level of skill with an opportunity to come back together to celebrate and reflect on learning through a plenary assembly or other whole-school activity.
A series of materials are based around each theme have been produced which provide teaching and learning resources for children at the Foundation stage, for Years 1-2, Years 3-4 and Years 5-6. There are also extensive supplementary materials for staff development related to each theme, and for work with small groups of children 'who need additional help in developing their social, emotional and behavioural skills' (ibid: 14) and for involving parents and carers in the learning process.

Thus, if the suggested mode of delivery is followed, the teaching of SEAL skills is realised by a series of concrete group learning experiences under a broad theme unifying learning that is being shared by the whole school and all at the same time. The group learning will be set in context and later reinforced through whole school activities such as assemblies and themed weeks (such as National Anti-bullying week), and through the use of whole school display work to celebrate what is being learned. Further reinforcement comes through the children's dealings with other members of the school community such as teaching assistants and lunchtime supervisors as well as via parents and carers through home learning tasks. This highly structured and unifying approach serves to achieve the aims of SEAL being explicit and whole-school.

Theme number and time of year	**Theme title**	**Key social and emotional aspects of learning addressed**
1. September/October	New beginnings	• Empathy • Self-awareness • Motivation • Social skills
2. November/December	Getting on and falling out	• Managing feelings • Empathy • Social skills
3. One to two weeks in the autumn term (to coincide with national anti-bullying week in November)	Say no to bullying	• Empathy • Self-awareness • Social skills
4. January/February	Going for goals!	• Motivation • Self-awareness
5. February/March	Good to be me	• Self-awareness • Managing feelings • Empathy
6. March/April	Relationships	• Self-awareness • Managing feelings • Empathy
7. June/July	Changes	• Motivation • Social skills • Managing feelings

Table 3: The structure of Primary SEAL Source: DfES (2005: 19)

Primary SEAL tends to be delivered through discrete lessons using the structured resources provided in the DCSF materials. This is both a potential advantage and also a potential weakness. The advantage is clearly the extremely high profile given to the learning of social and emotional skills across the whole school. The disadvantage may be a lack of appreciation how the same SEAL skills may be utilised in the general, wider learning environment such as in lessons in other curriculum areas. In the SEAL materials related to each theme there are suggested learning opportunities for relevant SEAL skills developed to be applied in other areas of the curriculum but these are far less developed than the discrete learning activities that make up the core of the materials. Secondary SEAL is organised differently. Aimed at predominantly KS3 students (ages 11-14) the programme is broader and less structured than its primary counterpart which in some way reflects the more fragmented nature of the secondary teaching experience but, in itself, presents challenges. Rather than develop the programme through the themes of Primary SEAL, the Secondary SEAL teaching resources are organised around the 5 aspects of SEAL.

Despite the work invested in developing a structure for the Primary and Secondary SEAL programmes the non-statutory nature of SEAL, combined with the ambitious scope of the programme, and the predilection of schools to tailor initiatives to their local context, can all lead schools to adopt something of a pick and mix approach to implementation. Thus SEAL can look very different from school to school. This was something alluded to by the Senior Adviser for the Primary and Secondary National Strategies, thus charged with giving the national lead for SEAL, when addressing to Headteachers at conference on Secondary SEAL:

> *… what we really want you to do is to actually take the essence of SEAL and to take it back to your schools. Do you remember when you were a youngster and looking at in science. And sometimes they gave you a little crystal, and you went home and you had to dangle it into a solution. And actually weird and wonderful crystals grew. And you'll dangle it in the nurturing solution of your school and build your own SEAL crystal, which will all be unique to your own school because your contexts are different. But everyone will encapsulate the essence of SEAL.*
>
> (Michel 2007)

An explicit, whole-school approach

Despite this loose approach to what constitutes SEAL, a great emphasis is placed on developing a common, shared vocabulary for use by every member of the school community to address the issues related to development of social and emotional skills. SEAL is not an intervention focused on remedial skill development but rather squarely aimed at developing skills of all students in the school.

> *The resource is intended for the whole school community…In all settings, all the adults who have contact with children – teachers and teaching assistants, lunchtime staff and support staff – need to be aware of the vocabulary used and the key ideas, for example about solving problems, that are introduced to the children.*
>
> (DfES 2005: 5)

The National Strategies operated on a three wave model (see Figure 5 below) in which wave one referred to *universal* interventions that are delivered to all students, wave two focused on *targeted* intervention for small groups of selected children, and wave 3 refers to interventions delivered to individual students.

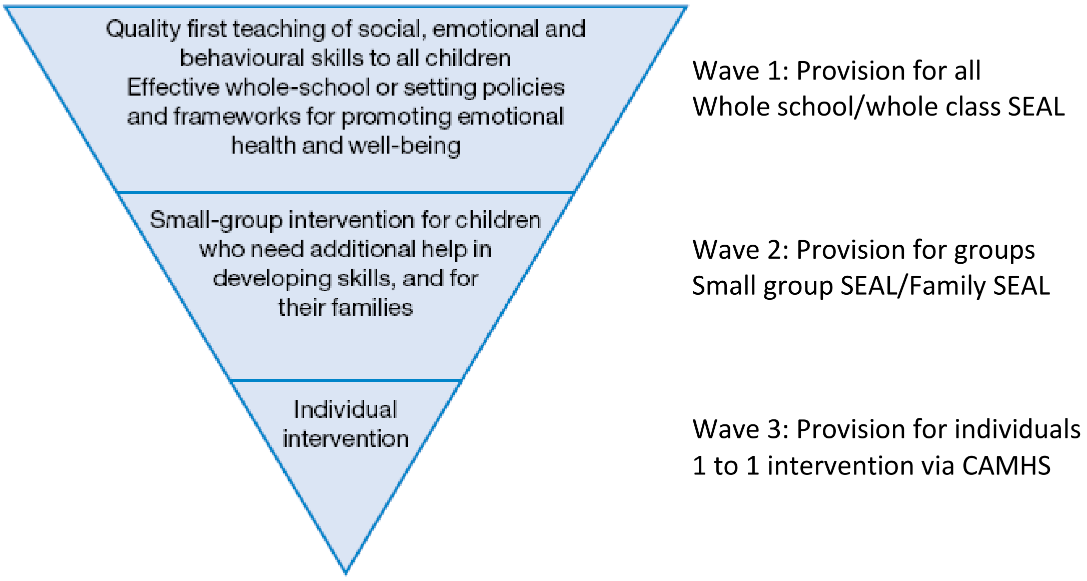

Quality first teaching of social, emotional and behavioural skills to all children
Effective whole-school or setting policies and frameworks for promoting emotional health and well-being
Small-group intervention for children who need additional help in developing skills, and for their families
Individual intervention
Wave 1: Provision for all
Whole school/whole class SEAL
Wave 2: Provision for groups
Small group SEAL/Family SEAL
Wave 3: Provision for individuals
1 to 1 intervention via CAMHS

Figure 5: The National Strategies Waves of Interventions model (Source: DFES, 2005, annotations added)

Both Primary and Secondary school operate using an explicit spiral curriculum model in which opportunities to develop specific non-cognitive skills, such as empathy or the ability to understand one's own feelings, return to the fore as you work through the materials. This is normally on a year to year basis.

> *Most social, emotional and behavioural skills are developmental and change over time. For example, if we think about the experience of loss, we know that children's capacity to manage the feelings involved, and the range of strategies at their disposal, will be very different in the early years than, for example, their experience at the age of 11. We cannot therefore 'teach' these skills as a one-off. There is a need to revisit and develop the concepts, understanding and skills over time, building on what has been learned previously.*
>
> (DfES 2005: 7)

There is an explicit understanding revealed here that these social and emotional competences can be developed and taught, and, as with cognitive outcomes, there is a sense of sequencing of leaning in age appropriate or learning stage appropriate facets a particular competence may be in view within any specific lesson or learning activity.

2.2 Distinctive Features of Family SEAL

Family SEAL (DfES, 2006) was added to the suite of resources a year after the national launch of Primary SEAL. The Family SEAL programme seeks to engage parents as partners in developing children's social and emotional competence through an introductory presentation followed by a series of seven, 2-hour sessions which broadly tackle each of the seven themes in Table 3 above (with the exception of the 'Say no to bullying' theme which is replaced by a second session on 'Going for Goals'). Each weekly session consists of an hour-long workshop, led by teacher facilitators. Each parent workshop typically begins with a variety of ice breaker games followed by regular time given over to negotiating and reviewing group expectations and confidentiality. This would be followed by a description, and sometimes modelling by facilitators, of the range of Primary SEAL approaches used by the school in its universal provision for all children. Parents would then be encouraged to consider how such SEAL approaches might be applied in family and home situations through discussion and sometimes through role-play. There might also be opportunity to reflect and discuss how these approaches compare with the parents own childhood experiences, both at home and at school.

After the workshop with parents the children enter the session in order to engage in an hour of structured activities with their parents. The activities are designed to provide opportunities to consolidate some of the strategies for social and emotional development that were covered in the preceding workshop. Suggested activities include making and playing a board game, shared relaxation exercises, producing star charts and negotiating rewards, constructing a feelings refrigerator-magnet, participating in team games and kite building (DfES, 2006: 14-15). Towards the end of each session there is encouragement for participating parents and children to apply the learning and experienced gained that week in their home settings. Parents are invited the following week to share their experience of taking the learning home. The materials give guidance to school facilitators on how to run a successful session, how to foster confidentiality within the group setting and on various practical issues such as venue and resources.

Although SEAL was designed as a universal approach to developing social and emotional competence in schools it also adopts the ‘waves of intervention’ model that underpins the whole National Strategies programme (DCSF, 2009) including the National Literacy and Numeracy Strategies. The model shown in Figure 5 suggests that Family SEAL could be considered a ‘Wave 2’ intervention, but this would also imply some sense of targeted provision for those children considered in need of specific additional skill development and, as will be made clear, such a targeted approach is not necessary for Family SEAL and was not the approach adopted in the pilot programme that was the focus of part of this study.

2.3 What are the specific outcomes in terms of social and emotional and behavioural skills?

In each programme, the 5 broad aspects of SEAL skills were broken down further into 42 ‘I can’ statements in primary SEAL and 50 learning objectives in secondary SEAL. These statements and objectives describe the fine detail of what SEAL is trying to achieve. The learning objectives are quite demanding, even at the primary level. For example “*I can make, sustain and break friendships without hurting others*” and “*I can resolve conflicts to ensure that* *everyone feels positive about the outcome*" suggest that win-win type resolutions are always possible rather than just representing a particularly desirable outcome.

As well as the spiral nature of development within each programme, there is also a sense of development across the programmes, thus, in the section on understanding my feelings in the self-awareness aspect "*I know that it is OK to have any feeling, but not OK to behave in any way I feel like*" in Primary SEAL becomes "*I can recognise conflicting emotions and manage them in ways that are appropriate*" and "*I can use my knowledge and experience of how I think, feel, and respond to choose my own behaviour, plan my learning, and build positive relationships with others*" in Secondary SEAL.
Similarly, "*I can express a range of feelings in ways that do not hurt myself or other people*" in Primary SEAL becomes "*I have a range of strategies for managing impulses and strong emotions so they do not lead me to behave in ways that would have negative consequences for me or for other people*" in Secondary SEAL

Sometimes, in an attempt to demonstrate progression in learning, the outcomes seem to cloud what is trying to be achieved. For example, under to aspect of empathy in Primary SEAL "*I can understand another person's point of view and understand how they might be feeling*" becomes "*I can see the world from other people's points of view, can feel the same emotion as they are feeling and take account of their intentions, preferences and beliefs*" in Secondary SEAL. There is a subtle, but nonetheless considerable difference between understanding how another person might be feeling, and feeling the same emotion that they are feeling, especially if the implication intended here is that it is possible for a person to experience the same emotion as another when faced with the same stimulus. This is something the Secondary SEAL objectives concede in another of the learning outcomes associated with empathy; "*I understand that people can all feel the same range of emotions, but that people do not necessarily respond in the same way to similar situations, and that different people may express their feelings in many different ways.*"

A clear theme running through the complete list of statements and objectives is an emphasis on social conformity, with less of an emphasis on autonomy. The emphasis lies heavily over to the side of social harmony, and compliance to group norms is taken as being the measures of positive emotions and behaviour. Any ground given to autonomy is couched in the need to maintain a harmonious social environment. For example, from Secondary SEAL "*I can achieve an appropriate level of independence from others, charting and following my own course* ***while***

maintaining positive relationships with others" and "*I can be assertive* ***when appropriate***." [emphases added].

Such statements and objectives suggest that there is limited place for any notion of creative tension or conflict. This is despite the fact that the imbalance of compliance versus autonomy was highlighted in research pre-dating SEAL, and commissioned by the DfES, into the assessment of social and emotional competence in preschool and primary settings (Edmunds & Stewart-Brown 2003). In this review the authors identified a prevalent assumption that the social competence, particularly of children 'implies an element of social compliance' (*ibid*: 48). Weare has also raised the need to balance other aspects of social and emotional competence with a clear emphasis on autonomy when she stated at a conference on Secondary SEAL (Weare 2007) that autonomy was the key balancing feature required to prevent an approach based on a combination of relationships, clarity of purpose and participation from becoming a form of social coercion.

Critics of SEAL have highlighted this as a major issue in the agenda behind SEAL. One of the most vociferous critics points out that an over emphasis on compliance will come at the expense of creativity and entrepreneurial endeavour which would require students to have the confidence to be different from others (Craig 2007).

> *It is not too difficult to see why encouraging children to read others' feelings, and develop empathy, could at least for some children lead them to be overly concerned with other people's views and feelings.*
>
> (Craig 2007: 65)

Craig goes on to assert that such an emphasis on social conformity may actually lead to rebellion in a minority of children and so, rather than facilitating a more harmonious learning environment SEAL could contribute to the very behaviour problems it seeks to address. Edmunds and Stewart Brown (2003) make an important distinction between socially competent, socially desirable and socially conformist behaviour. They suggest (ibid: 14) that social desirability often includes an element of social conformity, including behaviours that may suit the assessor of such abilities such as teachers. They stress that social or emotional competence should not necessarily imply conformist behaviours and give the example of taking an ethical, but unpopular stance over a particular issue.

> *Conformity in children makes the job of parents, practitioners and teachers easier, but it may be counterproductive in terms of the development of desirable attributes such as positive mental health and good citizenship.*
>
> (Edmunds & Stewart-Brown 2003) (15)

Craig acknowledges that training in emotional literacy may be of benefit to leaders and managers and also for teachers and other professionals working with children and young people. She also concedes that it may be of benefit to older secondary aged students where an informal approach is employed (*ibid*: 4). However, she has major reservations with the formal, whole-school and spiral curriculum approach of SEAL in developing the emotional literacy of younger children.

> …we believe that *the DfES Guidance is encouraging a major psychological experiment on England's children which we think could unwittingly backfire and undermine some young people's well-being in the longer term.* (Craig 2007, emphasis in the original)

Craig then turns her critical attention to the rationale behind SEAL, aiming her critique at the DCSF and in particular the research commissioned by the then DfES to carry out a review of studies and programmes aimed at developing the social and emotional competence of children (Weare & Gray 2003). Craig's criticism is aimed not so much at the fruits of the review, although she does bemoan the lack of notes of caution and temperance that she feels are present in other writings by one of the principal researchers (Weare 2004), but rather, at what she feels is the lack of caution exhibited by Government in promoting such a comprehensive, whole-school approach to the development of social and emotional skills exhibits based on what she feels is a scant evidence base. Craig calls for a scientific approach to evaluation the effectiveness of SEAL building clear evidence of based on control studies of the effectiveness of SEAL. She feels the evidence base marshalled by Weare and Gray (2003) is too wide a mix of broad-based, multi-factorial studies lacking controls and based mainly in the US. Inadequate evidence for the potential effectiveness of a programme like SEAL argues Craig. She is therefore highly critical of Weare and Gray's recommendations that the then DfES should prioritise a whole-school programme to develop the social and emotional competence of both students and staff through an explicit, curriculum-based programme that adopts a long term developmental approach (Weare and Gray 2003: 5-8). It is clear from familiarisation with the Primary and Secondary SEAL materials that the DCSF has taken on board the recommendations of Weare and Gray. Craig reserves her strongest criticism for the fact that SEAL is set to become an year on year developmental programme that spans across the complete school age range from pre-school (aged 3) to sixth form (aged 18).

> *So despite the very limited evidence at their disposal, Weare and Gray prescribe a course of treatment which will be as intense as possible in its effect. In other words,* ***this is no low dosage pill but a massive infusion of ingredients which they cannot know with any certainty will work but which has the potential to inflict serious damage on the patient – both the education system itself, as well as individual teachers and students.***
>
> (ibid: 29 emphasis in original)

Whilst proponents of SEAL like to make much of the way they feel it underpins the learning process it has had to fight for its share of curriculum time against the 'big guns' of literacy and numeracy in primary curriculum or, English maths and science in the secondary curriculum which suggest that Craig's hyperbolae such as "massive infusion" is far too extreme.

One could argue that SEAL is more a *pedagogical experiment* than a psychological one. The premise is that the acquisition and development of social and emotional skills enables both students and staff to better participate in the emotional endeavour of learning in the social environment that is school. As Goleman puts it, perhaps more forcefully than the SEAL materials, "emotional aptitude is a *meta-ability*, determining how well we can use whatever other skills we have, including raw intellect." (Goleman 1996: 36).

The place of SEAL in the wider school curriculum

There has been some acknowledgement of the part SEAL has to play in the curriculum by the Qualifications and Curriculum Authority (QCA). In a speech to a conference of Headteachers the Director of Curriculum for QCA (Waters 2007) linked SEAL to the revised secondary National Curriculum, and in particular to the Personal, Learning and Thinking Skills (PLTS) that form part of the new curriculum:

- Independent enquirers
- Creative thinkers
- Reflective learners
- Team workers
- Self-managers
- Effective participators

(QCA 2007)

There is some clear cross over between the success criteria for PLTS and the outcomes of SEAL, for example, as independent enquirers young people are expected to "*consider the influence of circumstances, beliefs and feelings on decisions and events*", as reflective learners they should "*invite feedback and deal positively with praise, setbacks and criticism*" and as self-managers they should "*manage their emotions, and build and maintain relationships*".

> *The SEAL outcomes link to the work we've been doing on personal, learning and thinking skills … For the first time, the curriculum talks about skills as something that we expect youngsters to experience. And the SEAL outcomes are a way forward in terms of planning and developing those personal, learning and thinking skills in context.*
>
> (Waters 2007)

In a forward look at the potential curriculum for 2020 (Gilbert 2006), a set of skills and attitudes valued by employers are listed

- being able to communicate orally at a high level
- reliability, punctuality and perseverance
- knowing how to work with others in a team
- knowing how to evaluate information critically
- taking responsibility for, and being able to manage, one's own learning and developing the habits of effective learning
- knowing how to work independently without close supervision
- being confident and able to investigate problems and find solutions
- being resilient in the face of difficulties
- being creative, inventive, enterprising and entrepreneurial.

(Gilbert 2006: 8)

It is clear that SEAL is designed to develop some of these skills and attitudes, especially under the aspects of motivation and social skills in group work settings but SEAL goes well beyond the skills listed here in terms of how it promotes understanding of one's own feelings and the feelings of others and managing the same. The 2020 document makes no explicit reference to SEAL and there is only a single mention of social skills in the context of primary education and no mention at all of emotional skills as part of the curriculum.

This is in stark contrast to the report by the Practitioners Group on School Behaviour and Discipline (Steer 2005) which has ten explicit references to SEAL as a resource and numerous references to both social and emotional skills.

> *We see the development of pupils' social, emotional and behavioural skills as integral to good learning and teaching. It is also integral to making classrooms orderly places for learning. This means teaching all pupils, from the beginning of education, to manage strong feelings, resolve conflict effectively and fairly, solve problems, work and play cooperatively, and be respectful, calm, optimistic and resilient. We have seen evidence that the social and emotional behavioural skills programmes being promoted by the Primary National Strategy (SEAL), and about to be piloted within the Secondary National Strategy (SEBS), are proving successful in developing these crucial skills and attitudes and 'growing good learners'. These are new programmes and we believe that schools would benefit from wider knowledge of, and access to, them. We welcome the extension of the primary work to the secondary phase. We believe it important that the SEAL work should be further promoted and embedded by the DfES.*
>
> (Steer 2005: 34-35)

There is, therefore, a much clearer emphasis on the role SEAL can play on influencing behaviour than there is on its potential influence on learning in school settings and beyond. This is not surprising as SEAL was developed as part of the Behaviour and Attendance strand of the National Primary and Secondary Strategies and was originally referred to as Social, Emotional and Behavioural Skills (SEBS). This produces something of a double edged sword for SEAL in terms of its primary purpose. Is SEAL about supporting children to conform to the expected norms of behaviour in school settings, or is it about providing skills to thrive in the social and emotional learning environment of school? There are strong links between this double edged approach to the core rationale for SEAL and the concerns raised in the conformity versus autonomy debate mentioned above. Which side of this sharp edge SEAL falls is likely to be in the hands of teachers working individual schools and how they interpret and implement SEAL in their local context.

2.4 Summary

This chapter has provided an overview of the development and nature of both Primary and Secondary SEAL and how the developers have proposed that the programme is implemented in the curriculum. It has indicated that strong links exist between the original Primary SEAL and Secondary SEAL to support effective transition across this critical interface between phases of schooling. This helps to make the case for viewing SEAL as a unified system with common aims and purpose across the range of KS1 to KS3 (ages 5-14) and potentially beyond to KS 4 (ages 14-16). The three-wave National Strategies model adopted by SEAL illustrates the mixed approach of both universal and targeted provision that will be addressed in this study by considering estimation of non-cognitive outcomes of both universal and Family SEAL. Before moving to the research design for the study we will consider, through a review of the literature, how the approach adopted by school effectiveness researchers has been used to estimate both cognitive and non-cognitive outcomes of schooling.

Chapter 3: Literature Review

This chapter consists of two main sections. In the first part it charts the application of school effectiveness research findings and analytical principles in the development of "value added" measures of school performance, specifically in England, over the period 1990 to 2006. The school effectiveness approach has become characterised by the application of analytical techniques known as multilevel modelling, which account for the clustering of students in schools, and help to portion the variance between the student and school levels and so determine the school effect. This narrative account also serves the purpose of indicating the types of data and visualisations that schools have become accustomed to utilising in order to monitor and evaluate the progress of students and to determine their relative effectiveness of their school in terms of the progress made by their students in cognitive outcomes. The size and interpretation of the school effect determined in school effectiveness studies is reviewed and the relatively smaller number of school effectiveness studies that also include the intermediate level of the class grouping are considered. The second section of the review takes a similar look at school effectiveness studies of non-cognitive outcomes, including some of the methodological challenges posed by measuring non-cognitive outcomes, and also compares the magnitude of school and class level effects observed for cognitive and non-cognitive outcomes of schooling.

3.1 The development of school performance measures in England – applying the school effectiveness framework to inform self-evaluation and accountability within and between schools

Over the last two decades successive UK Governments have sanctioned the development of an increasingly sophisticated suite of school effectiveness measures for schools in England and Wales. As a result schools now have access to value-added metrics that utilise advanced statistical techniques, the fruit of decades of international school effectiveness research, which are designed to provide robust and reliable measures of the educational value added by individual schools to the educational outcomes of their students. Schools and classroom teachers are actively encouraged to use these measures to evaluate their institutional and personal educational effectiveness. Politicians describe schools as 'data rich' environments (Miliband 2003) equipped for and era of 'intelligent accountability'(Miliband 2004) and the use of both national and local statistical measures has recently been incorporated into a suite of revised professional standards for newly qualified teachers (TDA 2007c) through to Excellent Teachers (TDA 2007b) and Advanced Skills Teachers (TDA 2007a), currently the highest levels of classroom practitioner. Teachers and School Leaders are encouraged to use data to evaluate the effectiveness of their teaching and learning development initiatives, which, as we have already noted in Chapter 1, includes broad-based interventions such as the Social and Emotional Aspects of Learning programme (SEAL).

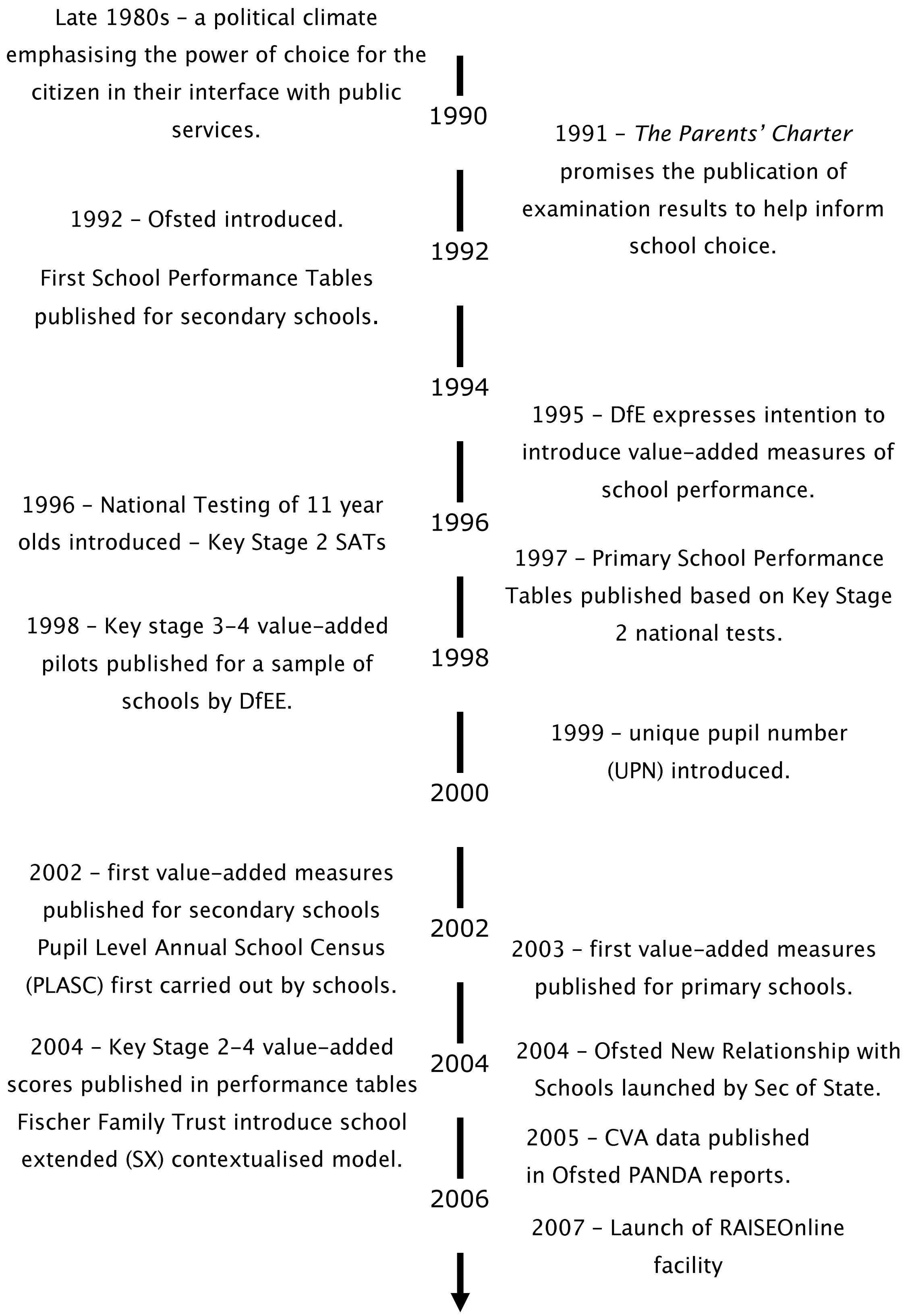
As Miliband's phrase 'intelligent accountability' suggests, there is another agenda at work here, that of using value-added measures to hold schools and teachers to public account for the educational outcomes of their students. The original, raw attainment metrics such as the proportion of students crossing a particular academic threshold (e.g. 5 or more A*-C grades at GCSE), were developed for publication in School Performance Tables (more commonly known as League Tables) and have been utilised by the Office for Standards in Education (Ofsted) for accountability purposes through its inspection regime. As more sophisticated value-added measures have been developed, these too have been adopted for the accountability agenda. Ofsted uses value-added measures to form pre-inspection hypotheses before visiting a school for inspection and schools can be ranked by value-added scores in League Tables published by the national news media.

This section of the literature review will chart the move from raw attainment measures of school effectiveness published in the first school performance tables published in the early 1990s through to the highly contextualised value-added measures available today generated by relatively sophisticated statistical modelling. It will focus on the two key sources of contemporary value-added data available to schools, the contextualised value-added (CVA) measure produced by the Government Department for Children, Schools and Families (DCSF) and Ofsted, and the value-added models developed by the Fischer Family Trust (FFT), 'lifting the lid' on both measures to give technical insights into the statistical models employed. The review will raise the tension inherent in using the same data for the potentially crossed purposes of school self-evaluation for school improvement and for the purposes of external accountability of school effectiveness. Finally, it will draw on the preceding material to make a case for the use of Fischer Family Trust value-added measures assist in the evaluation of such broad school improvement initiatives as the SEAL programme.

3.1.1 The background to the development of value-added measures

At the end of the 1980s the then Conservative government were pursuing policies to increase the accountability of public sector bodies to members of the public with the introduction of the Citizens' Charter. The Citizens' Charter was comprised of three areas of public sector delivery in transport (the passengers' charter), health (the patients' charter) and education (the parents' charter). The Parents' Charter was published in 1991 (Hoyle & Robinson 2002). The charter reflected the desire on the part of the Conservative government to put information into the hands of parents to enable them to make an informed choice of schools to which they might send their children. The charter contained five promises to parents, two of which related to information coming directly from government to inform school choice; the publication of raw examination results as a measure of school performance and the publication of extensive reports of school inspections. The year after publication of the charter saw both the implementation of both measures as school performance tables and the Office for Standards in Education (Ofsted) were introduced and both had a major impact on how secondary schools were viewed by those within and outside the system. Ofsted was introduced under the leadership of Chris Woodhead as Chief Inspector for Schools in order to provide qualitative reports of every state school in England and Wales which would be made available for publication and accessible to parents.

Secondary school performance tables were published in the form of percentage of students in each school attaining 5 or more passes at GCSE grades A to C. The tables were published by the Department for Education as alphabetical listings of schools (Hoyle & Robinson 2002) but these listings were often adjusted by newspapers to a "League Table" format with schools ranked on their percentage 5+ A-C GCSE pass figures. Immediately schools were named as belonging to top or bottom percentage groups and intense focus was placed on the schools at the top and bottom of the tables with less attention paid to the difference in raw scores separating the "top" school from those below or the "bottom" school from those above. This practice of ranking schools still continues despite the development of refined school performance measures. The BBC News website allows its visitors to sort schools in a particular Local Authority by performance measures such as raw examination scores and value-added scores, all at the click of a mouse, as well as publishing reviews of the "best" and "worst" schools as given by the various performance measures with links to "schools that add the most value".



Late 1980s – a political climate emphasising the power of choice for the citizen in their interface with public services.

1991 – *The Parents' Charter* promises the publication of examination results to help inform school choice.

1992 – Ofsted introduced.

First School Performance Tables published for secondary schools.

1995 – DfE expresses intention to introduce value-added measures of school performance.

1996 – National Testing of 11 year olds introduced – Key Stage 2 SATs

1997 – Primary School Performance Tables published based on Key Stage 2 national tests.

1998 – Key stage 3-4 value-added pilots published for a sample of schools by DfEE.

1999 – unique pupil number (UPN) introduced.

2002 – first value-added measures published for secondary schools
Pupil Level Annual School Census (PLASC) first carried out by schools.

2003 – first value-added measures published for primary schools.

2004 – Key Stage 2-4 value-added scores published in performance tables
Fischer Family Trust introduce school extended (SX) contextualised model.

2004 – Ofsted New Relationship with Schools launched by Sec of State.

2005 – CVA data published in Ofsted PANDA reports.

2007 – Launch of RAISEOnline facility

Figure 6: Timeline showing the development of value-added measures of school effectiveness in England and Wales.

A synthesis of (Hoyle & Robinson 2002; Schagen & Hutchison 2003; Taylor & Nguyen 2006)

Schools have clearly been a testing ground for public sector performance indicators. It was nearly a decade later before such performance rankings for health providers were released in the form of 'high level' and 'clinical indicators' for NHS hospitals in 1999 and 'star ratings' for hospitals in 2001 (Hoyle & Robinson 2002). Since then League Tables generated from such performance rankings have become the ubiquitous measure of public sector performance.

It soon became clear that such raw measures of school performance were not popular with many schools as they failed to compare "like with like". Thus grammar schools were ranked in the same tables as secondary moderns and comprehensives with no consideration of the prior attainment of students on entry to different schools. Inner city schools were ranked with those in 'leafy suburbs' with no account made of the socio-economic backgrounds of students in the schools despite a strong body of research evidence from the international field of school effectiveness that suggested that socio-economic status (SES) factors were strongly correlated with the academic attainment of students.

The publication of school 'league tables' based on raw examination results had been foreseen by school effectiveness researchers (Goldstein & Cuttance 1988; Smith & Tomlinson 1989) and their eventual publication provoked very critical responses both from schools and from the school effectiveness research community. Researchers had already been arguing convincingly for a distinction between the standards attained and the progress made by students. Gray had posed key questions about the circumstances under which one school might be considered to have done better than another (Gray *et al.* 1986). In answer to such questions a growing number of school effectiveness researchers were arguing for a sophisticated statistical methodology that accounted both for school context and also for the hierarchical or clustered nature of school data, reflecting the fact that students are nested within schools so that any set of students in the same school should be more similar than the general population of students from all other schools. One statistical technique that adjusts for such clustering of school data is a variant of regression known as hierarchical linear modelling or multilevel modelling (Goldstein 2003; Kreft & De Leeuw 1998).

The government responded to these criticisms in different ways. In 1995 the Department for Education expressed its intention to develop "value-added" measures that adjusted for the prior attainment of students before entering secondary school (Ray 2006). An extensive period of consultation and research began that culminated in the first publication of valued-added measures in school performance tables seven years later. A key aspect of this consultation was the Value-added National Project (Fitz-Gibbon 1997) commissioned by the

Schools Curriculum and Assessment Authority (SCAA). The project was led by Carol Taylor Fitz-Gibbon, a leading UK school effectiveness researcher who was Director of the Curriculum, Evaluation and Management (CEM) Centre at the University of Durham. The CEM centre had been producing value-added analyses to aid school improvement and self-evaluation such as the Advanced Level Information System (ALIS) since 1983 (Fitz-Gibbon 1991). The results of these analyses were well respected by schools. The Final Report for the project makes a comprehensive set of recommendations related to reporting value-added progress measures for both internal school improvement and as externally published measures of accountability (Fitz-Gibbon 1997: section 7)

In essence Fitz-Gibbon argues for a system based on:

- Residual Gain Analysis (RGA) based on ordinary least squares (OLS) regression methods
- school value-added scores as the simple mean of the scores of their individual pupils
- omission of further contextualising variables such as gender, SES proxy variables such as percentage free school meal entitlement (FSM) to prevent stereotyping and lowering of expectations, but some consideration given to schools with disproportionate numbers of students from such groups (such as boys schools or schools with >60%FSM)
- external publication of value-added measures summarising the progress of at least three cohorts of students to counter the inherent year-on year instability of value-added scores
- adoption of minimum threshold number on role below which school value-added scores would not be published
- publication of both subject and syllabus specific regression lines to facilitate interpretation of the confounding effects present due to the variety of subjects and syllabi available in the UK school examinations system

The additional recommendations cover wide range of issues from the need to develop the statistical literacy of teachers as part of initial teacher training to enable them to understand and interpret value-added analyses, through to the rigorous monitoring of the examinations and assessment system to ensure comparability of outcomes between subjects, syllabi and examination sessions. Fitz-Gibbon also calls for an acknowledgement of the narrow view of education that value-added measure represent calling for other measures of educational outcomes to be developed (*op cit.*).

In a annexes to the report (*ibid*: annex C-annex D) there is a some technical coverage of whether the multilevel regression techniques referred to above that adjust for the clustered nature of school performance data are to be preferred over the less complex OLS methods recommended by the report. This section of the report is largely based on earlier work of Fitz-Gibbon responding to criticisms of the CEM Centre's development of value-added analyses such as ALIS (Fitz-Gibbon 1991). The conclusion of the report is that multilevel modelling (MLM) presents little statistical benefit over simpler OLS techniques, especially for schools with reasonably sized cohorts, and that the corresponding complexity introduced by MLM detracts from the ability for teachers and school leaders to understand and the statistical procedures employed. Fitz-Gibbon identifies two flaws in the calls to employ MLM in value-added analysis. The first is that the use of school as the upper level in the data hierarchy (i.e. students nested within schools) is not valid. If MLM is to be employed, according to Fitz-Gibbon it is the class or teaching group that should be the correct upper level in the data hierarchy. The second issue is that of the shrinkage factor applied to residuals in multilevel analyses. The shrinkage factor enables more robust parameter estimates to be calculated for higher level units (schools) with smaller samples (small student cohorts). This is done by 'borrowing strength' from the larger data set and the resulting shrunken estimates for small schools are drawn closer to the mean. More detailed implications of the issues related to the choice of the higher level unit and the application of shrinkage factors in value-added measures will be discussed below.

Whilst the value-added consultation was underway the Department for Education and Ofsted continued to employ benchmarking techniques (grouping of similar schools) to adjust school effectiveness measures for prior attainment and SES factors, but only for data intended for *internal* school self-evaluation purposes via the Performance and Assessment reports (known as PANDAs) published directly to schools and local authorities by Ofsted. Figure 3.2.1 shows a typical PANDA analysis of the percentage of students attaining National Curriculum level 5 and above in nationally administered tests (KS3 SATs) for the National Curriculum core subjects of English, science and maths. When the school is benchmarked with schools admitting students with a similar prior attainment range (at the end of KS2) the reported performance improves considerably, with science, for example, moving from between the 5th percentile and the lower quartile to just under the 60th percentile.

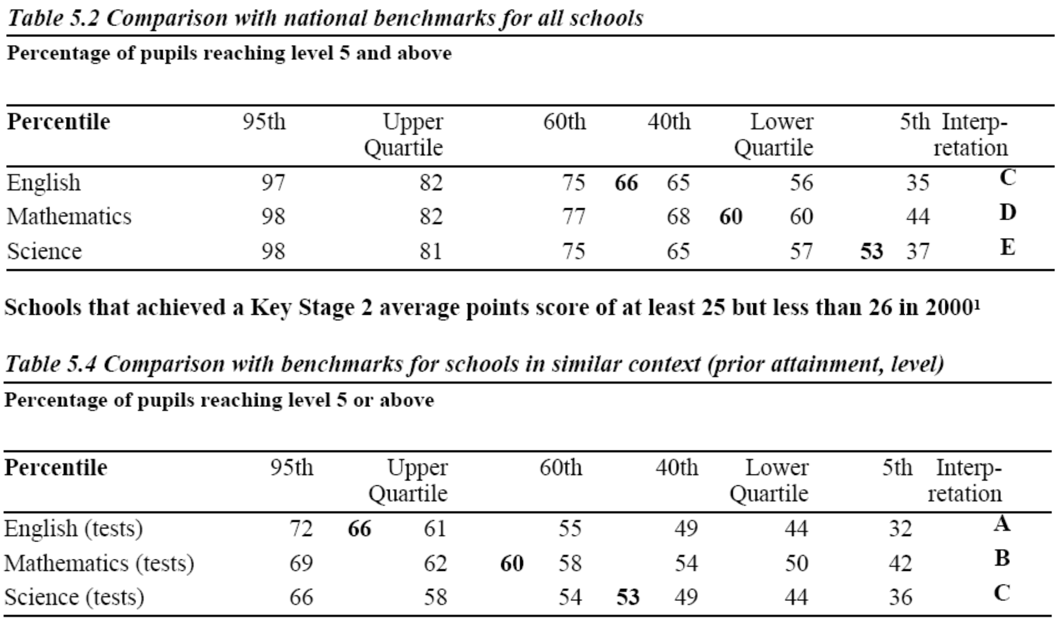


Table 5.2 Comparison with national benchmarks for all schools

Percentage of pupils reaching level 5 and above

Percentile	95th		Upper Quartile		60th		40th		Lower Quartile		5th	Interpretation
English	97		82		75	**66**	65		56		35	**C**
Mathematics	98		82		77		68	**60**	60		44	**D**
Science	98		81		75		65		57	**53**	37	**E**

Schools that achieved a Key Stage 2 average points score of at least 25 but less than 26 in 2000[1]

Table 5.4 Comparison with benchmarks for schools in similar context (prior attainment, level)

Percentage of pupils reaching level 5 or above

Percentile	95th		Upper Quartile		60th		40th		Lower Quartile		5th	Interpretation
English (tests)	72	**66**	61		55		49		44		32	**A**
Mathematics (tests)	69		62	**60**	58		54		50		42	**B**
Science (tests)	66		58		54	**53**	49		44		36	**C**

Figure 7: Extracts from a typical PANDA report (Ofsted 2003)

Figure 7 shows extracts from a typical PANDA report showing the effect of benchmarking (in this case for prior attainment) on the judgement of school effectiveness. The grades given under the heading 'Interpretation' correspond to the position of the percentage attainment for each subject on the percentile ranges.

When value-added scores were published for a pilot group of schools in 1998 the methodology used was different to that recommended by Fitz-Gibbon. A very simple median model was employed rather than OLS regression techniques.

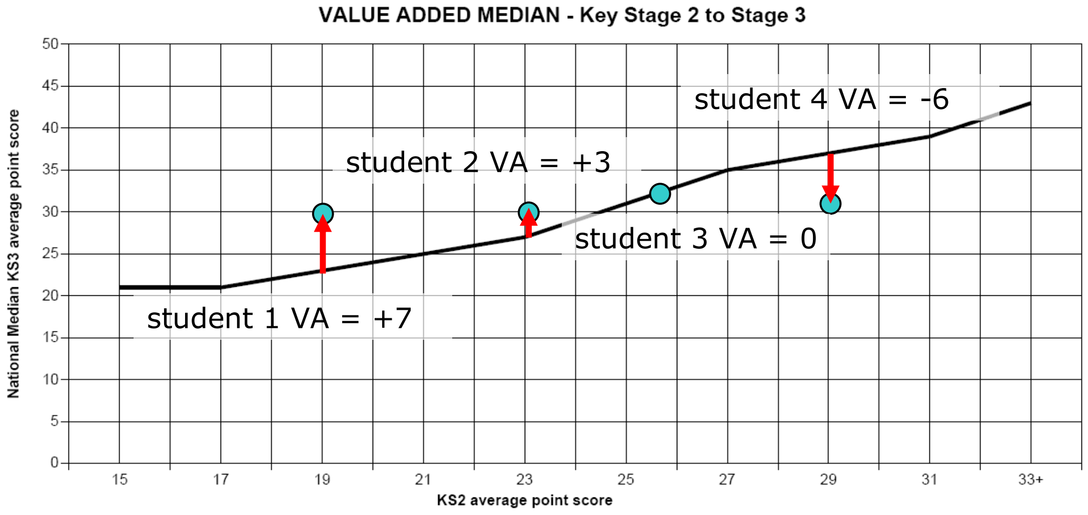


Figure 8: A median value-added plot (DfES 2004) [annotations added]

In Figure 8 above the solid line represents the median performance of students in KS3 national curriculum assessments (SATs) based on their prior attainment at the end of KS2. The outcomes for four typical students have been plotted on the figure together with annotations showing how their VA scores have been calculated. The school value-added score is simply the mean of the student VA scores with a value of 100 added (or 1000.0 for KS2-4 and KS3-4 value-added scores) for cosmetic reasons so that all school VA scores are positive.

school VA score = (7 + 3 + 0 – 6)/4 + 100.0 = 4/4 + 100.0 = 101.0

One consequence of this mix of median and mean averages in the VA methodology was the apparently odd result that the mean of the school mean VA scores could (and usually was) less than 100.0 (or 1000.0). This was seen by some critics (Gorard 2006) as evidence of a slide in standards rather than the more likely explanation of ceiling and floor effects skewing the distribution of scores. Some researchers argued for a consistent approach, urging the use of the mean rather than the median throughout (Critchlow & Coe 2003).

A key advantage of the median VA method was its simplicity. As soon as median lines were published schools were able to calculate VA scores for their own students and hence work out their own school VA scores. They were thus able to interact with the data, calculating VA scores for groups of students such as teaching groups or students with special educational needs. They could check the veracity of published scores and could determine the influence of 'outlier' VA scores on the overall school VA score. Eventually an electronic tool called the Pupil Achievement Tracker (PAT) (DfES 2004) was developed to help schools carry out such in house analyses.

The first school performance tables incorporating school value-added scores were published in 2002. Despite being heralded as an advance in measuring school performance by adjusting for student prior attainment allowing schools to be compared on more of a like-for-like basis, the median VA metric was not without its critics. One of the most vocal of these was Stephen Gorard from the Department of Educational Studies at the University of York. In a paper entitled "Value added is of little value" (Gorard 2006) Gorard argued that school value-added scores were little better than raw attainment measures of school effectiveness. To provide evidence for this claim he investigated the degree of association between VA scores and the national GCSE benchmark of percentage of students attaining 5+A*-C passes.

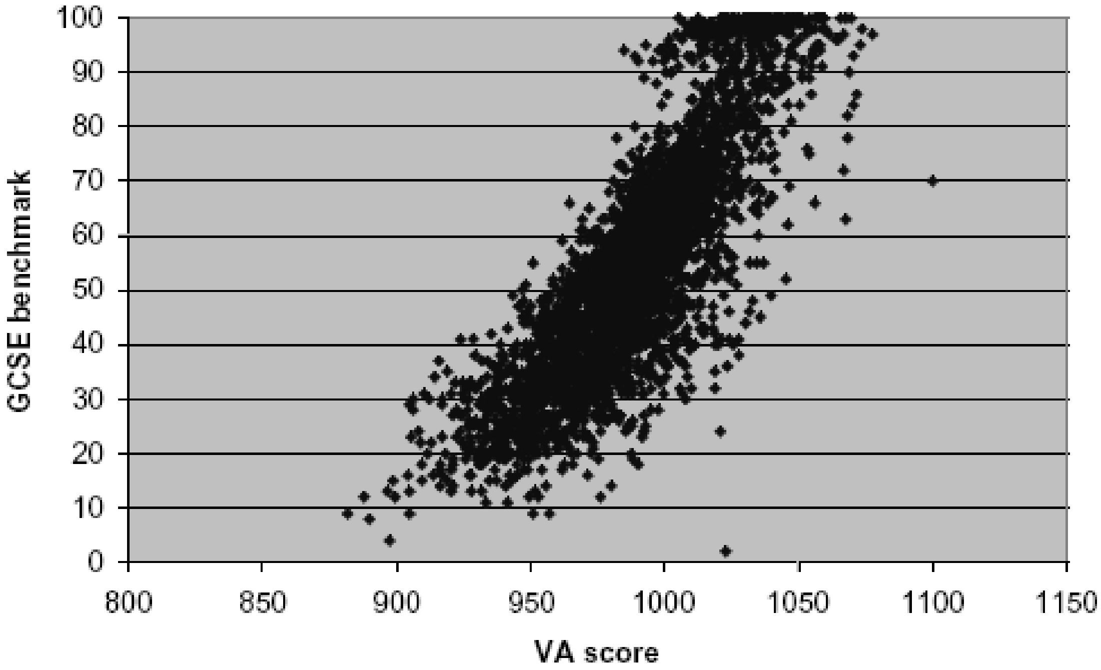


Figure 9: Analysis of the association between GCSE benchmark attainment and school KS2 to GCSE VA scores

Figure 9 uses the 2004 published data by the DfES (schools with under 30 students in their cohort are not included). (Source – Gorard 2006: 240).

Gorard found that the two measures had a Pearson correlation of r = 0.84 signifying that 71% of the variance (r^2) in VA scores could be explained by the raw attainment measures. Gorard argued that such a high correlation between the old and the new measures was evidence that VA scores were adding little to our understanding of school performance:

> *…there is a clear pattern of low attaining schools having low VA, and high attaining schools having high VA. Value-added scores are no more independent of raw-score levels of attainment than outcomes are independent of intakes.*
>
> (*op cit.*)

Gorard concluded that such measures were therefore “worse than pointless” (ibid: 241) as they could lead to politicians and families being misled about the relative effectiveness of schools. He was also critical of calls to add further complexity to VA models by the use of multilevel modelling techniques and the introduction of contextualising variables, “This may mask, but will not solve, the problem described here” *(op cit)*.

A wider ranging critique (Moody 2001) of the whole rationale behind the development of value-added indicators and use of predictive models for the purpose of target setting concluded that VA measures were "being driven by political necessity, and by the desire to 'prove' that educational standards are rising, rather than by any demonstrable evidence of their predictive validity or reliability." (p. 100) and warns about the use of such measures to formulate judgements on schools, teachers and students.

Moody points out that an initial report of the VA feasibility study carried out by the forerunner to the QCA known as the School Curriculum and Assessment Authority(SCAA 1994), concluded that National Curriculum levels should not be used as baseline predictors of attainment as they would lead to unreliable predictions due, at least in part, to concerns about the parity of levels across key stages. The SCAA study recommended instead that some finely differentiated but unspecified prior attainment measure be used as a baseline. Three years later (SCAA 1997) an about turn occurred with the Authority stating that standardized tests were no longer necessary as "the end of KS2 tests are now sufficiently established and reliable to lend confidence to their results' although the report did still express the criticism of some that "the KS2 tests is that the levels are not equivalent to the same levels at KS3" (*ibid*: 5).

Moody also criticises the final report (Fitz-Gibbon 1997) of the SCAA national study on VA measures which he says represents "a classic case of using a range of statistical methods to examine data, and then choosing the method which produces the outcomes which 'prove' that the existing preconceptions are well founded, whilst at the same time rejecting outcomes which tend to undermine, challenge or complicate those preconceptions." (Moody 2001: 84). It seems that Moody has some misunderstandings of the differences between OLS and MLM, possibly confusing multiple regression with multilevel modelling, and also with related issues such as shrinkage. It may be that these misunderstandings are inherited, at least in part, from the earlier SCAA reports which Goldstein criticises as employing "weak statistical procedures" (Goldstein 1997).

Moody seeks to show that the results of Cognitive Ability Tests (CATs), as developed by National Foundation for Educational Research (NFER), give a better baseline predictor of KS3 performance than either KS2 NC levels or teacher assessments. His study is seriously limited however, as the data is based predominantly on analysis of two years' of input/outcome data in a single school (a rural, girls comprehensive). Nevertheless, the debate about the suitability of NC test data as a baseline measure is an important one, for the reason that Moody highlights of implied transferability of NC levels between key stages, but also other issues such as the objectivity of the process by which level thresholds are set for the tests and subjectivity of the marking process, especially for English tests but also the fact that a whole-school's test papers are allocated to a single marker rather than randomly distributed across a range of markers which would be possible if the papers were scanned and marked electronically as an increasing number of GCSE and A-level examinations currently are. This may well result in greater internal consistency between scores within the cohort and thus facilitating some aspects of internal self-evaluation within the school, but it raises questions of reliability when making comparisons between schools based on the outcomes of NC tests. This is particularly important when the fine grade decimalised levels are used to calculate the input prior-attainment measures for the most recent incarnations of VA measures. When the resulting VA measures are used to produce comparative indicators such as percentile rankings for whole schools or sub-groups of students then the differences between each decimalised score and the associated percentile rankings can be very fine indeed!

A more methodologically focused critique (Prais 2001) of the 1998 KS3-4 median value-added pilot discusses the perils introduced by ceiling effects and differences in the spread of data when trying to compare different types of schools using value added scores (in this case the focus is on comparisons of the performance of comprehensive schools with that of grammar schools). Prais also raises the problems introduced by assigning a continuous scale (the points score system) to the ordinal scale of GCSE grades that assumes that the GCSE grade spectrum represents a continuum of challenge. As Prais notes, the points score scale makes no distinction between the student who achieves an F rather than their predicted G grade and one who achieves an A* rather than an A grade. Which of these one GCSE point score value-added scenarios represents the greatest improvement may well depend on individual circumstances reflecting issues of motivation as well as academic challenge. The value-added models developed by the Fischer Family Trust make use of ordinal regression techniques to generate estimates of GCSE performance from prior attainment in NC tests. Such techniques make no presumption of a linear continuum of challenge that is inherent in the GCSE points score system.

Gorard's critique had tapped into the impression that the median VA metric was a stepping stone to a more complex approach. His critical views of such further sophistication were heard but not heeded by the DfES and Ofsted. The median VA methodology made no adjustment for other variables such as SES. One way round this was to employ the benchmarking approach described above.

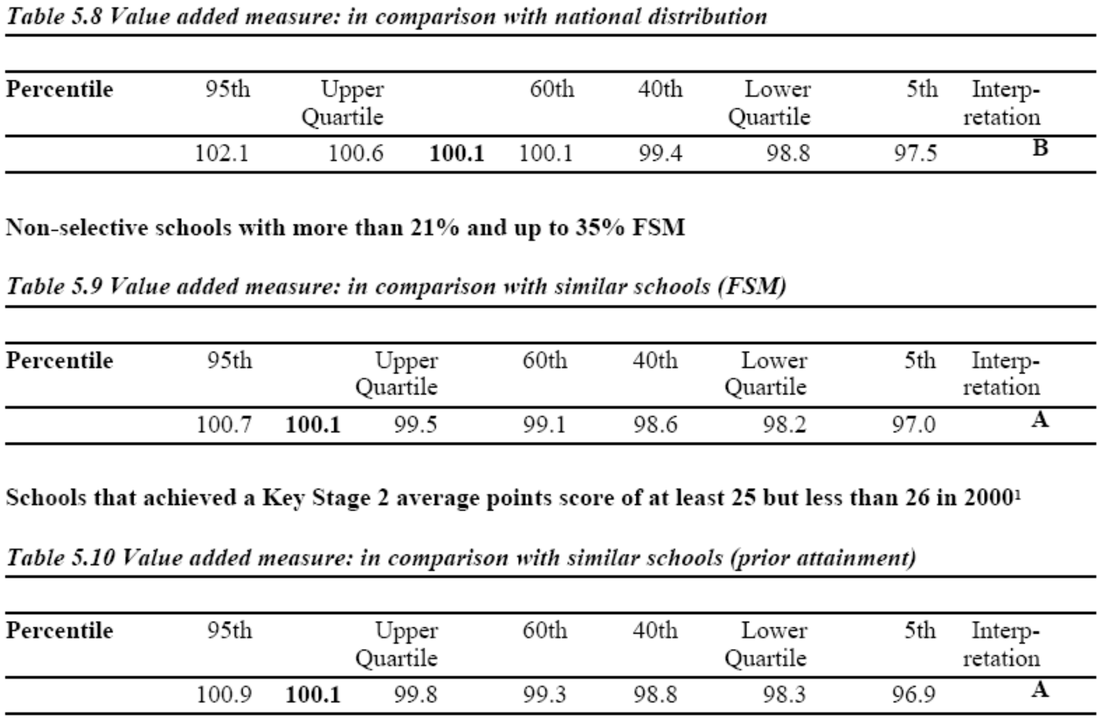


Table 5.8 Value added measure: in comparison with national distribution

Percentile	95th	Upper Quartile		60th	40th	Lower Quartile	5th	Interp-retation
	102.1	100.6	**100.1**	100.1	99.4	98.8	97.5	**B**

Non-selective schools with more than 21% and up to 35% FSM

Table 5.9 Value added measure: in comparison with similar schools (FSM)

Percentile	95th		Upper Quartile	60th	40th	Lower Quartile	5th	Interp-retation
	100.7	**100.1**	99.5	99.1	98.6	98.2	97.0	**A**

Schools that achieved a Key Stage 2 average points score of at least 25 but less than 26 in 2000[1]

Table 5.10 Value added measure: in comparison with similar schools (prior attainment)

Percentile	95th		Upper Quartile	60th	40th	Lower Quartile	5th	Interp-retation
	100.9	**100.1**	99.8	99.3	98.8	98.3	96.9	**A**

Figure 10: Extracts from a typical PANDA report showing the effect of benchmarking of school value-added scores by SES (Ofsted 2003

Figure 10 is an extract from a PANDA and shows how benchmarking could be used to contextualise the median value-added score by grouping schools into bands of similar percentage FSM entitlement as a proxy variable for SES. In this example a school with a VA score of 100.1 is positioned on the 60th percentile when compared with all schools nationally. When benchmarked with other non-selective schools with a similar %FSM entitlement it is positioned between the 75th and 95th quartiles of VA scores. Bizarrely, in a confusing example of double accounting, the value-added score has also been benchmarked against schools with similar prior attainment, despite the fact that VA scores are, by definition, already adjusted for prior attainment. A practice that lends some weight, albeit obliquely, to Gorard's pessimistic view of the potential misuse of VA measures.

Critiques such as that of Gorard had tapped into the realisation that the median VA metric was something of a stepping stone to a more complex approach. His critical views on the introduction of further technical sophistication, like the calls of Fitz-Gibbon for simplicity and clarity were heard but not heeded by the DfES and Ofsted. This may well be due at least in part to a growing group of voices representing calls for the most robust statistical methods to be applied to value-added measures.

Lesley Saunders, then Principal Research Officer at the NFER, charts the background to the development of more complex value-added measures that incorporate a range of contextualising variables and employ multilevel modelling techniques (Saunders 1999). A group of school effectiveness researchers were reflecting on methodological developments of the 1980s and drawing conclusions about what they felt were the essential components of value-added models. As early as 1991 in a piece written for the Times Educational Supplement Nuttall gave a checklist of factors that value-added models should incorporate:

- measures of prior attainment;
- a range of outcome measures (not merely five or more A-C grades at GCSE, for example);
- socio-economic variables;
- analyses of differential effectiveness;
- allowance for the possibility that schools' results are not stable over time;
- use of multi-level modelling as the statistical technique.

(Nuttall 1991) - adapted from Saunders 1999

A paper for the National Commission for Education (McPherson 1992), which defines value-added explicitly as a 'calculation of the contribution schools make to pupils' progress', contains a similar value-added model checklist to that produced by Nuttall. Such models should allow for:

- pupils' prior attainment;
- the longitudinal nature of progress;
- the multilevel nature of schools;
- the multivariate nature of the factors involved, especially 'non-school factors that boost or retard progress', such as pupils' socio-economic background;
- differential effectiveness for different groups of pupils.

(adapted from Saunders 1999)

The National Commission report also states that performance measures based on raw examination results may have their place but that such measures can lead to 'mistaken judgements, needless anxieties and fruitless 'further investigations'…triggered by false signals'. It also highlights the potential tension between providing useful information to parents and giving school teachers evidence for raising attainment in their schools and noting that value judgments will need to be made by all audiences of a VA system. Crucially for charting our journey to the development of complex and contextualised VA measures, it concludes that issues of complexity should not prevail over the need to aim for the best possible indicator.

> *Any attempt to improve schooling by means of informing choice presupposes that parents are capable of understanding at least the complexity of an adjusted outcome score. To reject that possibility is to reject the possibility of informing parents.*
>
> (McPherson 1992) cited in Saunders 1999

Such optimism was not shared by the school effectiveness researchers involved in developing the measurement methodology. Two longstanding UK school effectiveness researchers, Harvey Goldstein and Sally Thomas seem less than sanguine about the capacity of VA models, based on the fruit of school effectiveness research, to differentiate between schools on the basis of performance (Goldstein & Thomas 1995). They point to the historical nature of VA measures which are, by definition, based on past data rather than reflecting the contemporary practice and policy of schools and conclude that “research into school effectiveness is a useful activity in our attempts to obtain knowledge about the process of education, but a very poor tool for holding schools to account.” (*ibid*: 37)

Saunders summarises the situation at the end of the 1990s stating that, whilst the *principle* of value-added was comprehensible, the increasingly rigorous *analysis* had only revealed the complex and inconsistent nature of school effectiveness that lay at the centre of value-added. Drawing her words from the recommendations of the SCAA report (1994) Saunders argues that that ‘better information’ and ‘public consumption’ are incompatible especially if public consumption depends on ‘simple and straightforward’ measures of progress (Saunders 1999: 253). This is a conclusion that school effectiveness researchers (Thomas & Goldstein 1995) would concur with, “research emphatically demonstrates that the measurement of progress or value added… is neither simple nor straightforward” (*ibid*:17). Such voices formed a powerful lobby with policy makers and statisticians at the DfES:

> *Views of a selection of academics in the field were sought on the future direction of the value added work and, although there was no consensus of opinion, there was strong support from some for the development of more complex models that used the new data.* (Ray 2006: 10)

The new data Ray refers to comes from the Pupil Level Annual School Census (PLASC) which the DfES began gathering in 2002 in order to facilitate the development of contextualised value-added (CVA) models. Information such as late entry into a Key Stage (mobility), ethnic background, level of Special Educational Need (SEN), whether the student was considered to be in the care of the Local Authority or spoke English as an additional language (EAL) was gathered as well as the more classic contextualising information such as entitlement to free school meals (FSM). This much criticised SES proxy variable was bolstered by the addition of another Government produced measure known as the Income Deprivation Affecting Children Index (IDACI) calculated by the Communities and Local Government department. IDACI is a measure of the proportion of children under 16 living in families in receipt of a specified range of income support measures. The index (which ranges in value from 0.0 to 1.0) is linked to the student's postcode but the area of measurement is much wider, the lower-level super output area (LLSOA), which has on average a population of 1,500 people and therefore subsumes a number of postcodes.

The CVA model made substantial use of the data gathered from the PLASC, although some information, such as data collected on absences and exclusions, was not included in the CVA model because although it would improve the explanatory power of the model, 'schools should to some extent be responsible for these factors' (Ray 2006: 21). This illustrates the tension between the accountability and improvement agendas highlighted earlier. It would be useful, for school improvement purposes, for schools to have a measure of the mean impact of an increase in absence by a single percentage point, or an extra day of fixed term exclusion, but inclusion of such variables would muddy the waters for the use of CVA as an accountability measure of school performance.

The model for the CVA pilot included variables listed in Table 4below and multilevel modelling was employed to account for the clustered nature of school data. The majority of the variables were pupil level factors but two additional school level factors were included (the mean and the spread of prior attainment of the school cohort) to capture peer effects.

Pupil Factors

Mean *[intake]* Test Level based on decimalised National Curriculum levels

Subject Variations (take into account differential attainment in Eng, Ma and Sci)

Student’s Gender

Month of Birth

EAL *[English as an Additional Language]*

FSM *[Free School Meal Entitlement]*

SEN *[Special Educational Needs]* Stage or Statemented for SEN

Ethnicity

Mobility (late entry to the Key Stage)

In Care at Current School

Geodemographic Data (IDACI Score)

School Factors

Mean Intake Test Level of cohort

Range (Standard Deviation) of Intake Test Level of cohort

Table 4: Factors included in the DfES/Ofsted CVA multilevel model

The parameters form the 2005 CVA pilot model are summarised in Table 5 below.

Explanatory factor/variable	**Estimate**	**Std. Error**	**p value**
Intercept	147.81	13.20	0.00 **
Prior attainment			
KS2 student APS	-8.55	0.24	0.00 **
KS2 APS (using fine grades) – squared	0.45	0.00	0.00 **
KS2 English PS deviation	1.94	0.07	0.00 **
KS2 Maths PS deviation	-0.32	0.07	0.00 **
Deprivation/SEN			
Does student have FSM?	-21.36	0.30	0.00 **
Deprivation indicator – IDACI score	-65.14	0.70	0.00 **
Does student have SEN - Action Plus?	-64.02	0.42	0.00 **
Special Educational Needs			
Does student have SEN - school action?	-37.91	0.35	0.00 **
Student joined other than Jul/Aug/Sep?	-27.09	0.44	0.00 **
Mobility/Gender/Age/EAL			
Student joined within last 2 yrs?	-74.98	0.67	0.00 **
Gender-Is student female?	15.80	0.20	0.00 **
Age within year	-14.20	0.31	0.00 **
Language-Is English not the student's first language?	23.83	0.65	0.00 **
Ethnic group/in care			
Is the student White Irish?	-0.40	1.48	0.79
Is the student a White Irish traveller?	-43.76	6.84	0.00 **
Is the student White Gypsy/Roma?	-43.05	4.50	0.00 **
Is the student White other?	14.68	0.79	0.00 **
Is the student Mixed White/Black Caribbean?	-1.25	1.05	0.24
Is the student Mixed White/Black African?	4.91	2.19	0.02 *
Is the student Mixed White/Asian?	7.78	1.49	0.00 **
Is the student any other Mixed ethnic group?	6.08	1.09	0.00 **
Is the student Indian?	22.58	0.85	0.00 **
Is the student Pakistani?	24.50	0.91	0.00 **
Is the student Bangladeshi?	30.92	1.27	0.00 **
Is the student any other Asian ethnic group?	27.06	1.41	0.00 **
Is the student Black Caribbean?	17.13	0.84	0.00 **
Is the student Black African?	34.22	1.02	0.00 **
Is the student any other Black ethnic group?	8.07	1.49	0.00 **
Is the student Chinese?	29.01	1.66	0.00 **
Is the student any other ethnic group?	25.44	1.27	0.00 **
Is the student in an unclassified ethnic group?	-11.82	0.60	0.00 **
In care-Has the student ever been in care at this school?	-32.85	1.35	0.00 **
School level factors (Peer Effects)			
Level of school prior attainment School KS3 APS (using fine grades) for CVA	3.04	0.36	0.00 **
Spread of school prior attainment std dev of KS3 APS	-5.45	0.95	0.00 **
Random components:	**Estimate**	**Std.Error**	
Between school variance	351.16	9.63	
Within school variance	4444.83	8.51	
Variance partition coefficient	0.07		
Overall model parameters	-2*log likelihood = 6168797	n = 548,222 pupils	

Table 5: Output of the 2005 pilot CVA multilevel model showing parameter estimates, standard errors and significance values.(Modified from Ray 2006: 37-38)

Inspection of the model parameters shows that the vast majority of parameters included in the model are significant either at the 99% significance level ($p<0.01$) or the 95% level ($p<0.05$). The two non-significant ethnic origin parameters in the model were retained for political reasons (*ibid*: 42-43).

The estimates for the random components of the multilevel model show how the variance in the data is partitioned at the student level (within school variance) and the school level (between school variance). This shows that only 7% of the total variance is at the school level.

The value of the -2*log likelihood parameter is a measure of the fit of the model. This was compared with a simpler, prior attainment factor only model (a multilevel VA model) which had a -2*log likelihood = 6251331 (*ibid*: 41), larger in value than the parameter estimate for the multilevel CVA model suggesting a poorer fit to the data. Clearly the addition of contextualising factors explain more of the variance. This was demonstrated by inspecting the R^2 values for OLS versions of both the models which showed that the prior-attainment only model explains 49% of the variance, whereas the CVA model explains 57% (*op cit*). Interestingly the simpler multilevel VA model showed that 11% of the variance was at the school level (*ibid*: 42) so the addition of contextualising variables, almost entirely at the student level, reduces the variance between schools proportionately more than the variance within schools.

In line with the pilot approach employed in the development of published median VA scores a pilot group of over 350 secondary schools were selected in 2005 to form a CVA pilot group and CVA scores for these schools were made publicly available the following year (DfES 2006). This presented the opportunity to carry out an analysis, after Gorard (2006), of the association between the old GCSE benchmark raw attainment measure (%5+A*-C grades) against both the old VA and the new CVA measures for the CVA pilot schools (Downey & Kelly 2007). Figure 10 below shows the association between the old raw attainment benchmark and the schools' KS2-4VA scores (N = 370). The Pearson correlation for the two measures was r = 0.77. This is similar to Gorard's analysis (r = 0.84) of the same association for the national set of 2004 school data shown in Figure 9 above.

When a similar association is examined between the benchmark raw attainment measure and the new school CVA score (Figure 11) the Pearson correlation coefficient was r = 0.37. These results suggest that CVA was producing a different picture of the performance of schools.

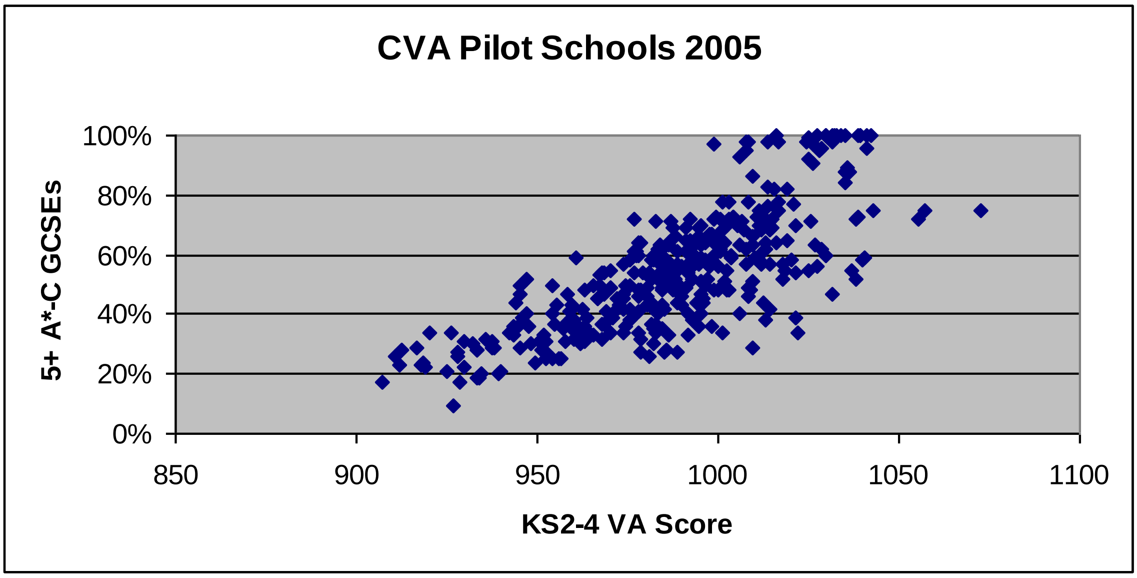


Figure 11: A ‘Gorard type’ plot of KS2-4 value-added score against the raw attainment measure of percentage of students attaining 5+ A*-C grades at GCSE for the 2005 CVA pilot schools. Source data:(DfES 2006)

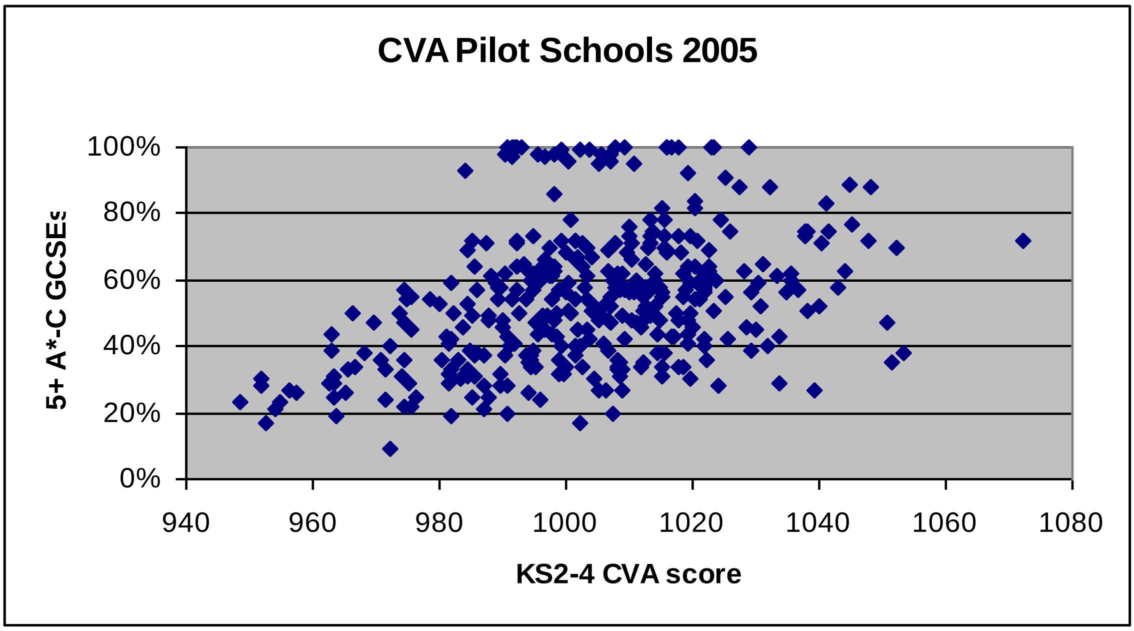


Figure 12: A ‘Gorard type’ plot of KS2-4 contextualised value-added score against the raw attainment measure (%5+ A*-C grades at GCSE) for the 2005 CVA pilot schools. Source data:(DfES 2006)

Despite the development of value-added measures moving increasingly towards the complex contextualised multilevel models called for here the criticism of published performance tables for schools in England has continued unabated. Harvey Goldstein, whose work has clearly influenced the development of the current CVA methodology, has been one of the most vocal and longstanding critics of what he continues to call ‘league tables’ (Goldstein 2007). In a wide ranging critique of pupil and school performance measures and the plethora of uses to which it is applied (Goldstein 2001) he discusses the benefits and limitations of the basic value-added systems that were being introduced at the time over raw attainment measures. His strongest criticism is reserved for the use of such measures to bolster those “politically driven” (*ibid*: 442) policies related to accountability (teacher performance management, target setting and annual league tables) with no reference to the limitations (especially in terms of statistical uncertainty) and inappropriateness of such practices. He concludes:

> *What is required is a commitment to phasing out current procedures which serve a purpose which is largely politically driven, which is widely viewed as irrelevant and which, in its misleading nature, may be doing fundamental harm to education.*
>
> (*op cit*)

Three years later, after the basic value-added measure had become more established, Goldstein is still highly critical of the use of these VA scores for the purposes of public accountability (Myers & Goldstein 2004). “…the best use of value-added comparisons is for LEAs and schools to provide additional information, in confidence, about school performance, set alongside, and not dominating, other factors, especially when disaggregated to individual school subjects or departments.”

Another three years on, after the publication of school KS2-4 CVA scores as measures of secondary school progress Goldstein (2007) remains critical, although perhaps, in this piece, less strident in tone. He acknowledges the inclusion of key contextualising factors in the model and also the publication of 95% confidence intervals in the official DCSF tables. Despite these advances Goldstein states that “there remain considerable problems” (*ibid*: 4). For Goldstein these problems centre around the continued publication of value-added measures in the public domain. He refers to a “league table culture” (*ibid*: 5) that is gives rise to “…a surface precision associated with numerical data, is used, sometimes unscrupulously, sometimes in ignorance, as a substitute for serious and well-informed debate.” (*op cit*).

Part of this critique centres on the view that any single measure of school effectiveness can capture the performance of an individual school. A decade before the first value-added pilot Goldstein and Cuttance (1988) argued that a comparison of school averages

> *tells us nothing about the relative achievements of different types of pupils within the schools… Consequently schools which perform well relative to other schools for the average pupil in the population may perform less well for disadvantaged or advantaged pupils.*
>
> (*ibid*: 198).

This view was reiterated with a more methodological slant (Nuttall et al. 1989):

> *…school effectiveness varies in terms of the relative performance of different sub-groups. To attempt to summarise school differences, even after adjusting for intake, sex and ethnic background of the students and fixed characteristics of the schools, in a single quantity is misleading...[T]he concept of overall effectiveness is not useful."*
>
> (*ibid*: 775–776)

The school effectiveness researchers have also strongly criticised the public ranking of schools based on value-added analyses. Goldstein and Cuttance (1988) contend that the specific method employed to calculate value-added measures will, at least in part, determine the rank order. Saunders (1999) elaborates on this as covering both the choice of outcomes used (to differentiate between different groups of students for example) as well as the type of statistical methodology employed. As Saunders notes, there have been calls for contextualised VA models to include as many factors as possible that might be construed as affecting a school's performance. This, Saunders argues, whilst understandable, fails to appreciate the differences in purpose between school effectiveness research and value-added analysis, the latter of which she says, is essentially "eliminative rather than an accumulative" (*ibid*: 249), aiming to eliminate those factors that are extraneous, part of the 'noise', from the analysis.

CVA scores are presented using a similar approach to that adopted with median VA scores, namely addition of the CVA score (positive or negative) to 100.0 for CVA up to KS2 and to 1000.0 for CVA up to KS4 (GCSE). Rather than the school score being calculated as the mean of the individual student scores (as for VA and the RGA approach suggested by Fitz-Gibbon (1997)) CVA residuals for each school as part of the multilevel modelling output. These residuals are adjusted by incorporating the shrinkage factor referred to earlier and corresponding 95% confidence intervals are calculated for every school score so that a judgement can be made as to whether the school score is significantly above, in line with, or significantly below the national mean CVA score. There is no longer any need for benchmarking of scores as both prior attainment and an extensive range of contextualising variables have been accounted for. In keeping with the use of percentiles in the benchmarking tables from the PANDA reports schools are given a *percentile rank* (a scale inversely proportional to percentiles with a percentile rank of 1 being highest and 100 lowest).

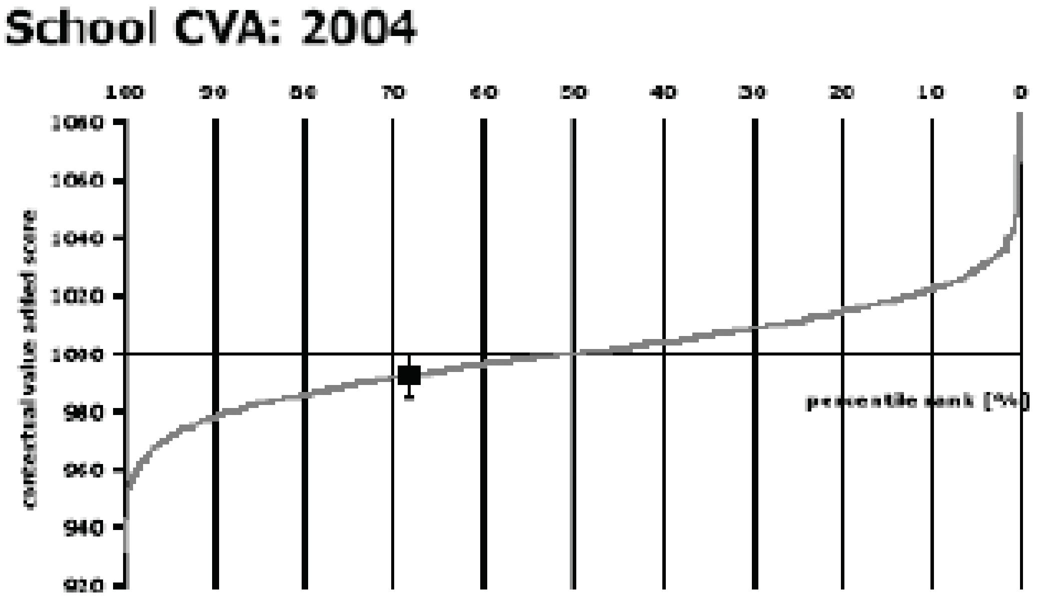


Figure 13: A typical presentation of CVA score (with 95% confidence interval) and associated percentile rank. (Source: Ray 2006: 63)

Early investigators of CVA models based on multilevel models (Sammons & Smees 1998; Thomas & Mortimore 1996) had warned against the use of MLM residuals to produce league tables. They stressed the importance of applying confidence limits when interpreting school residuals.

> *Only schools in which, taking account of intake, results are significantly (p< 0.05) better or significantly worse than predicted on the basis of intake relationships calculated for the whole sample can be distinguished.*
>
> (Sammons & Smees 1998: 400).

As Sammons and Smees suggest, the choice of 95% confidence interval is an arbitrary one, and even though schools are allocated to three categories of significance (significantly lower than predicted, no significant difference and significantly higher than predicted) there is obviously a region close to the national mean value where a small change in the confidence limit will produce a crucial change in the designation of significance. It is likely that a number of schools in the 'no significant difference' category will have confidence intervals that overlap with schools in the ‘significantly above the mean’ or ‘significantly below the mean’ categories making it difficult to distinguish between these schools. Understanding the difference between ‘a significant difference’ in score and ‘a change of statistical significance’ renders the interpretation of residuals a fairly fraught (aside from the fact that they are loaded expressions in terms of colloquial usage of the terms related to significance). Use of such indicators in the public domain set high stakes for schools, their staff and their students. Sammons and Smee’s research on the use of baseline assessments on pre-school children to give indications of progress at KS1, commissioned by Surrey LA, was carried out under a strict code of conduct ensuring voluntary participation, confidentiality of individual schools’ results, no ranking of schools by the LA based on the value-added analysis and that the participating schools would not use their value-added results for marketing and publicity purposes. Within the last decade things have clearly moved from this tentative approach to the use and interpretation of VA/CVA scores which are now used well beyond the purposes of internal self-evaluation reported by Sammons and Smees, informing such varied and wide ranging decisions such as local and national government funding allocations, inspection judgements, performance management of teachers and of course, the initial driver for the publication of such performance indicators and, the original purpose of published school performance indicators, to inform parental choice of schools.

After running their contextualised value added model Sammons and Smees concluded that less than 5% of the total number of schools in their sample were significantly below estimated KS1 outcomes for all four assessments (reading, writing, mathematics and science) and less than 5% were significantly above across the four assessments. They also found statistically significant evidence of differential effectiveness between schools based on gender and FSM eligibility. Whilst their sample was self-selecting due to the voluntary participation clause in the code of conduct, this suggests the use of a single measure to capture the effectiveness of schools across a range of outcomes hides a great deal of fine detail which, as raised a decade before by other school effectiveness researchers (Goldstein & Cuttance 1988; Nuttall et al. 1989) questions the utility of a single school (C)VA score as a performance indicator. This clearly has implications for the use of whole school value-added measures in any evaluation of school improvement initiatives such as SEAL. Even when such measures are calculated with some of the most sophisticated statistical methodology available, one single indicator, even when used as part of a longitudinal analysis of trends in school performance, cannot be the exclusive metric of successful school improvement.

Issues of data hierarchy and shrinkage

As Multilevel modelling is a technique that deals with the fact that not all pupils are "islands" but that the data appears to be clustered into groups of pupils that are more similar than those who are members of a different cluster. Thus it is unsatisfactory to produce a single regression line by ordinary least squares (OLS) regression for a large data set and then relate all pupils to this single line calculating a residual that is a measure of how far the individual pupil's data differs from the regression line. OLS regression does not take into account the clustered nature of the dataset and tends to underestimate the standard error associated with each parameter estimated by the model.

The standard approach developed in recent decades of school effectiveness research has been to acknowledge the hierarchical arrangement of the school system where all pupils are nested within schools. Generally speaking pupils within the same school are more similar in terms of outcomes than pupils in different schools. Multilevel modelling allows for this clustering effect by dividing the variation in the data into a school level residual and pupil level residual.

Multilevel modelling, unlike the OLS method, offers a more complex set of options that take into account the data-structuring fact that pupils are grouped by school. One of the substantive advantages of MLM with such data is that it thus produces more robust estimates of the standard errors for factors in the model, whereas OLS tends to underestimate them. This means that the judgement of whether a factor is deemed to be statistically significant (say at the 95% level) is correspondingly more robust. While such judgements of significance are the bread and butter of substantive school effectiveness research, it is not clear how such judgements play a role in deciding the method of choice in calculating CVA scores, particularly when other influences play a major role in deciding whether factors are included or excluded from the model, regardless of their statistical significance.

Some recent research as (Luyten 2003; Sharp 2006) on the effect of introducing the class as a level in multilevel models (either as an intermediate level between student and school or as the higher level itself) has generated some important insights into the claim that the school is not the most appropriate level use to adjust for clustering in school data. Sharp (2006) researched the progress of students during the first year of formal schooling in Edinburgh Primary schools. The study involved a dataset consisting of baseline and end of year progress scores in numeracy and literacy for 3,569 pupils grouped in 161 classes within 101 schools. The study gave Sharp the opportunity to investigate some key methodological issues associated with the statistical models used to calculate VA progress measures. The nature of the dataset enabled Sharp to construct a 3 level MLM with students nested within classes nested within schools. One limitation was that approximately half of the schools had only one class thus conflating the upper two levels for those schools. Therefore Sharp also produced two level MLMs (using the same MLwiN software used to run the CVA MLM) with students nested within schools and students nested within classes. Finally Sharp compared the residuals calculated in these analyses with those produced via an OLS regression model (using SPSS software) in which class and school level residuals are calculated as a mean of the student residuals contained in the within the level, a form of residual gain analysis (RGA). Thus issues related to the inclusion of an intermediate level (the class), a potential source of useful information for school improvement purposes, could be explored together with issues surrounding the interpretation of OLS vs MLM residuals. The end of year progress scores for literacy demonstrated a pronounced skew towards the upper end of the range so they were transformed so that they were normally distributed with a mean of zero and a variance equal to one.

Sharp produced a MLM in which both the intercept and slopes were allowed to vary, an analysis which would show if schools were differentially effective across the range of prior attainment. He found that the slope residuals were significant (at the 95% level) for numeracy but non-significant for literacy. Although statistically significant in the case of the numeracy dataset, the differences between the school slopes were small in comparison to the intercept differences in the fixed slopes model so Sharp decided to restrict the main models so that they had fixed slopes (this is the current procedure used for the calculation of CVA residuals by the DCSF).

Sharp found that school effects were greater than those produced by classes. In other words the variation between schools was greater than the variation between classes within schools (nearly twice as much for numeracy and almost three times for literacy). When the single class schools were omitted this finding remained true. These findings were also confirmed by carrying out a one-way ANOVA (with school as the grouping variable) on the class level residuals from the MLM with classes as the level 2 units.

Sharp reports that this finding is in contrast to the research literature summary produced by Luyten (2003) who concluded that teacher/class/grade (year group) effects are larger than school effects from a review of a number of studies. Luyten himself seems a little more reserved in his conclusions, especially as studies involving parallel classes (where confounders such as age, subject and curriculum content are controlled) proved far less conclusive than studies involving differences between subjects or grades (year groups). It is interesting itself to note that Luyten was only able to identify 16 studies (ranging from 1987-2001) that enabled him to investigate the differences between school effects and those at an intermediate level between that of the school and the students. This was out of what Luyten considered to be hundreds of school effectiveness studies during that period (*ibid*: 37)

Sharp's results are also in contrast to his own earlier work on literacy progress during the first year in Edinburgh Primary schools (Sharp & Croxford 2003). Here the authors found that class effects were slightly greater than school effects in a multilevel analysis of 2583 students in 69 Edinburgh Primary schools. The MLMs employed involved the use of prior attainment and a number of contextualising factors including entitlement to FSM, EAL and SEN. In this study some schools were found to have widely differing class level residuals suggesting that the effectiveness of classes/teachers may vary more within schools than between classes (meaning that "averagely" performing schools are not necessarily made up of "averagely" performing classes). Sharp & Croxford performed a one-way ANOVA, similar to that described above, on the class level residuals grouped at the school level. This showed the within school variance to be twice that of the between school variance (*ibid*:224), a reverse of the effect shown in the later study described above. It seems there remains more work to be done to gather evidence for the effect of teachers/classes compared with schools and the value of adding the intermediate level in MLMs of school effectiveness. This also suggests that longitudinal studies may produce more definitive information regarding the size of class vs school effects due to the limitations cross-sectional studies like this. Looking south of the border to England Sharp & Croxford suggest that their results call into question the publication of school performance data to aid parental choice of schools (a practice that never occurred in Scotland) as parents have no control over the allocation of teachers to classes. They highlight the potential power of 3 level models to provide information to assist schools and LAs in the process of school improvement but call for caution in the development and application of models analysing at the class/teacher level urging that a range of qualitative and quantitative evidence be used to inform professional judgments in a 'spirit of trust and co-operation', "a spirit which would be highly unlikely to survive the publication of the evaluation process." (*ibid*: 230)

As described earlier, accounting for the clustering effect observed in hierarchical datasets produces more robust estimates of standard errors which will reduce the chance of making Type I errors as OLS techniques will underestimate the standard errors and so effects will be judged to be significant when this in fact not the case. MLMs partition the variance between the different levels in the data hierarchy allowing inferences to be drawn at the level of analysis (class or school) rather than analysing at the student level and drawing conclusions from aggregated values calculated at student level (what is known as the 'ecological fallacy'). These benefits need to be weighed up alongside the desired elements of comprehensibility and interpretability for the outputs of VA models.

Another issue raised by Sharp (2006) is that of interpretation of MLM residuals in comparison to those generated by OLS analyses. Sharp's analysis makes it clear that, even in the setting where schools were found to have more of an effect than classes, the omission of the intermediate level from the model has a significant impact on the size of the variance between schools reducing the variance between schools from 34% to 24% for numeracy data and 37% to 33% for the literacy data. There is also an effect, albeit smaller, on the variance between students. Such findings are in line with those of other researchers investigating the effects of including intermediate levels in MLMs (Noortgate et al. 2005; Opdenakker & Van Damme 2000). One of the important knock-on effects of the change in variance values is a change in the number of schools judged to be significantly above or below the mean and therefore judged as high or low performing schools. Using a two level MLM (students nested within schools), for the numeracy data the number of schools significantly below the mean was 31 out of 99 and above the mean were 33 schools. When classes were introduced as an intermediate level this reduced to just 5 schools below the mean and 8 schools above (*ibid*: 339 – the results for the literacy data are less extreme but the number of schools at the significance thresholds is at least halved by the introduction of classes as an intermediate level). Now it is clear that schools are, to a certain degree at least, responsible for the variation at the class level as well as at the level of the school. What is striking from these results is that, for a study in which the school was seen to have a greater effect than the class, when the class is accounted for as a level in the model, it becomes much harder to differentiate between schools in terms of value-added progress scores. To put it another way, what these data seem to call into question, as hinted at by Sharp's earlier analysis (Sharp & Croxford 2003), is the ability of school level residuals (of any kind) to inform parental choice. Not only do such residuals fail to capture any indication (currently) of differential effectiveness of schools across the prior ability range, between various sub-groups of students or across subjects, they also "hide" the variation between classes/teachers. This surely provides further evidence that the value of any single numerical score as a VA measure of school effectiveness is at best severely limited, but potentially considerably misleading. When the then DfES published value-added scores for primary schools for the first time in December 2003 the Schools Standards Minister, David Miliband, stated

> *We have always said that we will listen to the views of heads, teachers and parents about how the performance tables can provide a more comprehensive and rounded picture of school performance. Including value added information does just that. It shows the rates of progress that children make between 7 and 11 in different schools.*
>
> (GNN 2003)

It may be that Miliband's use of the term 'rounded' carried more meaning that he thought at the time. This one-number-fits-all approach to publicising school performance surely carries with it 'rounding error' of a most severe kind and does far less to inform parental choice than its proponents claim. Whether or not the published value-added scores are accompanied by confidence intervals (and they were certainly not at the time), the publication of VA scores in a form such as "98.7" (or even the true underlying value of the school residual, -1.3) suggests a degree of precision and 'confidence' in measurement of 'school performance' that the statistical methodology simply won't support.

The reality of schools is clearly far more complex than even the DCSF's latest two-level, random intercept, contextually adjusted value-added model can capture. Of course, any attempt to capture more of this complex reality would require, as Ray (2006) is quick to point out, far more complex models and measures, or, at very least, multiple measures for each school. This does not always serve the aim of interpretability, especially for a public audience, and possibly also for the professional audience.

Sharp (2006) raises the example of class level residuals in his 3-level MLM. These are centred around the school mean rather than the mean of classes in the whole dataset. This may be a useful feature for school self-evaluation as the relative progress made by each class in the school is then apparent, but it can make the interpretation difficult if a class is found to be significantly below the school mean in a school that is significantly above the sample mean. Is such a class still making good progress, even though other classes in the same school have made better progress? Also, due presumably to the greater effect of shrinkage on the smaller classes, it is possible for all the class residuals to have the same sign which, as will be illustrated by the work of Thomson (2007) below in the example of pupil subgroup CVA scores from RAISEonline, is counter intuitive.

Sharp's research is based around the progress made by students during the first year of primary school. It is probably fair to say that, in this case, classes can be equated with teachers. This raises the crucial but nonetheless controversial prospect of factoring in the effect of teachers on student progress. This may be a reasonable assumption at this early stage of education as a single class may well spend a substantial amount of time together as a unified group being taught the range of the curriculum by a single teacher. This will not be the case in every school. There may be two-form entry schools where the teachers team-teach the two classes and there will undoubtedly be other models of teaching employed in schools. During the later years of primary schooling, and certainly in secondary education, the concept of the class becomes far more fluid as subject teaching is delivered by 'specialist' teachers and children may be placed in a variety of ability groupings for different subjects and so rarely stay as a unified group throughout the school day. The utility of the class level to explain variance in the data will be diminished. This may well be an argument for ignoring intermediate levels between the student and the school but such a judgement would need to be supported with further evidence, possibly employing more advanced techniques such as multiple-membership and cross-classified multilevel models (Goldstein 2003) that allow for the fact that students will be members of a number of classes within their school and may also change classes over the period of focus. In one study using a two year data set of 4,500 students from 36 Scottish secondary schools (Thomas 2001) the impact of the class on outcomes was found to differ between subjects explaining more of the variance in maths outcomes than English and in maths the class level variance was similar to that at the level of the school (just over 5%). Thomas found that the resulting school residuals for the three level model were highly correlated (r>=0.99, *ibid*: 311) with those from two level models and suggests that there is "little difference in interpretation" (*op cit*) between the two sets of residuals for this particular data set.

Sharp compares the OLS residuals that his findings residual gain analysis (RGA) based on OLS residuals for schools and classes may well be easier to interpret as they are weighted averages of the student residuals from which they are comprised. Thus they are always with reference to a common mean whether at the school or class. Sharps analysis showed correlations very close to 1 (the lowest value was 0.991 – *ibid*: 340) for the RGA residuals calculated via OLS and the class and school level residuals in his two-level models. This was despite the fact that the class sizes ranged from 6-30 with an inter-quartile range of 17-26 which would have resulted in different shrinkage factors being applied to the classes. Thus for situations where a two level fixed slope MLM in the model of choice (such as the DSCF's 2006 CVA model) Sharp concludes that "the ease with which residual gain analysis can be applied and interpreted may make it a more flexible and straightforward way of reaching essentially the same results as are provided by ML modelling." (*ibid*: 345)

Another difference between OLS and MLM is in the way the latter 'shrinks' the value-added estimates, which depend, in part, on the size of the school: smaller schools are 'shrunk' towards the national mean (Ray, 2006: 47). Application of the shrinkage factor means that the CVA score is reduced to a percentage of its 'raw' size, closer to the national mean. OLS also has a problem dealing with small cohorts so that a small school's value-added score can only be given with a wide confidence interval. With MLM, the estimate is calculated from both the school's pupils and from the national data, which is then used to modify the estimate when robust information on the school is limited because of size. This process of shrinkage towards the mean has been described by Kreft (1996) as akin to small schools 'borrowing strength' from the data of larger schools. The resulting shrunken CVA scores generated by the MLM thus have narrower confidence intervals, and these are now given in Performance Tables (for KS2-4). Shrinkage also prevents schools at the extremes - those with residuals which suggest that they are (relatively) very effective or very ineffective - from registering a very high or a very low CVA score. Supporters of MLM say that this is not problematic because the raw residuals for small schools are not known to be good estimates of effectiveness from one year to the next, but this is surely an argument against modelling added value in the first place. Indeed, Fitz-Gibbon (1991) quotes Raudenbush, one of the early developers of MLM in the field of school effectiveness, as saying that shrinkage causes school effectiveness scores to be pulled "in the socially expected direction, demonstrating a kind of statistical self-fulfilling prophecy." (ibid: 19)

Also it seems strange the Ofsted provides its inspectors with the means of unshrinking the data to see what the raw residual looks like. One would be tempted to ask if inspectors can judge from raw data and formulate from their own impressions of the school 'the extent to which the raw residual' is 'an accurate reflection' of the school's effectiveness, then why shrink the CVA residuals? The answer lies in the school improvement / public accountability tension. The application of the shrinkage factor seems to offer further evidence that CVA is more about public accountability for Performance Tables than anything Ofsted espouses in its New Relationship with Schools about a school centred, self-evaluation.

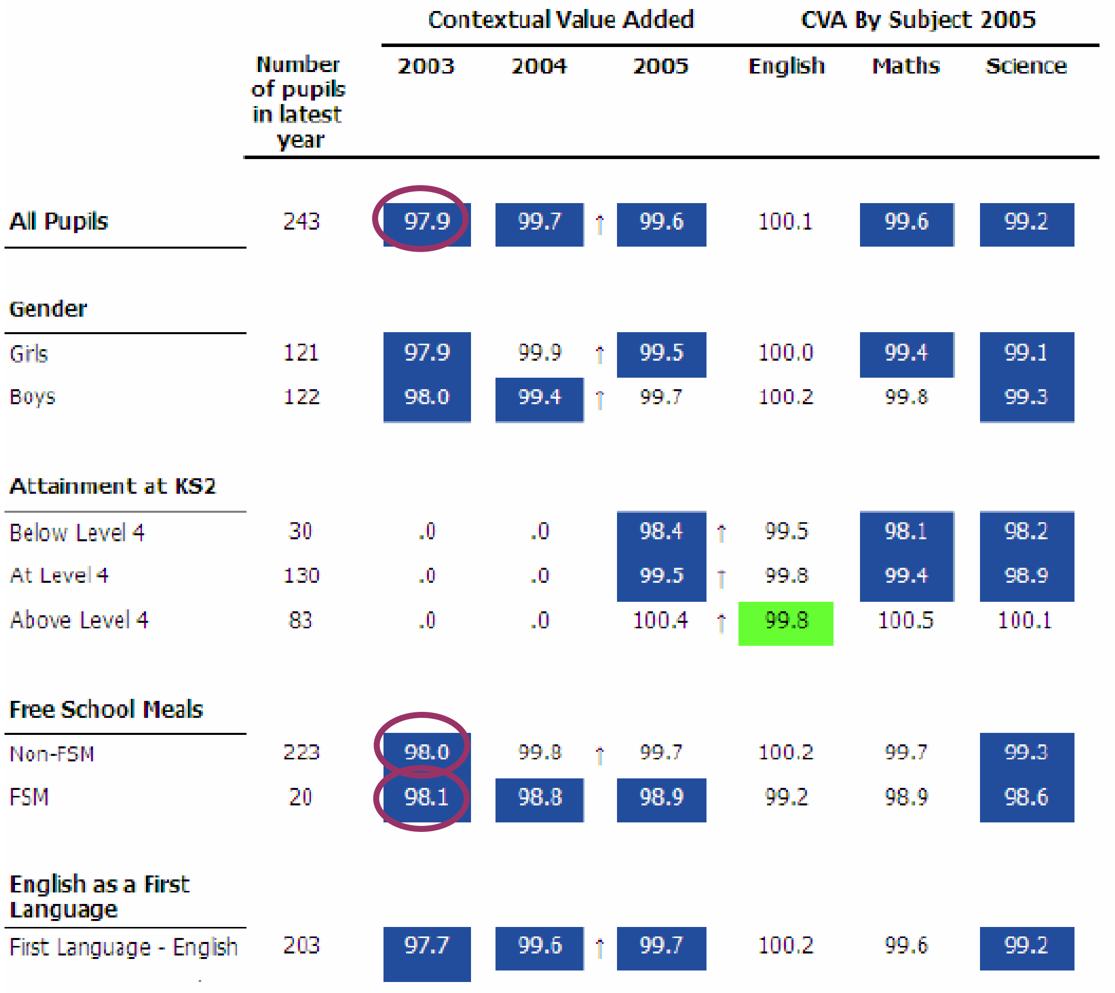
Fitz-Gibbon (1991) made an interesting distinction in delineating the hierarchy in the data set, namely, that students are nested within departments rather than within schools, as the purpose of the ALIS system was to provide feedback to teaching departments and not to provide an indicator of school performance. In this study estimates were calculated for a series of outcomes including A-level grade and attitude measures using a model based on prior attainment, gender as pupil level factors the two school level factors; the mean values of the prior attainment and a SES proxy variable (the occupation of the head of the household). The resulting multilevel modelling (MLM) residuals were compared with those calculated using the OLS regression model that was employed at the time in the ALIS analyses. The largest differences in OLS and MLM residuals were observed for departments with small samples (small numbers of students studying a particular subject at A-level). Fitz-Gibbon attributes the key difference in residuals to the effect of the *shrinkage factor* that is applied to MLM residuals. The uptake of A-level subjects in schools can be highly variable from subject to subject and school to school (in comparison with FE and 6th From Colleges where numbers tend to be higher and more stable). Since the nature of ALIS was to report progress to departments Fitz-Gibbon argues that the effect of the shrinkage factor would have made both benchmarking between departments more difficult with departments having to restrict their comparisons not just to those with similar characteristics but also to those of a similar size. It would also make year-on-year comparisons more difficult due to the fluctuations in uptake from year to year.

Proponents of MLM would argue that this is exactly the benefit of shrinkage, that by borrowing strength from whole dataset such variations are accounted for in the analysis giving residuals that are more stable to the effects variable sample size in the level 2 units, both within the annual dataset when making comparisons between departments/schools, and also when comparing the values of a school's residuals from year to year in a longitudinal study.

Variability in sample size is less of an issue for school longitudinal studies of school performance as school admission numbers tend to remain relatively similar over time. However, there is an issue of sample size when comparing schools due to the inherent error in small sample sizes and therefore large confidence intervals. This affects any type of value-added residual but shrinkage compounds the issue for CVA scores calculated using MLM. Application of shrinkage factor makes it possible for schools with the same value of raw residual to have different CVA scores. The effect of pulling the scores towards the mean value will make it difficult for small schools to achieve significantly high CVA scores and correspondingly very small schools are less likely to have CVA scores significantly below the mean. In England the issue of sample size is particularly acute for primary and first schools, a number of which, particularly in rural areas, have very small numbers of students in each year group. Current DCSF practice excludes schools with 10 or fewer students in a cohort (or those where less than 50% of the students can be matched to the data for calculation of value-added) from being reported in Performance Tables of value-added. Commentators have suggested that this figure is too low. Fitz-Gibbon herself (1997) suggested 30 as the minimum cohort size whereas other researchers (Tymms & Dean 2004) have set the limit at 50 students. As Ray concedes “There is clearly a trade-off between statistical reliability and the desire to include data on as many schools as possible.” (2006: 34). Such compromises, though practicable, may call into question the need to employ such statistically robust methods as MLM. When the CVA methodology is adopted in primary schools for calculation of KS1-2 CVA scores for Performance Tables it will be important to know in more detail what the magnitude of the trade-off between inclusivity and reliability is. The issue of sample size becomes an acute one for all schools when CVA scores are ‘sliced and diced’ down to sub-groups of students such as those with a particular class of special educational needs or those from a specific ethnic background.

An example of this is given in Figure 14 below; a sample output form RAISEonline illustrates the confusing effect of shrinkage on sub school level analyse of CVA scores. The 2003 KS1-2 CVA score for all students in the school (N = 243) is given as 97.9. When the scores for students with FSM entitlement (N = 20) and those not entitled to FSM (N = 223) are examined both scores are above the score for all students (98.1 and 98.0 respectively). This is not a rounding issue but rather the differential effect of the application of shrinkage to groups of students of differing sample size. It will be important for schools to be briefed very carefully on the issues involved in the application of shrinkage before they use such analyses for self-evaluation purposes.

Even in large secondary schools the number of students in each of these sub-groups may be small. RAISEonline gives the facility to calculate CVA scores for these sub-groups regardless of sample size. Not surprisingly, such small groups do have very large confidence intervals, making it unlikely that they will be flagged as significantly above or below the mean unless the scores are extreme, however, the application of the shrinkage factor to the residuals from which the scores are derived makes it unlikely that extreme scores will be reported. The value of such information for school self-evaluation is at best limited and at worst misleading. It is hard to understand why such a facility is present within RAISEonline and it is hoped that Ofsted inspectors and LA staff, who also have access to the data, will be fully briefed on the limitations of these analyses. It will be important for schools to be briefed very carefully on the issues involved in the application of shrinkage before they use such analyses for self-evaluation purposes.



		Contextual Value Added					CVA By Subject 2005		
	Number of pupils in latest year	2003	2004		2005		English	Maths	Science
All Pupils	243	97.9	99.7	↑	99.6		100.1	99.6	99.2
Gender									
Girls	121	97.9	99.9	↑	99.5		100.0	99.4	99.1
Boys	122	98.0	99.4	↑	99.7		100.2	99.8	99.3
Attainment at KS2									
Below Level 4	30	.0	.0		98.4	↑	99.5	98.1	98.2
At Level 4	130	.0	.0		99.5	↑	99.8	99.4	98.9
Above Level 4	83	.0	.0		100.4	↑	99.8	100.5	100.1
Free School Meals									
Non-FSM	223	98.0	99.8	↑	99.7		100.2	99.7	99.3
FSM	20	98.1	98.8		98.9		99.2	98.9	98.6
English as a First Language									
First Language - English	203	97.7	99.6	↑	99.7		100.2	99.6	99.2

Figure 14: A sample output from RAISEonline with outputs highlighted to illustrate the potentially confusing effects of applying shrinkage to multilevel residuals.

Such an attempt by Hillingdon Borough Council (Thomson, 2007) to assist its schools in interpreting RAISEonline outputs highlights some of the issues raised when interpreting the student sub-group scores. In one sample output from RAISEonline (p. 41) the CVA score for all matching pupils (182 in total) is given as 978.4. The scores for the 170 students with English as their first language is 978.7 whilst the remaining 12 students who do not have English as their first language had a CVA score of 987.3. This is potentially confusing for schools who would intuitively expect to see the scores for subgroups of students distributed around the mean value for all students. In this case the application of shrinkage will have pulled all three CVA model residuals closer to the mean (1000) but shrinkage has affected the sub-group residuals more due to their smaller sample size. The author's conclusion is that:

> *the statistical model underpinning CVA was not designed to analyse the performance of pupil groups. This gives rise to unreliable scores for some groups. FFT analyses are recommended instead.*
>
> (*ibid*: 76).

If, as argued above, the use of CVA measures for analysis at the student group level is problematic due to potential confusion deriving from the application of shrinkage, then it seems wise to heed the advice of Thomson and the Hillingdon LA to consider the development of Fischer Family Trust value added measures. So at this point in the review we turn to consider the parallel development and key aspects of the Fischer Family Trust suite of value added models.

3.1.2 Fischer Family Trust value-added models

Origins of the FFT and the Performance Data Project

The trust was set up by Mike Fischer, an Oxford graduate and co-founder of Research Machines (RM) who are best known for their provision of information and communication technologies (ICTs) to schools. Initially the trust focused on three projects, one aimed at improving literacy levels in the early years of primary education, one focused on raising the performance of schools in North Islington and the third related to the provision of, according to Fischer, "the best possible performance data" (Hague 2005) to schools and Local Authorities. Since then the trust has expanded its school improvement focus to include further literacy projects and an analysis of the impact of ICTs on learning in schools.

The performance data project is managed by Mike Treadaway, a former ICT teacher and Local Authority (LA) adviser for both Suffolk and Glamorgan. He developed the school performance measures that lay at the heart of the early FFT analyses whilst working with Glamorgan LA. The FFT performance data project began in 2000 and supplied analyses to 55 Local Authorities

(LAs) by 2001. By 2004 the project had extended to include all LAs in England and Wales. The data for the FFT analyses is provided by the Department for Children, Schools and Families (DCSF) and the Welsh Assembly Government.

The Fischer Family Trust website gives more detail on the aim of the performance data project:

> *The project aims to provide analyses and data which help LAs and schools to make more effective use of pupil performance data. Using a database which now contains performance information on over 10 million pupils in England and Wales we have developed a range of analyses to support the processes of self-evaluation and target-setting.*
>
> (FFT, 2007)

These twin purposes of self-evaluation and target-setting are at the heart of the performance measures developed by the FFT and it is important to hold these twin purposes in mind when working through the wealth of training and guidance materials the trust has produced, and in seeking to understand the choice of statistical methods employed by the project's developers. The measures are not designed primarily to be indicators of school performance for public consumption in the way that DCSF value added (VA) and contextualised value added (CVA) measures are (Ray, 2006). They are intended to be a tool to inform school self-evaluation and improvement, and as an aid to setting aspirational targets for academic outcomes at both the student and school level.

As a result of this approach, the data has always been made available to schools via their LA and is not released into the public domain. It is clear from our own experience working both with schools and LAs that the analyses produced by the trust are held in high regard and are widely used to inform school self-evaluation and improvement. This was confirmed by a survey conducted by National Foundation for Educational Research (Halsey *et a*l. 2005) on the impact of the increased emphasis on school self-evaluation as part of Ofsted's New Relationship with Schools (NRwS). The survey was conducted with a group of trial LAs and included 68 schools. FFT data was the most highly regarded external source of data used by the respondent schools (*ibid*:16). 69% of the schools gave FFT data the highest rank of "very helpful" for the purpose of self-evaluation. Only internally produced school data, lesson observations and departmental reviews scored higher. By contrast the next highest scoring external data sources (LA produced data and Ofsted Performance and Assessment Reports or

PANDAs) were scored as "very helpful" by 38% of respondent schools. A primary headteacher summarised this impression of FFT data with the comment that

> *The Fischer Family Trust data is versatile, reliable, and accurate*
>
> (*ibid*:17).

Targets, predictions and estimates

Representatives of FFT will often go to great lengths to explain that the predictive aspect of their models produces *estimates* of pupil attainment not targets or predictions. They state that these estimates should always be used in conjunction with the professional and local knowledge of teachers to inform the process of predicting pupil potential and moving from such predictions to target-setting with individual pupils, or for setting departmental and school level targets for accountability to Governors, the Local Authority and the like. This may seem like semantics to some but FFT representatives have rationalised the differences as follows:

> Estimate + Professional Knowledge → Prediction
>
> Prediction + Challenge → Target
>
> "Targets should be aspirational"
>
> (Spradbery & Cashin, 2006)

The FFT suggested approach to target setting is based on triangulation of FFT estimates with other external sources of data such as Cognitive Ability Tests[1] or MidYIS/Yellis[2] data, together with teacher professional judgement. Delegates at FFT training meetings are urged to use the calculated estimates as an aid to inform planning and delivery of their lessons; identifying pupils who may need extra support, differentiating lesson content for a teaching group, assessing the progress of individual pupils and teaching groups against expectations and

[1] Cognitive ability tests (CATs) are widely utilised in UK schools. The most popular, produced by GL assessment (formally nferNelson). consist of a battery of assessment items that are not curriculum specific which form tests of a student's verbal, non-verbal and quantitative skills. The results are usually adjusted for age and standardised. For more details of GL assessment CATs see: http://shop.nfer-nelson.co.uk/icat/7616main

[2] The Middle Years Information System (MidYIS) and Year Eleven Information System (Yellis) are predictive tools designed to help the setting appropriate academic targets for students at the end of KS3 (MidYIS) and for GCSE (Yellis). They are produced by the Curriculum, Evaluation and Management Centre (CEM) based at the University of Durham. For more details see http://www.cemcentre.org/

highlighting potential underachievement whilst there is time to take remedial action. The language is very much that of development and improvement at all organisational levels of the school, but with a particular focus on the pupil and class levels.

Statistical models used to calculate FFT estimates

FFT produce estimates of pupil attainment for the national tests at the end of Key Stage 2 (KS2) at age 9, Key Stage 3 (KS3) at age 13, and for the General Certificate of Secondary Education examinations (GCSEs) at age 16. These estimates are calculated for every pupil on role in LA schools[3] and the estimates are calculated using two multivariate regression models applied to the attainment of the previous year's pupils.

Early analysis of pupil attainment data by the trust revealed a number of pupil and school level factors that were associated with the level of attainment of pupils in National Curriculum tests and GCSE examinations. The results will hold little surprise to readers familiar either with the literature associated with the development of contextualised value added measures of pupil progress or with the core literature of school effectiveness research.

The key factors identified by FFT that have been incorporated into their predictive statistical models can be summarised as follows.

Pupil level factors

- Prior attainment – based on the assessments given at the end of the previous Key Stage.
- Gender
- Age in the school cohort – children born early in the school year make better progress than their younger, later-born peers.

School context factors

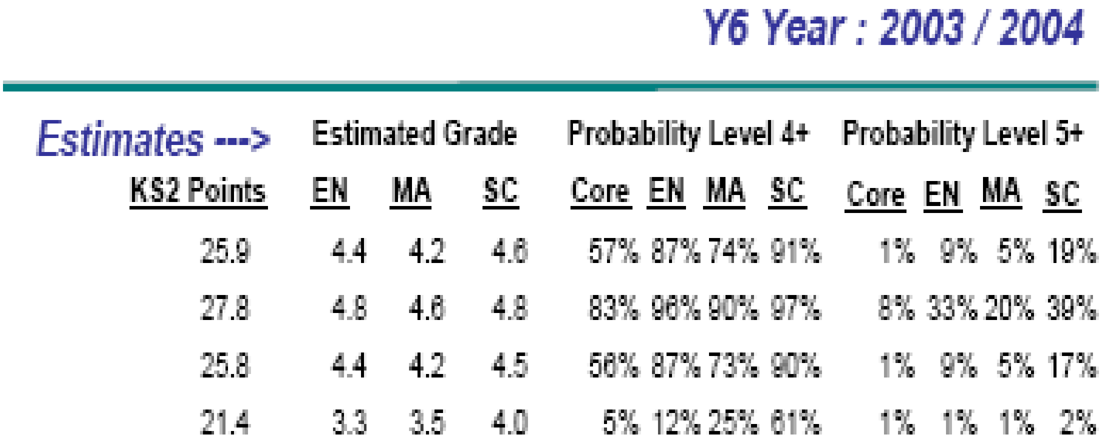
- Social and economic status (SES) – Pupils from disadvantaged backgrounds generally make less progress than their more advantaged peers.
- Peer effects – both the level and spread of prior attainment in a particular school cohort is associated with the level of attainment.

[3] The data is provided for state maintained schools in England and Wales. It is therefore not available for independent schools in England and Wales, nor for schools in Scotland and Northern Ireland.

As a result of these conclusions FFT have developed two models for generating pupil estimates. Their *PA model* (Prior Attainment) employs only the pupil level factors outlined above whereas the *SE model* (Socio-Economic) incorporates both the pupil and school level factors into the model. It can be seen then that neither of these models produce a 'pure' prior attainment only value added measure. The PA model is contextualised to a certain degree by the addition of the pupil level factors of gender and age. The SE model includes the three pupil level factors of the PA model and adds school context factors related to socioeconomic status (SES) and the attainment of peers into the mix. The outputs from the estimate models are as follows:

KS 2 and 3 estimates

- average points scores[4] for all core subjects (English, maths and science)
- an estimate National Curriculum level for each subject expressed as a "fine-grade" decimalised figure
- percentage probability of achieving national attainment benchmarks (levels 4 and 5 at KS2 and levels 5 & 6 at KS3) based on the performance of pupils with similar prior attainment during the previous year.



Y6 Year : 2003 / 2004

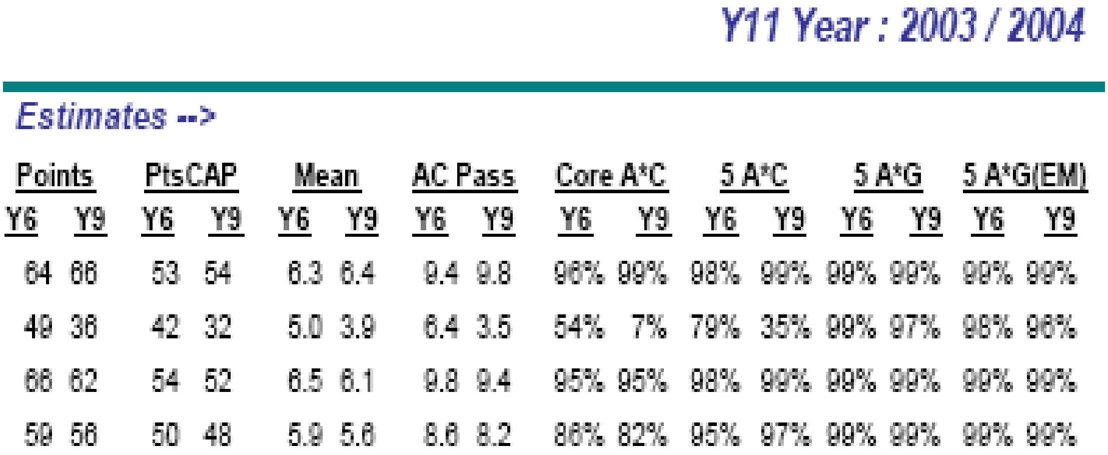
Estimates --->	Estimated Grade			Probability Level 4+				Probability Level 5+			
KS2 Points	EN	MA	SC	Core	EN	MA	SC	Core	EN	MA	SC
25.9	4.4	4.2	4.6	57%	87%	74%	91%	1%	9%	5%	19%
27.8	4.8	4.6	4.8	83%	96%	90%	97%	8%	33%	20%	39%
25.8	4.4	4.2	4.5	56%	87%	73%	90%	1%	9%	5%	17%
21.4	3.3	3.5	4.0	5%	12%	25%	61%	1%	1%	1%	2%

Figure 15: FFT exemplar anonymous KS2 estimates report (Source: FFT website http://www.fischertrust.org/)

[4] Average points score (APS) is a system used to quantify attainment in national curriculum assessments in the 'core' subjects of reading, writing and maths at KS1 and English, maths and science at KS2 & 3. The points score for a subject is calculated from the formula 6L+3 (where L = the NC level in that subject). The simple mean of the three subject scores is then calculated to produce the APS. The national expectation is that it will take more than a year for students to progress from one National Curriculum level to the next so the points score system allows progress within a level to be quantified.

GCSE estimates

- estimates of both total and capped (best 8) points score[5]
- an estimate of the number of GCSE passes at grades A* to C
- percentage probability of achieving national attainment benchmarks (5 or more A*-C grades, 5 or more A*-G grades and 5 or more A*-G including English and maths qualifications) based on performance of pupils with similar prior attainment during the previous year
- in a separate report, specific estimated grades (based on ordinal regression techniques) for a wide range of GCSE subject groups



Y11 Year : 2003 / 2004

Estimates -->

Points		PtsCAP		Mean		AC Pass		Core A*C		5 A*C		5 A*G		5 A*G(EM)	
Y6	Y9	Y6	Y9	Y6	Y9	Y6	Y9	Y6	Y9	Y6	Y9	Y6	Y9	Y6	Y9
64	66	53	54	6.3	6.4	9.4	9.8	96%	99%	98%	99%	99%	99%	99%	99%
49	36	42	32	5.0	3.9	6.4	3.5	54%	7%	79%	35%	99%	97%	96%	96%
66	62	54	52	6.5	6.1	9.8	9.4	95%	95%	98%	99%	99%	99%	99%	99%
59	56	50	48	5.9	5.6	8.6	8.2	86%	82%	95%	97%	99%	99%	99%	99%

Figure 16: FFT exemplar anonymous GCSE estimates report (Source FFT website http://www.fischertrust.org/)

[5] The GCSE points score system allows a quasi-continuous scale to be produced from the ordinal scale of grades A* through to G that are the reported results of GCSE assessments. The original points system, used in the example here, assigned 8 points to an A* grade, 7 points to an A grade, 6 points to a B grade and so on down to G which scores 1 point. The *total points score* is the simple sum of GCSE points in all subjects achieved by the student. The *capped points score* is the total points derived from the best 8 GCSE results for the student. The national qualifications framework at age 16 in England and Wales includes a whole raft of non-GCSE examinations. These assessments are assigned a GCSE points score equivalence so that overall attainment at age 16 can be quantified. Over time the growing portfolio of qualifications has required the point system to be modified so that an A* grade is now worth 58 points, an A grade worth 52 points and so on with the gap between each grade worth 6 points so that a G grade is worth 16 points. The wider range of points on the new system allows for points to be allocated to short courses, worth only half a standard GCSE through to vocational based course worth up to the equivalent of four GCSEs. Interestingly, the shift in points system caused the rankings of schools to shift between 2 and 15 percentile points (Ofsted 2004) due to the greater weighting of lower grades that results from the new system, a point picked up by the BBC at the time (BBC 2004).

Lifting the lid on the estimates models

In the technical summary that follows, much greater attention will be given to factors that incorporate compound measures that might be considered unique to FFT models.

Pupil level factors Prior attainment

FFT estimates models draw on a several measures related to pupil prior attainment.

- mean national curriculum test level (decimalised National Curriculum level)

This is the basic prior attainment measure utilising the score achieved in the three core national curriculum tests at each key stage. The individual pupils' scores, together with level thresholds calculated after marking of National Curriculum tests are used to generate a decimalised National Curriculum (NC) level or "fine grade" which is then used as the primary input variable in a statistical regression model. This is in contrast to earlier valued added measures where a national median line was generated using the much broader full NC levels as the input and output measures of attainment (Ray: 2006). The use of decimalised levels, however, raise questions about the robustness of the National Curriculum assessment framework and the validity of fine grades generated after marking of individual student scripts. The National Curriculum tests were not originally designed to produce such fine grade assessments, although schools are familiar with the practice of subdividing levels into finer bands (normally three bands denoting upper, middle and lower, or "a", "b" and "c" band attainment within the same National Curriculum level). The KS2 estimates, based on prior attainment at KS1, use these broader sublevels as KS1 test scores are not available in the national dataset since KS1 tests are internally marked rather than the external marking process employed for the KS2 and 3 tests.

Whilst idiosyncrasies in the marking process may cancel themselves out over the whole national dataset there will clearly be a certain degree of subjectivity associated with the marking process, especially in subjects such as English. The current practice associated with National Curriculum test marking is that a single marker will assess the test scripts for at the pupils in an individual school so a certain degree of bias is likely to be inherent in the system at the pupil and school levels.

- differences between National Curriculum Test scores and Teacher Assessments

The regression model used to calculate estimates includes a term that incorporates the difference (if any) between the teacher's assessment of national curriculum level and the pupil's performance in the National Curriculum test. FFT data is unique in utilising the results of statutory Teacher Assessments (TAs) as a prior attainment measure in their models. FFT analyses show that the teacher assessments are statistically significant and their addition as inputs produces a better fitting model for the data. The inclusion of teacher assessments may well be one of the reasons why FFT data is held in high regard by practitioners in schools. Where a pupil may have been absent for a National Curriculum test the Teacher Assessment is used to impute a fine grade score based on the median of the fine grades awarded for the same Teacher Assessment level.

- subject differentials

In their early analyses FFT found that attainment in some subjects has a greater influence on future attainment than others. For example, prior attainment in English at KS1 is more influential than maths in estimating attainment at the end of KS2. This is another significant difference that sets FFT estimates apart from similar predictive measures generated by the DCSF/Ofsted for example "Progress Charts" (sometimes known as "Chances Graphs") in their *Autumn Package* materials sent to schools and now available via RAISEonline. The DCSF methodology involves the conversion of National Curriculum levels into the average points score (APS) for the purpose of estimating outcomes at the end of the next key stage with no account taken of subject differentials. This method misses out on the subtle but significant differences in outcomes by pupils with a different mix of levels in the core subjects but with the same average point score on the DCSF measure.

Figure 3.3.3 below illustrates the impact of subject differential. Students with the KS2 levels of 3,4,5 (any combination of subjects English, maths and science) or 4,4,4 have an average KS2 points score of 27. Using the progress charts available via RAISEonline would produce the same range of estimates for each of these children but the subject differences at KS2 NC levels relate to widely differing outcomes at GCSE. 50% of pupils with a combination of 3,4,5 in English maths and science respectively went on to gain 5 or more A*-C grades at GCSE whereas for those pupils who scored 5,4,3 (i.e. English and science levels reversed) 75% achieved 5 or more A*-C passes at GCSE. (FFT, 2004a)

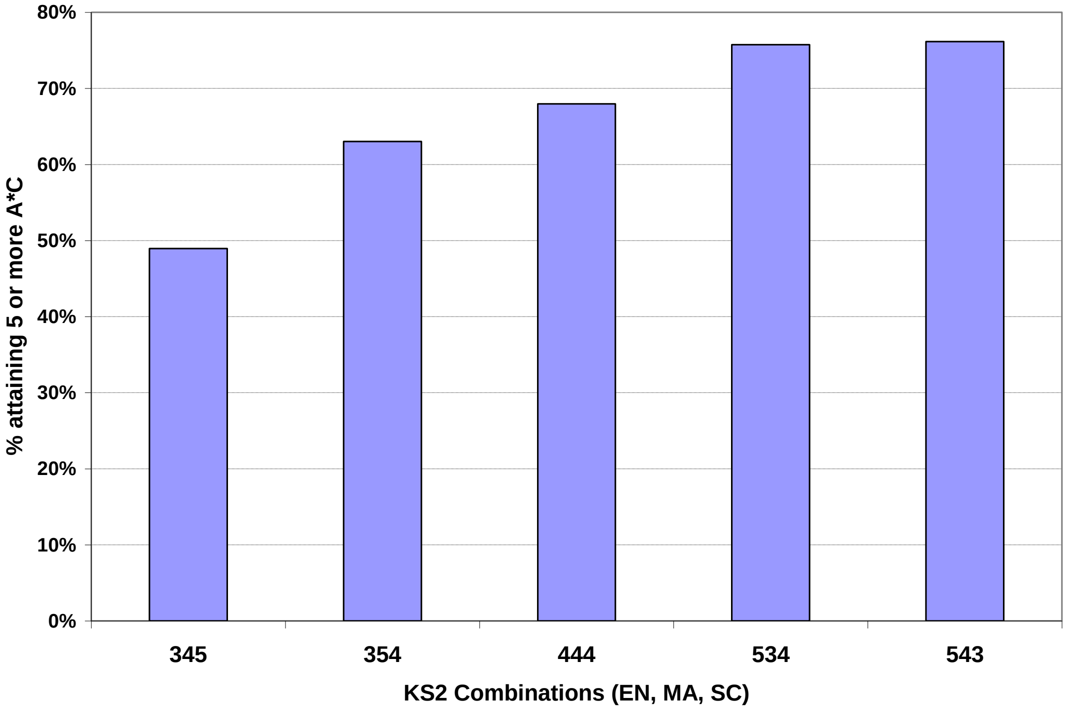


Figure 17: A three year analysis of pupil outcomes carried out by the FFT illustrating the effect of differential performance in the three core National Curriculum subjects. (Source: FFT 2004a)

Gender

There has been much research and media interest in the effect of gender on pupil attainment. Despite several waves of policy initiatives in UK schools the 'gender gap'in achievement, quantified both by raw attainment and value-added measures, proves to be stubbornly resistant to change, leading some researchers to conclude that the source of the gender effect lies outside the realm of school practice (Burgess *et al*: 2004).

FFT models, at all levels, factor in the gender of pupils (as a binary variable) which explains a statistically significant proportion of the variance in the data. Like many contextual variables utilised in estimate models the inclusion of gender as a predictor will act as something of a double-edged sword, lowering expectations for some pupils and raising them for others when compared with pure prior attainment based estimates.

Month of birth

Students born earlier in the school academic year (from September to August in the UK) tend to achieve higher outcomes than their younger peers. This means that a June born pupil with an identical prior attainment score to a November born peer had made better progress than expected during the prior key stage and so estimates of progress by the end of the next key stage would be correspondingly higher.

School level factors present in the FFT SE (School Effects) model

Socioeconomic status (SES)

Originally the only school level SES factor employed by FFT was percentage free school meal (%FSM) entitlement. Since 2004 school level SES measures have been based on both %FSM entitlement and a modified form of the ACORN deprivation measure (the acronym stands for 'A Classification of Residential Neighbourhoods') developed by CACI International[6] (formally known as Consolidated Analysis Centers, Inc) using a mix of 2001 census data and information derived from the company's Consumer Lifestyle Datbases. ACORN is predominantly used as a geo-demographic indicator for private sector company marketing purposes but also being increasingly used by public sector bodies.

ACORN divides the whole of the UK into three levels of deprivation measure. At the top level there are 5 overarching categories of deprivation. These 5 categories are subdivided into a total of 17 groups which are further divided to give a total of 56 types of classification. The measure is linked to postcode and so is a finer measure of deprivation than the IDACI (income deprivation affecting children index) measure used by the DCSF in their CVA measures. IDACI is based on Lower Layer Super Output Area (LLSOA, over 34,000 in the country consisting of at least 1,000 residents with a mean of 1,500 residents per LLSOA see National Statistics (2006) for details).

The FFT performed some preliminary regression analyses of ACORN deprivation (FFT, 2005) against value-added scores. The association between the 17 ACORN groups and a mean value added score based solely on prior attainment for the pupils in each ACORN group was

[6] For more information on the ACORN geo-demographic measure see http://www.caci.co.uk/msd.html

investigated. The 17 ACORN groups (designated by the letters A-Q with Q representing the most deprived group) were then gathered into 4 broader value added groupings where marked transitions occurred between residual VA scores:

Broad VA Group	ACORN Group	VA score range
1	A B C D	>=0.3
2	E F G H I J	<0.2 but >0.0
3	K L M Q	<0.0 but > -0.2
4	N O P	< -0.4

Table 6: VA scores for FFT ACORN deprivation groups (Source: adapted from FFT (2004a & 2005))

Interestingly there was some degree of mismatch which bucked the general trend of negative association between deprivation and value-added score. Closer inspection revealed that these were almost exclusively groups that were associated with a higher proportion of pupils from minority ethnic backgrounds who made more progress than their ACORN deprivation measure suggested. In particular the ACORN group Q (the group representing the highest level of deprivation) was mismatched with the VA scores and placed with less deprived groups in the FFT VA based groupings described above. CACI have titled group Q “Inner City Adversity” and subdivided it into two types which describe on multi-ethnic communities that are 25% black and 14% Asian.

When percentage free school meal entitlement (%FSM) was added to the resulting model both SES factors remained significant with the ACORN *group* explaining slightly more of the variance in VA scores. There was clearly a high degree of colinearity between the ACORN and FSM entitlement with a measured correlation of 0.82 (FFT 2005). For some schools however, FFT reported (2004a) that the percentage FSM entitlement and ACORN score “provide significantly different pictures” of SES. Despite this the decision was made to retain both variables as SES measures of school context in the models.

A more detailed analysis involving the 57 ACORN *types* was then carried out against value added scores representing combined progress of pupils at all key stages (between KS1-2, KS2-3 and KS3-4). Once again, the correlation between the ACORN types (ordinal categories) and value added score was strong with any substantial anomalies due almost exclusively to the presence of higher proportions of minority ethnic pupils in certain ACORN types. As a result of this analysis the ACORN types were re-ordered to match the pattern in VA score and these 57 modified types used as the postcode based geodemographic factor employed in contextualised FFT models. To reflect the non-significant effect of living in an area assigned to the middle set of modified ACORN types (type numbers 20-39 out of 57) these types are all allocated an ACORN score of 30. The FFT analysts noticed that the majority of these types shared the common property that school age pupils made up a relatively low proportion of the population (FFT 2005a).

They also looked at the variation in capped GCSE points score (based on the best 8 GCSE results for each pupil) with school FSM entitlement and found less variation between capped GCSE points scores across the range of school FSM entitlement for all minority ethnic groups except Caribbean (Figure 18). This would suggest that deprivation has a differential impact on pupil progress depending on their ethnic background with the impact of deprivation greatest on pupils from Caribbean and White British ethnic backgrounds. As a result of this analysis interaction terms between percentage school FSM entitlement (as a measure of SES) and ethnicity were included in FFT fully contextualised models (see later section on the SX model) as these were found to be statistically significant in the regression analysis. Such interaction terms have also been included in the DCSF CVA model developed since the 2005 CVA pilot[7] .

[7] see http://www.standards.dfes.gov.uk/performance/1316367/CVAinPAT2005/?version=1

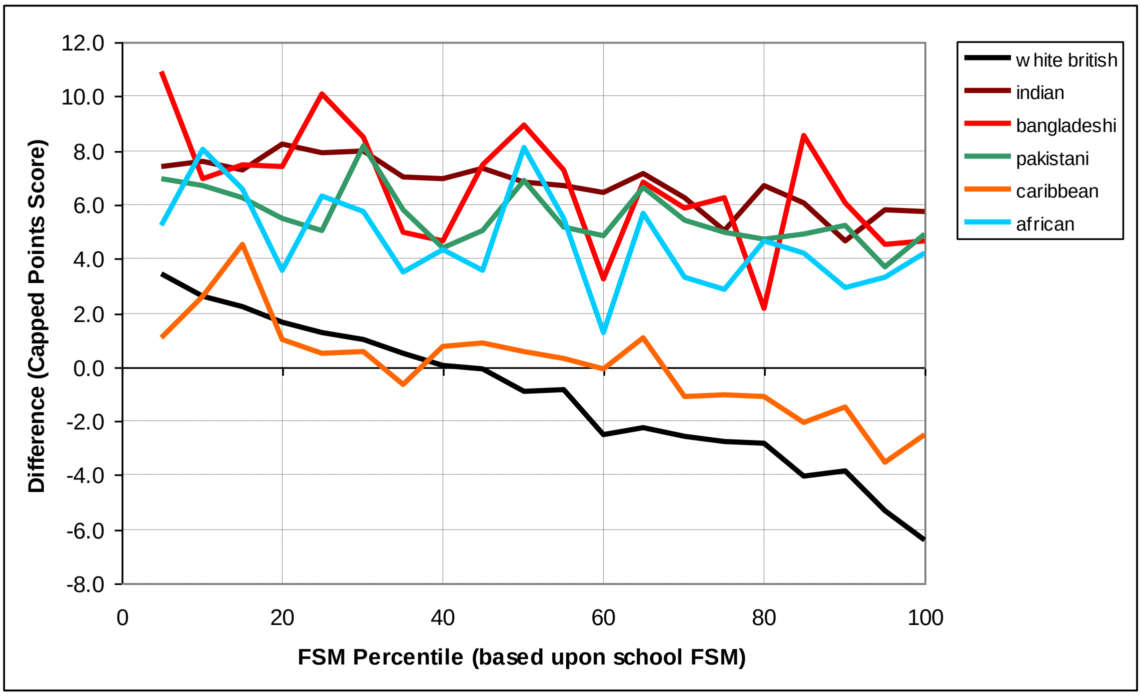


Figure 18: Association of %FSM entitlement with difference in capped GCSE points score for a variety of ethnic backgrounds (Source FFT 2005a)

Another key finding from the analyses of ACORN and FSM data was the regional variation in the proportion of pupils entitled to FSM in each of the four broad ACORN VA groups (FFT 2004a). This was particularly the case for schools in London where inner London schools had a disproportionately high proportion of children with FSM entitlement in the ACORN VA bands with lower deprivation (Bands 1, 2 and 3 in the table above).

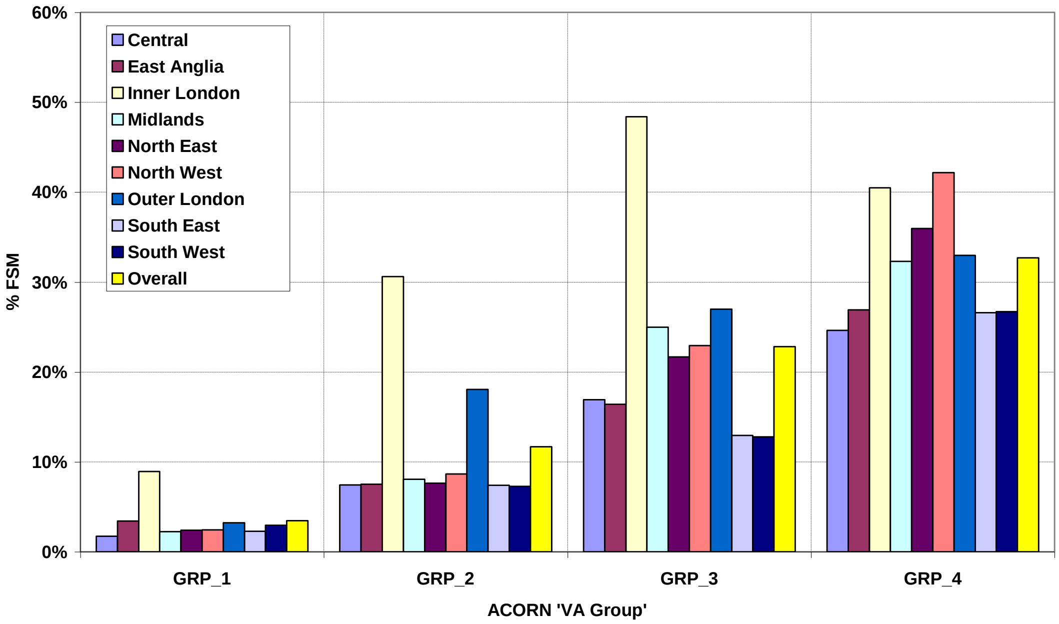


Figure 19: Distribution of FSM entitlement for the FFT ACORN VA groups (Source: FFT 2004a)

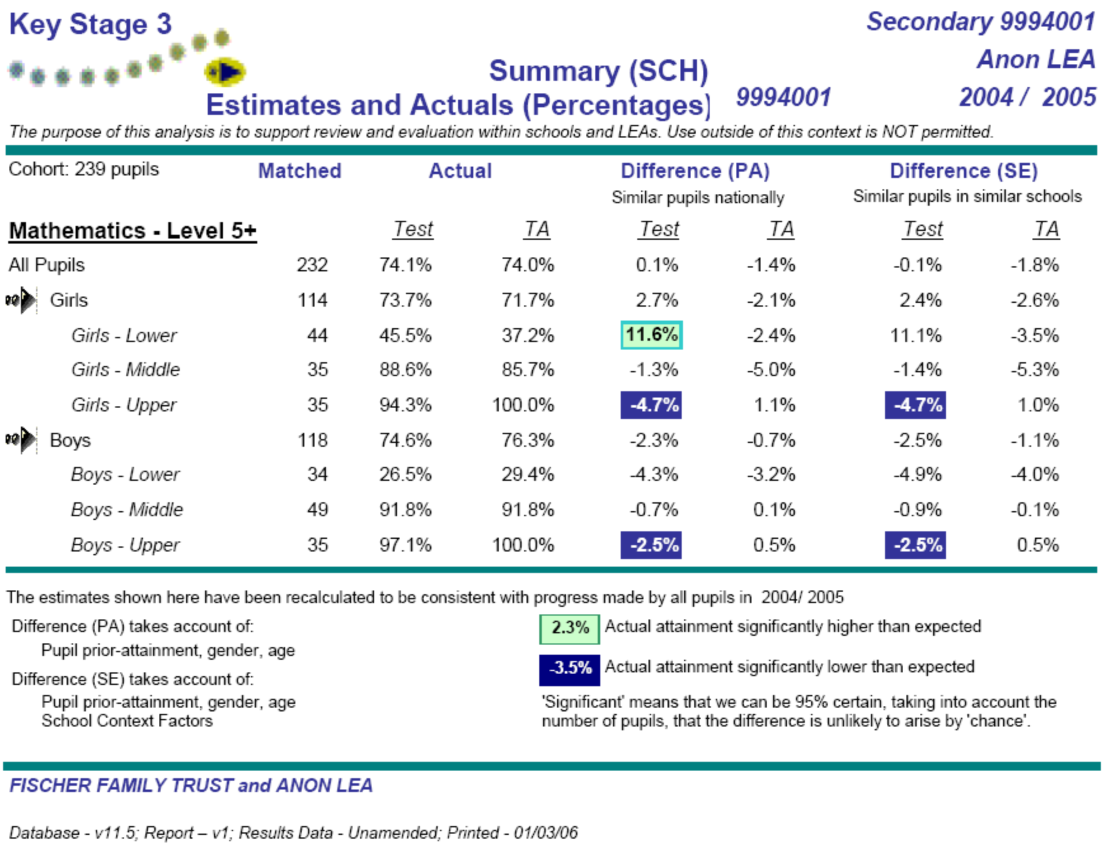
Peer effects

Both the mean and the distribution of the prior attainment of pupils at the end of the previous key stage are included in the SE model. Both terms are statistically significant. In common with the DCSF/Ofsted CVA model only the peer effects of the students in the same assessment cohort are included rather than all students in the school. It is not known whether any school wide peer effect measure was considered and found to be statistically non-significant.

An illustration of the use of FFT estimates model data for school self-evaluation

Estimate/Actual reports

The PA (prior attainment) and SE (socio-economic) models described above provide a baseline from which estimates of pupil attainment can be calculated and progress towards targets can be assessed. It is also possible to use residuals calculated by subtracting the actual attainment scores of pupils from such estimates to calculate value-added scores at the pupil, group, class and school level. FFT produce reports, known as *estimate/actual reports*, displaying such calculated residuals.



Key Stage 3

Secondary 9994001

Anon LEA

Summary (SCH)

Estimates and Actuals (Percentages) 9994001

2004 / 2005

The purpose of this analysis is to support review and evaluation within schools and LEAs. Use outside of this context is NOT permitted.

Cohort: 239 pupils	Matched	Actual		Difference (PA) Similar pupils nationally		Difference (SE) Similar pupils in similar schools	
Mathematics - Level 5+		*Test*	*TA*	*Test*	*TA*	*Test*	*TA*
All Pupils	232	74.1%	74.0%	0.1%	-1.4%	-0.1%	-1.8%
Girls	114	73.7%	71.7%	2.7%	-2.1%	2.4%	-2.6%
Girls - Lower	44	45.5%	37.2%	11.6%	-2.4%	11.1%	-3.5%
Girls - Middle	35	88.6%	85.7%	-1.3%	-5.0%	-1.4%	-5.3%
Girls - Upper	35	94.3%	100.0%	-4.7%	1.1%	-4.7%	1.0%
Boys	118	74.6%	76.3%	-2.3%	-0.7%	-2.5%	-1.1%
Boys - Lower	34	26.5%	29.4%	-4.3%	-3.2%	-4.9%	-4.0%
Boys - Middle	49	91.8%	91.8%	-0.7%	0.1%	-0.9%	-0.1%
Boys - Upper	35	97.1%	100.0%	-2.5%	0.5%	-2.5%	0.5%

The estimates shown here have been recalculated to be consistent with progress made by all pupils in 2004/ 2005

Difference (PA) takes account of:
Pupil prior-attainment, gender, age

Difference (SE) takes account of:
Pupil prior-attainment, gender, age
School Context Factors

2.3% Actual attainment significantly higher than expected

-3.5% Actual attainment significantly lower than expected

'Significant' means that we can be 95% certain, taking into account the number of pupils, that the difference is unlikely to arise by 'chance'.

FISCHER FAMILY TRUST and ANON LEA

Database - v11.5; Report – v1; Results Data - Unamended; Printed - 01/03/06

Figure 20: Extract of a sample FFT Estimate/Actuals report (Source: Adapted from FFT 2006a)

The report in Figure 20 above is for an anonymised secondary school and shows actual and residual scores for the percentage of pupils achieving the national benchmark of NC level 5 in mathematics from one cohort of pupils. The analysis has been subdivided by gender and by broad prior ability bands based on prior attainment in national curriculum tests (in this case, at the end of Key Stage 2, normally at age 11). Such an analysis would allow a school, and the mathematics department in particular, to evaluate the progress of this cohort of pupils against the PA and SE type estimates which are based on the performance of the previous year's pupils within the whole national dataset. The colour-coded significance boxes show where the attainment of a group of pupils is significantly (at the 95% level) above or below the national mean for that group of pupils, based on each model. This school might conclude that the attainment of boys at level 5 and above in mathematics is an area for consideration and review for development planning, but, more specifically, the attainment of both boys and girls with high prior attainment (who achieved high scores in KS2 NC tests) is an area of concern. There may be valuable lessons to be learned from reflecting on the significantly good progress in the attainment of girls with low prior attainment, especially in informing the teaching of boys with similarly low prior attainment, who have a negative residual for both the test and Teacher Assessment measures.

The slightly lower residuals resulting from subtraction of the SE model estimates from the actual percentages suggest that the school level factors added to the SE model (peer prior attainment mean and spread measures, the mean ACORN score and percentage free school meal entitlement) suggest that the school is above national means in at least some of these measures and so the value-added measure of pupil attainment is slightly lower against the SE model estimates than the PA model estimates.

The term “similar pupils in similar schools” under the heading “Difference SE” needs to be read with caution. It does not refer to the type of benchmarking analysis in which only pupils from schools identified to be similar the school under study on the basis of socio-economic measures are used as a comparator. Rather, it is still a comparison with all pupils in all schools nationally but with the school level socio-economic and peer-effect factors allowed for. See the response section below for a more detailed discussion of this issue and potential problems stemming from misunderstanding the comparisons being made.

The FFT valued-added models

It’s important to reiterate the key distinction between the estimates models described above and the valued added models that follow, especially as in terms of statistical procedure, they are very similar. The estimates models calculate pupil and school level estimates based on the performance of pupils from the previous year’s cohort, whereas the value-added models that follow produce pupil and school level value-added scores based on a regression analysis of the outcomes of the current cohort. Essentially the estimates models are predictive giving an indication of what might be achieved, whereas the value-added models are retrospective, giving an evaluation of what has been achieved

The PA and SE models described above are used to produce value-added scores as well as estimates in FFT reports. Thus schools are provided with effectiveness measures based on prior-attainment plus the pupil level factors of gender and month of birth via the PA model, and also, via the SE model, with effectiveness measures that include these same pupil factors together with the school context factors based on SES and peer-effects related to mean and spread of prior attainment.

A third, highly contextualised model is used to calculate value-added scores in FFT reports, the SX (or school extended) model. This model is closest to the DCSF/Ofsted CVA model which was developed after the SX model. The two sets of models are so close in fact that the FFT has devoted considerable time to researching and explaining the subtle, but in some cases significant differences between the two sets of scores that the models produce. What follows is a summary of the development of the SX model and a discussion of some of the key differences between the SX and CVA models, followed by an illustration of how the SX model is used to produce data for school self-evaluation.

The SX model and the FFT *Analyses to Support Self-Evaluation* Reports

Over the period 2003-2004 analysts at the FFT investigated a wider range of school and pupil level contextualising factors (FFT 2004a). This research stemmed from the opportunity to generate highly contextualised models using the wide range of variables made available via the yearly data gathering exercise carried out in state schools known as the Pupil Level Annual School Census (PLASC), introduced in 2002.

At their 2004 annual conference the FFT presented a report of the outputs of the research (FFT 2004a). Much of what follows is based on this and other FFT reports (2005b) of this research. The main focus of the discussion that follows will be on the contextualised value added model resulting from that initial research into and analysis of Key Stage 2 to GCSE progress but very similar models are now applied to measuring contextualised valued added progress of pupils and groups of pupils between the other key stages. Where later research resulted in modifications or developments that have been carried over to the most current model employed in FFT analyses this will be highlighted. For the KS2-4 SX model the outcome variable is the capped GCSE points score. This consists of the grades from the best eight GCSE examination results, or their equivalents, with each grade converted to a points score and then summed (see DfES (2006) for details on calculating GCSE and equivalent qualifications points scores).

The initial variables included in the FFT research of contextualising factors are listed in Table 7 below.

Pupil Level Variables	School Level Variables
Mean GCSE points score for pupil prior attainment band[8]	Mean of Intake Test Level for cohort
Mean teacher assessment (TA) level	Standard deviation (SD) of Intake Test Level for cohort
Individual Subject Test Levels (fine grade)	
Gender	% Girls, Girls School/Boys School
Month of Birth	
English as an additional language (EAL)	Percentage of pupils with EAL
Entitled to free school meals (FSM)	Percentage of pupils entitled to FSM
Special Educational Needs (SEN) Stage School Action, Action Plus or Statemented	Percentage SEN % School Action, % Action Plus, % Statemented
Ethnicity	
Time in School	
Joined 'late'	% Joined 'late' in cohort
FFT modified ACORN group	Mean FFT modified ACORN group
	Mainstream or Special / Unit

Table 7: Pupil and School level variables included in the FFT SX model (Source: modified from FFT 2004a)

From this presentation it is clear to see the approach was to research the effects of each contextualising variable at both the pupil and school level where possible. It is interesting to note that this approach appears not to have been extended to the inclusion of percentage of pupils in each ethnic group at the school level.

The addition of these extra contextualising factors produced an improvement in the model in terms of percentage of variance explained. The previous KS2-4 SE model (using percentage of pupils entitled to FSM as the sole SES factor) explained 55% of the variance in the cohort of data under study whereas the revised SE model with the modified ACORN types included as a school level variable alongside %FSM entitlement explained 60% of the variance. The extra pupil and school level factors included in the more contextualised SX model raised the percentage of the variance explained to 65% (FFT 2004a). This figure has since been revised by FFT (2005b Thomson & Knight 2006) to 60% compared with the percentage explained by the DCSF/Ofsted CVA model quoted as 57% (FFT 2005b).

[8] The mean GCSE capped points score is a prior attainment proxy variable used to overcome issues arising from non-linearity of the association between KS2 fine-grade test scores and GCSE capped points score. The rationale behind the use of this proxy variable is explained below.

In common with other value added models of pupil attainment of this type, prior attainment was found to be by far the most important factor in explaining the variance in the data. The association between mean test level as the prior attainment factor at the pupil level and level/grade attained at the end of the next key stage was found to be non-linear. FFT address the issue of non-linearity in two key ways.[9]

Pupils are allocated to one of 96 prior attainment bands prior to regression analysis being carried out and the mean GCSE capped points score calculated for each of these 96 bands (Thomson & Knight, 2006). The mean capped points score is then used as the prior attainment measure for each student. Schagen (2006) has suggested using a similar approach to deal with the non-linearity and ceiling effects that are a common issue in value-added analyses. Pupils are then allocated to a number of 'broad bands' (FFT, 2005b) based on prior attainment at the end of KS2. Separate regressions are performed for each band of pupils. Initially four bands were used (FFT 2004a). This was later increased to five bands (FFT 2005b). These two techniques combined are what FFT refers to as 'multilevel modelling'. It is important to note that this is not the same statistical procedure as that applied in the DCSF/Ofsted CVA model, which employs hierarchical regression analysis techniques (Kreft and De Leeuw 1998, Goldstein 2003) in which all pupils are considered to be nested within schools. During their early research the FFT considered methodological issues relating to the difference between ordinary least squares (OLS) regression techniques and the hierarchical modelling approaches that were being considered by the DCSF in developing their CVA model whilst the FFT decided to retain OLS techniques. One issue driving the decision to use OLS regression techniques was the application of a *shrinkage factor* already identified in the previous section of this review. Application of the shrinkage factor has the effect of reducing the level of uncertainty associated with outlying residuals and residuals for schools with small cohorts, but shrinkage may also mask true high or low value added scores for these schools.

As well as the interaction terms between ethnicity and %FSM entitlement, described in the estimates model section above, other interaction terms were added to the model (Thomson & Knight 2006), both at the pupil level (such as interactions between the differentials in KS2 test scores in different subjects) and at the school level (such as interactions between mean KS2 prior attainment score and school %FSM or ACORN percentile ranks). These interaction terms were retained where they improved the model fit and/or were considered to be significant. A

[9] The DCSF/Ofsted CVA model employs a simpler mathematical approach to dealing with the non-linearity of data by introducing a quadratic term, namely the square of the prior attainment score.

comprehensive list of factors included in the SX model with the values of their coefficients and effect sizes can be found in Table 8 below.

Variable	Coefficient in SX model	Effect size of variable	Variable Type (S / P)
KS4 mean score	1.04	1.43	P
SEN- School Action Plus	-67.15	0.63	P
Joined late	-60.18	0.57	P
School GDF (ACORN) rank	-0.94	0.43	S
Interaction: KS4 mean and KS2 TA differential	0.04	0.41	P
SEN: Statement	-36.13	0.34	P
SEN: School Action	-34.66	0.33	P
Gypsy/ Roma	-32.51	0.31	P
FSM	-29.64	0.28	P
Bangladeshi	28.60	0.27	P
Irish heritage Traveller	-25.39	0.24	P
Black African	24.97	0.23	P
Interaction: KS4 mean and school FSM rank	0.00	0.23	
Same intake and output school	24.74	0.23	P
Chinese	23.68	0.22	P
Interaction: KS4 mean and school KS2 mean	-0.03	0.22	
EAL	23.20	0.22	P
Interaction: KS4 mean and school GDF rank	0.00	0.20	
Ethnic background not obtained	-21.01	0.20	P
Pakistani	18.98	0.18	P
Indian	16.95	0.16	P
Gender	16.76	0.16	P
School FSM rank	0.35	0.15	S
Any other Asian	16.26	0.15	P
Any other ethnic group	14.11	0.13	P
KS2 English differential	23.34	0.13	P
Any other white	13.60	0.13	P
Ethnic background refused	-12.67	0.12	P
Months at school	0.54	0.11	P
Black Caribbean	11.07	0.10	P
Interaction: School KS2 mean and GDF rank	0.04	0.08	S
KS2 TA differential	14.23	0.08	P
Interaction: Ethnicity and school FSM rank	-0.16	0.07	
Any other black	-6.72	0.06	P
KS2 Maths differential	9.84	0.05	P
Mixed white/ Asian	5.54	0.05	P
KS2 science differential	9.25	0.05	P
Mixed white/ any other	-4.81	0.05	P
School KS2 mean	9.06	0.04	S
School KS2 SD	-24.10	0.04	S
Interaction: KS2 English and KS2 maths differentials	-15.74	0.03	P
Interaction: KS2 English and KS2 science differentials	-13.56	0.03	P
Mixed white/ black African	-3.32	0.03	P
Interaction: KS2 maths and KS2 science differentials	-15.25	0.03	P
Mixed white/ black Caribbean	-2.20	0.02	P
Interaction: School KS2 mean and FSM rank	0.01	0.02	S
Interaction: Months in school and joined late	0.99	0.02	P
Interaction: School FSM and GDF ranks	0.00	0.02	S
Age (months)	1.26	0.01	P
Irish	0.88	0.01	P

Table 8: Source: Factors included in the FFT SX model Thomson & Knight (2006)

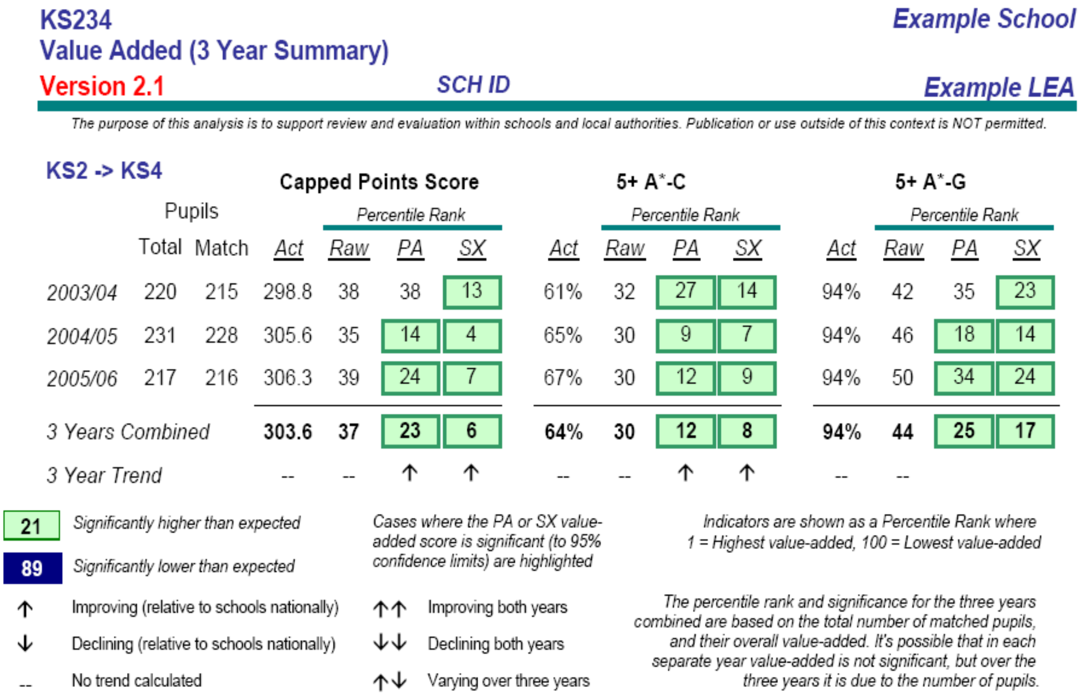
It is clear from the material presented at FFT conferences that there is a consultative process associated with FFT methodology and practice in which feedback from LAs and schools is sought, considered and addressed. This may well be a contributory factor to the value schools attribute to FFT data as shown in the NfER survey referred to above (Halsey *et a*l. 2005). 70% of respondents (FFT 2004a) expressed the desire to see more contextualising factors available from PLASC to be incorporated into the FFT models, although nearly half of these expressed some reservations on the use of such factors in generating estimates leading to issues related to 'labelling' and lowering of expectations with some respondents specifying that only those factors that raised the values of estimates should be incorporated in predictive FFT models. As a result of this feedback FFT proposed to retain the PA and SE models for the production of Type A-D estimates and use the new contextualised model (originally called KS but now known as SX – school extended) for retrospective measurements of value added progress. This avoided the potential criticisms that inclusion of factors such as SEN stage, FSM entitlement and certain ethnic groups would lower the value of estimates and thus lower expectations of pupil progress. Schools with a relatively high proportion of students in these categories will find that the standard Type A-D estimates produced from the PA and SE models don't necessarily match their school population well and so the FFT (2006c) recently published to LAs, plans to allow individual schools and LAs to use these contextualised models for generating pupil and school level estimates via a web based tool. A variety of options are planned that allow only pupil level contextualising factors to be included in estimate models, or both pupil and school level factors as in the full SX model. There will also be the facility to increase the estimates to place them in line with a user-specified centile of schools and to base the estimates on the previous three years of pupils value added progress rather than on just one year as in the current models.

Currently, the SX model is used in the main to produce an extensive report for schools entitled "Analyses to support self-evaluation". These reports were originally known as "Supplement to the PANDA" (FFT 2005d&e) as they were pitched alongside the Ofsted PANDA reports (Ofsted 2005a) as a tool for school self-evaluation and review. The reports provide a three year rolling value-added analyses of pupil attainment from KS1-2 for primary schools and KS2-3, KS2-4 and KS3-4 for secondary schools. Exemplar reports are available for download from the FFT public website (FFT 2007b&c). First schools also have analyses produced based on KS1 outcomes but this is not a value-added analysis as there is no use of a prior attainment measure as in all other cases. It is purely a contextualised comparison of the schools KS1 outcomes with those attained by students in other schools, based on the pupil and school level factors included in the SX model. FFT describe the statistical robustness of such context only analyses as "just good enough" (FFT 2005c) and places extensive notes to that effect in the introduction, a lengthy set of guidance notes, and the various headers and footers associated with each page of the report; a common, and we think highly commendable feature of all FFT reports.

The reports essentially provide three different analyses of the school's data:

- contextual value added summaries

a three year view of student attainment across a range of outcomes such as percentage of student achieving benchmarks in National Curriculum tests or GCSEs (including the new 5+A*-C benchmark including English and Maths) as well as outcomes involving the attainment of all students such as mean National Curriculum level or mean GCSE capped point score.



KS234 — Example School

Value Added (3 Year Summary)

Version 2.1 — SCH ID — Example LEA

The purpose of this analysis is to support review and evaluation within schools and local authorities. Publication or use outside of this context is NOT permitted.

KS2 -> KS4

	Pupils		Capped Points Score				5+ A*-C				5+ A*-G			
				Percentile Rank				Percentile Rank				Percentile Rank		
	Total	Match	Act	Raw	PA	SX	Act	Raw	PA	SX	Act	Raw	PA	SX
2003/04	220	215	298.8	38	38	13	61%	32	27	14	94%	42	35	23
2004/05	231	228	305.6	35	14	4	65%	30	9	7	94%	46	18	14
2005/06	217	216	306.3	39	24	7	67%	30	12	9	94%	50	34	24
3 Years Combined			**303.6**	**37**	**23**	**6**	**64%**	**30**	**12**	**8**	**94%**	**44**	**25**	**17**
3 Year Trend			--	--	↑	↑	--	--	↑	↑	--	--		

🟩 21 Significantly higher than expected

🟦 89 Significantly lower than expected

Cases where the PA or SX value-added score is significant (to 95% confidence limits) are highlighted

Indicators are shown as a Percentile Rank where 1 = Highest value-added, 100 = Lowest value-added

↑ Improving (relative to schools nationally)

↓ Declining (relative to schools nationally)

-- No trend calculated

↑↑ Improving both years

↓↓ Declining both years

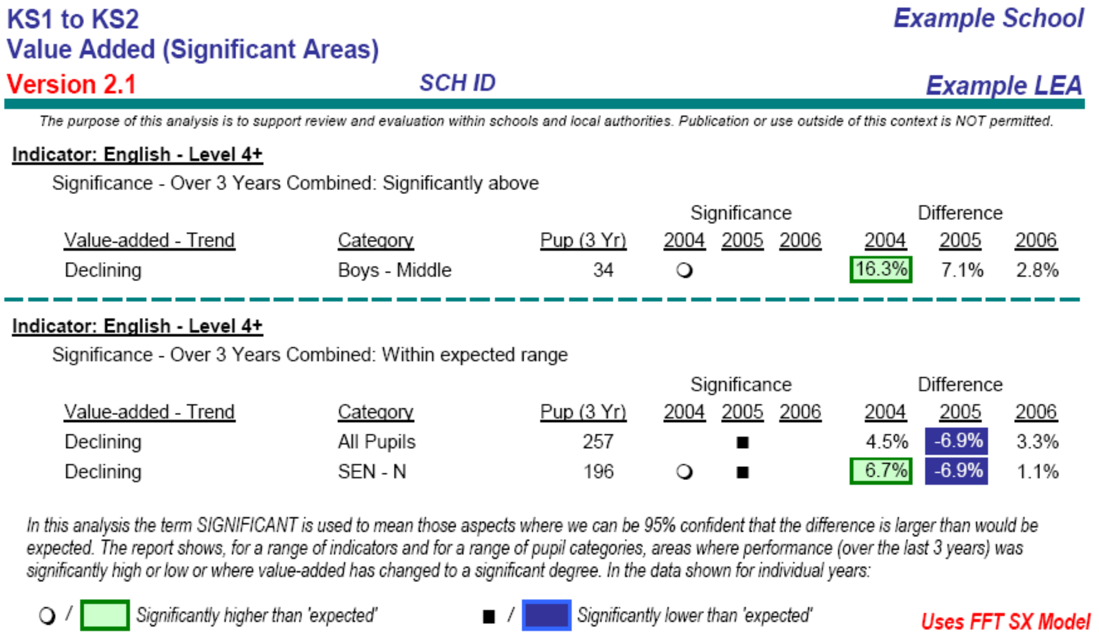
↑↓ Varying over three years

The percentile rank and significance for the three years combined are based on the total number of matched pupils, and their overall value-added. It's possible that in each separate year value-added is not significant, but over the three years it is due to the number of pupils.

Figure 21: Extract from the FFT Analyses to Support Self-Evaluation: three year value-added summary. Source: FFT (2007c: 4)

- significant areas grid

a breakdown of value-added measures of attainment by gender, prior ability 'band' (upper, middle and lower), SEN stage, entitlement to FSM and ethnicity. Only those groups where value-added progress was significantly high or low over the three year average or where changes to the state of significance occurred over the three years are included.



KS1 to KS2
Value Added (Significant Areas)
Version 2.1 SCH ID

Example School
Example LEA

The purpose of this analysis is to support review and evaluation within schools and local authorities. Publication or use outside of this context is NOT permitted.

Indicator: English - Level 4+

Significance - Over 3 Years Combined: Significantly above

Value-added - Trend	Category	Pup (3 Yr)	Significance 2004	Significance 2005	Significance 2006	Difference 2004	Difference 2005	Difference 2006
Declining	Boys - Middle	34	○			16.3%	7.1%	2.8%

Indicator: English - Level 4+

Significance - Over 3 Years Combined: Within expected range

Value-added - Trend	Category	Pup (3 Yr)	Significance 2004	Significance 2005	Significance 2006	Difference 2004	Difference 2005	Difference 2006
Declining	All Pupils	257		■		4.5%	-6.9%	3.3%
Declining	SEN - N	196	○	■		6.7%	-6.9%	1.1%

In this analysis the term SIGNIFICANT is used to mean those aspects where we can be 95% confident that the difference is larger than would be expected. The report shows, for a range of indicators and for a range of pupil categories, areas where performance (over the last 3 years) was significantly high or low or where value-added has changed to a significant degree. In the data shown for individual years:

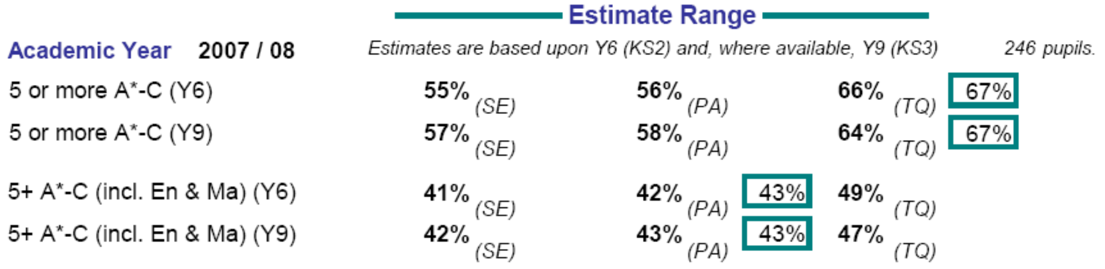
○ / 🟩 *Significantly higher than 'expected'* ■ / 🟦 *Significantly lower than 'expected'*

Uses FFT SX Model

Figure 22: Extract from the FFT Analyses to Support Self-Evaluation: significant areas grid. Source: FFT (2007b: 15)

- estimates

for a range of outcomes type A (PA model), B (SE model median) and D (SE model top quartile) estimates are given together with a fourth estimate of the potential attainment based on the value-added progress made over the three years covered by the report.



Estimate Range

Academic Year 2007 / 08	*Estimates are based upon Y6 (KS2) and, where available, Y9 (KS3)*				*246 pupils.*
5 or more A*-C (Y6)	55% *(SE)*	56% *(PA)*		66% *(TQ)*	67%
5 or more A*-C (Y9)	57% *(SE)*	58% *(PA)*		64% *(TQ)*	67%
5+ A*-C (incl. En & Ma) (Y6)	41% *(SE)*	42% *(PA)*	43%	49% *(TQ)*	
5+ A*-C (incl. En & Ma) (Y9)	42% *(SE)*	43% *(PA)*	43%	47% *(TQ)*	

Figure 23: Extract from the FFT Analyses to Support Self-Evaluation: estimates. Source: (FFT 2007c: 11)

The advantage of the three year rolling model is that it presents a measure of school progress over time providing both a richer and a more stable picture of school value-added. The merging of three years of student data also produces larger numbers of students in subgroups such as SEN stage or ethnic background so that shorter confidence intervals resulting from larger sample sizes produce more informative measures. For small rural schools this provides similar benefits when working with data at the whole pupil cohort level. It also avoids the need to resort to hierarchical modelling techniques that employ shrinkage in order to achieve these improvements in the robustness of estimates.

When the launch of the new electronic version of the Ofsted PANDA, RAISE online (Reporting and Analysis for Improvement through School Self-Evaluation, Ofsted 2005b) was delayed from its planned Summer 2006 launch date the FFT were commissioned by Ofsted and the then Department for Education and Skills (DfES now known as the DCSF) to produce their "Analyses to support self-evaluation" reports for every maintained school in England and Wales. RAISEonline eventually launched in a reduced format in January 2007. With no paper PANDAs issued by Ofsted during the Autumn term the FFT reports were, for some schools, the only detailed source of external data available to assist them in completing or updating their Self-Evaluation forms (SEF) and informing conversations with School Improvement Partners and Ofsted Inspectors, all of which are a crucial aspects of the Ofsted New Relationship with Schools (NRwS) inspection regime (DfES/Ofsted 2004).

Although no detailed PANDA analysis was made available, overall CVA scores for schools were published by the DfES, based on student attainment in the 2006 round of examinations. Thus schools were faced with two measures of overall school performance in the form of SX and CVA school level scores, but only one set of finer analyses giving valued-added progress measures for groups of students and by subject that schools had been utilising for self-evaluation and improvement. This served to highlight issues related to multiple measures of value added, particularly for those schools where analyses resulting from the FFT SX model and DCSF CVA calculations differed. Thomson and Knight's work for the FFT (2006) comparing FFT SX measures with the CVA scores concluded that 21% of schools had a different significance state for KS2-KS4 CVA compared to equivalent SX model (derived from a separate comparison of the 95% confidence interval for each individual measure with its national mean value), but only 4% (122 schools) had a significantly different result (comparing the two measures together to find that the 95% confidence intervals for the two measures did not overlap). The differences between these two states is a subtle but important one which is not easily grasped. As measures of significance in comparison with national means have been included in PANDA data in recent years they clearly have impact with school leaders and with inspectors. It would require a detailed working knowledge of the two statistical models to unpick why the significance state for the two value-added measures might be different. Not an easy job for a school leader at any time but especially when under the microscope of inspection. This has clearly brought FFT data more sharply into the arena of accountability measures, a departure from the FFTs stated primary purpose of supporting school self-evaluation. We have heard anecdotal reports of school leaders having to argue a case for the picture presented by FFT data in discussions with Ofsted inspectors. Whether this will have set a precedent for greater use of FFT data in Ofsted inspections, either by school leaders or by inspectors remains to be seen but it certainly has created a new set of tensions.

3.1.3 Issues related to the use of value-added measures in longitudinal analyses of school improvement trajectories

Almost all the school effectiveness studies discussed in this review were cross-sectional in nature, that is, they look at the relative performance of students/schools in a particular cohort or year. Longitudinal studies are rare as the data needed to produce contextualised studies, via UPNs and PLASC, is relatively recent development. Also, the changing nature of measures makes longitudinal comparison studies more difficult.

In a ten year study (1993-2002) of data from secondary schools from one large English Local Authority (Thomas *et al.* 2007) has resulted in what may be a seminal paper on the stability of value added school performance measures and sustainability of that aspect of school performance that such techniques are able to measure.

Two of the paper's co-authors, John Gray and Sally Thomas, have been particularly active in this area over the last decade or more (Gray et al. 1996; Gray et al. 2001; Gray et al. 1995; Thomas et al. 1997). Some of these studies refer to value-added measures of school progress but over shorter time spans such as three years. Another study (Mangan *et al.* 2005) can match the time span but made use of raw threshold measures (percentage 5A*-C at GCSE). This study employed a time-series analysis of the benchmark GCSE attainment for 541 schools from 20 English LAs. Their finding was that, although there was an aggregate trend of improvement over time in the benchmark outcome,

> *…'continuous improvement' is an aggregate phenomenon that does not survive disaggregation. At the school level there is very considerable variation in improvement paths.* (*ibid*: 47)

They found that few schools demonstrated a continuous improvement trend for more than three consecutive years and that there was evidence of general upward trends but with a large amount of year-to-year variation. One overall conclusion was that "'School improvement' seems to come in bursts, whether of energy or changing contexts." (*ibid*: 48) The main criticism of these findings, which the authors themselves acknowledge, is that no account was made to adjust for differing intake in terms of prior attainment and context in the way that value added models do. This is where the key advantage of Thomas *et al.*'s recent study comes to the fore.

Data from the 63 schools with a complete 10 year record of cohorts were reported although the full data set analysed in the model contained 142,574 students from 134 schools. Descriptive statistics showed that the 63 schools were considered to be representative of the whole set of schools. The authors found that there was a continuous trend of improvement in the percentage of students attaining 5 or more GCSEs at A*-C (the 'headline' figure reported by the Government) for the LAs secondary schools and that this trend was in line with the improvement trend nationally. In analysing the school data to investigate trends over time a multilevel model was used that assumed a linear trend over the time period and this was compared with a corresponding MLM that did not impose a linear trend by assigning dummy variables to each year (non-linearity was explored using quadratic and cubic terms but these were not found to be statistically significant). Thus the average trend over the whole 10 yr period could be compared with the model giving year-on-year fluctuations. School level residuals were calculated for both models in order to assign schools to one of three categories; performing in line, significantly below, or significantly above the average trend. The linear models calculated two residuals for each school; an intercept residual (a measure of baseline performance at the start of the study in 1993) and a slope residual giving a measure of the improvement trend over time. In the unconstrained models a single residual is calculated for each school for every year of the study. Finally, two sets of these models were produced, one representing a contextualised value-added model and one a "raw" model with no adjustment made for prior attainment or contextualising factors. The prior attainment measure used was score in Cognitive Ability Tests (CATs) at age 11 rather than the KS2 SAT average points score used in KS2-4 CVA scores. CATs were chosen as they were not considered to be 'high stakes' tests that the students were specifically prepared for, in contrast to the KS2 National Curriculum tests. The contextualising factors included in the model were month of birth, gender, FSM entitlement, ethnicity, SEN and mobility (full 5 years in the same secondary school).

The value-added linear trend model identified 16 of the 63 schools (25.4%) as having a slope residual (the measure of the improvement trend) significantly above the mean trend and 12 schools with a slope residual significantly below the mean. This was much larger than the number of schools identified in the same category at the midpoint of the study (Thomas 2001) when less than 5% of schools were in the corresponding improving schools category. Also of interest, of the 16 most improving schools 13 of these had an intercept residual significantly below the mean, indicating that they were in the group of poorer performing schools at the start of the study in 1993. Correspondingly 8 of the 17 schools with higher initial value added scores demonstrated a lower than average improvement trend. This seemingly inverse relationship between initial performance and improvement trend was confirmed by the negative value of the Pearson correlation between slope and intercept residuals (r = - 0.58)

In the unadjusted, “raw” model 14 schools were identified as showing an improvement trend significantly higher than the mean. 9 of these schools were in the same category based on the value-added model meaning that 7. The correlation between intercept and slope residuals was relatively weak (r = - 0.15).

In the non-linear value-added model just 4 of the 63 schools were found to have sustained 5 or more years continuous improvement (significantly above the average change for the corresponding years). A further 7 secured four consecutive years of improvement and then 15 schools with three consecutive years. The modal span of continuous improvement was just two years which was the case for 28 schools (44% of the sample). Whilst demonstrating continuous years of improvement in raw scores was harder, as might be expected, the figures for the number of schools demonstrating different spans of continuous improvement were surprisingly close with 52% being able to secure just two consecutive years as the longest span of improvement.

Correlation analyses of the school residuals for each year showed that the raw model residuals were more stable over time (r > 0.92 for 5 year span) than the VA residuals (0.6 ≤ r < 0.9 for 5 year time span). This suggests that schools have the capacity to transform their relative positions based on VA scores in a way which is less possible when measured using “raw” attainment measures.

Perhaps more compelling than the above is an analysis of the absolute allocation of schools to the categories significantly high performing, performing in line with the mean and significantly low performing at the beginning, middle and end of the study (1993, 1998 and 2002). Although the residuals measured at the three time points represent only snap-shots of school performance, the analysis demonstrates the capacity for a school to demonstrate such a level of improvement that it can change its ‘status’ in terms of performance measure. Of course, negative changes are also highlighted. Just under half (46%) of the schools were found to have the same designation across all three time points. Of the 16 significantly higher performing schools in 1993, 10 were still in this category in 2002 (one of which had 'dropped' to performing in line with the mean at the midpoint of the study). The other 6 schools were all in the in line with the mean performance category although one of these had dropped down to the significantly lower performing category at the midpoint.

Of the 21 significantly lower performing schools in 1993, eight had shifted to being in line with the mean performance by 1998. By 2002 2 of these 8 schools had dropped back to the lower performing schools category, 4 had maintained their performance gain and 2 schools had managed to consolidate their gain in performance and transform this into becoming higher performing schools.

Figure 24 charts the 'school improvement trajectories' of the 63 schools in the Lancashire longitudinal study. The lines are designed to show the movement of schools from one category to another across each 5 year span. The lines are dotted to express the fact that the actual year-on-year path is unlikely to be linear. The weight of the line is representative of the number of schools following each particular trajectory. The actual numbers of schools are given against each line. The figure shows that the modal trajectory from each of the three starting categories is 'no change'. Nevertheless, the positive news is that it is possible for schools in the lower performing category to transform their designation over time, and for most of those schools to consolidate and sustain these VA performance gains. Improvements for those schools starting from the mean performance category seemed somewhat more difficult to sustain. It must be noted however that relative comparisons of schools such as those generated in Thomas *et al.*'s study will result in 'losers' as well as 'winners'. Schools that show strong and sustained upward improvement trajectories will 'force' other schools to show negative trajectories in respect to the mean. As much as it is laudable for all schools to aim to be significantly *higher* performing schools at a future point in time, when the school residuals are reordered at the next time point, by definition only a limited number of schools will achieve the designation.

Thomas *et al.* found that 'raw' attainment measures produce a starkly different set of improvement trajectories (represented in Figure 25). None of the schools that began the study with significantly lower 'raw' attainment scores moved to the significantly higher category over the 10 year period. In fact only 3 schools from the 25 in the lower performing category managed to move into line with the mean attainment category. A further 2 schools were able to sustain their improvement from the mean attainment category at the mid-point of the study and retain their change of status to the group of higher attaining schools. At the other end of the picture 92% of the schools that began in the higher attaining schools category held their position over the 10 years (*ibid*: 295).

Some of the key findings of the study can be summarised as follows:

- Although some schools demonstrating significant improvement trends based on raw attainment measures also showed significant improvement trends based on value added scores, value-added does make a difference to the make-up of the group of schools in this designation.
- Schools with a low baseline VA score seem more likely to demonstrate high levels of improvement over time. It is harder to demonstrate such levels of improvement based on raw scores.
- A substantial period of time (more than five years) may be required for schools to establish above average improvement trends.
- Year-on-year improvements, even in value added scores, are difficult to sustain beyond 2-3 year bursts, which is in line with Mangan *et al.*'s findings using much cruder raw attainment threshold measures.
- Changing school status (in terms of broad effectiveness categories based on tests of statistical significance) is challenging but not impossible based on value added measures, whilst raw attainment measures seem to perpetuate past designations.

Summaries of school trajectories that I have produced in Figures 24 and 25 it is possible to conclude that there are a number of categories within which schools can fall, based on their performance trajectories over a specified period, as schools stay within, or move across selected performance thresholds. Thomas et al (2007) selected the threshold of statistical significance for their discussion of the longitudinal patterns of school performance in the study and this has been applied in the figures below. However, other threshold might be chosen such as the 25th and 75th percentile marking the boundaries between the upper quartile performance and the middle 50% and between the middle 50% and performance in the lower quartile. 5 categorisations of the trend over time, in attainment or (value-added) progress appear as marked in the diagram.

For the highest performing schools at the start of the sequence remaining in the group of school with significantly high outcomes (or the upper quartile based on the chosen metric) over the period would give the highest categorisation (5). A slow decline in outcomes, eventually moving down into non-significant outcomes toward the end of the period, or moving between significantly high and non-significant outcomes over time, would rate a 4. A steeper decline in outcomes, moving into the non-significant group relatively early in the time period would give rise to a category 3 trajectory. A steeper decline still, moving just into the group of schools with significantly low outcomes by the end of the period would be a category 2 trajectory and a very steep decline down to non-significant outcomes would give the lowest designation of category 1 trajectory.

A similar approach could be taken for schools starting in the middle band and those starting in the lowest band, as shown in the figures below.

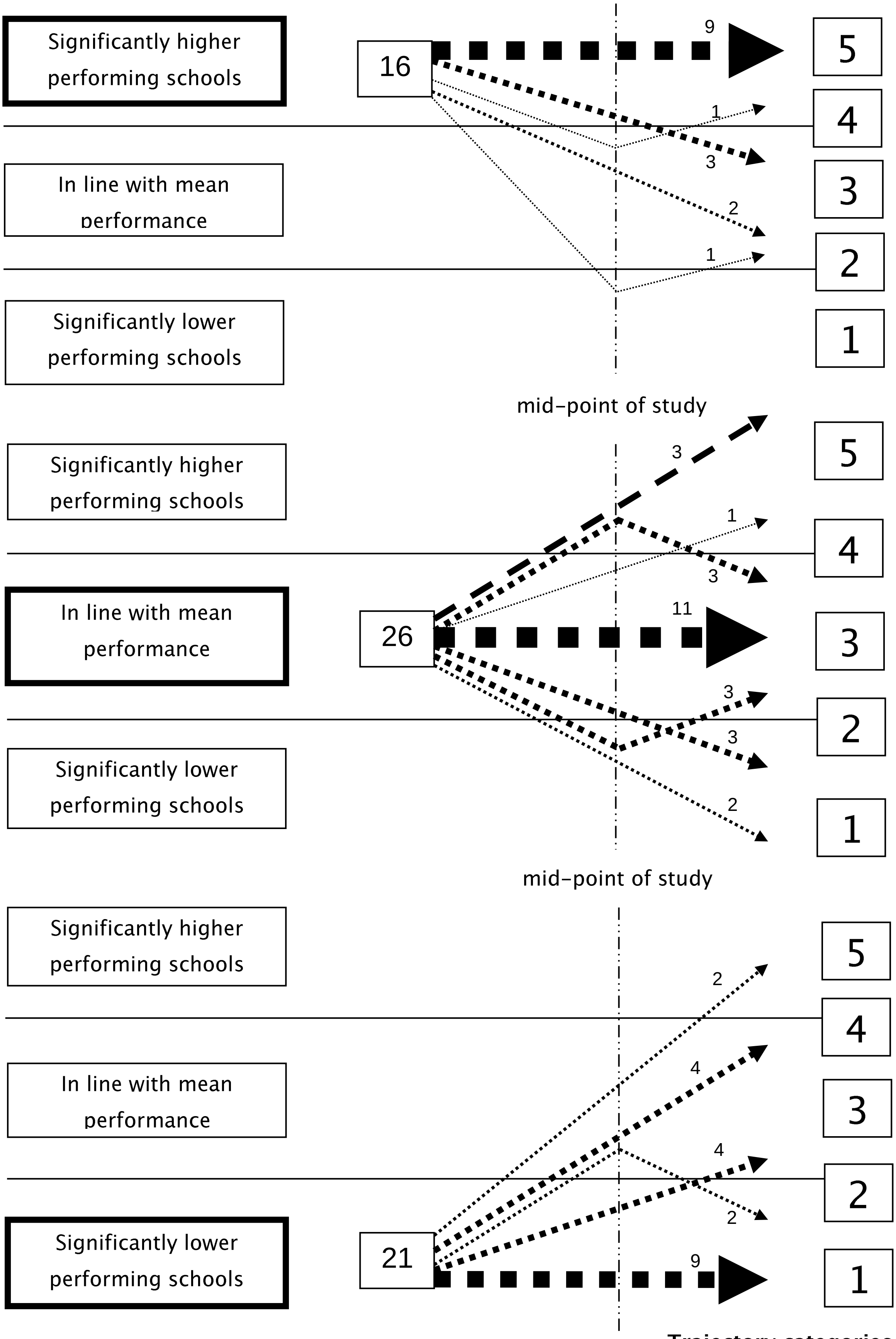

Significantly higher performing schools
16
9
5
1
4
3
In line with mean performance
3
2
2
1
Significantly lower performing schools
1
mid-point of study
3
5
Significantly higher performing schools
1
4
3
In line with mean performance
26
11
3
3
2
3
Significantly lower performing schools
2
1
mid-point of study
Significantly higher performing schools
5
2
4
In line with mean performance
4
3
4
2
2
Significantly lower performing schools
21
9
1
Trajectory categories

Figure 24: Value added 'school improvement trajectories' for the 63 secondary schools in the Lancashire longitudinal study (based on data in Thomas et al, 2007: 278).

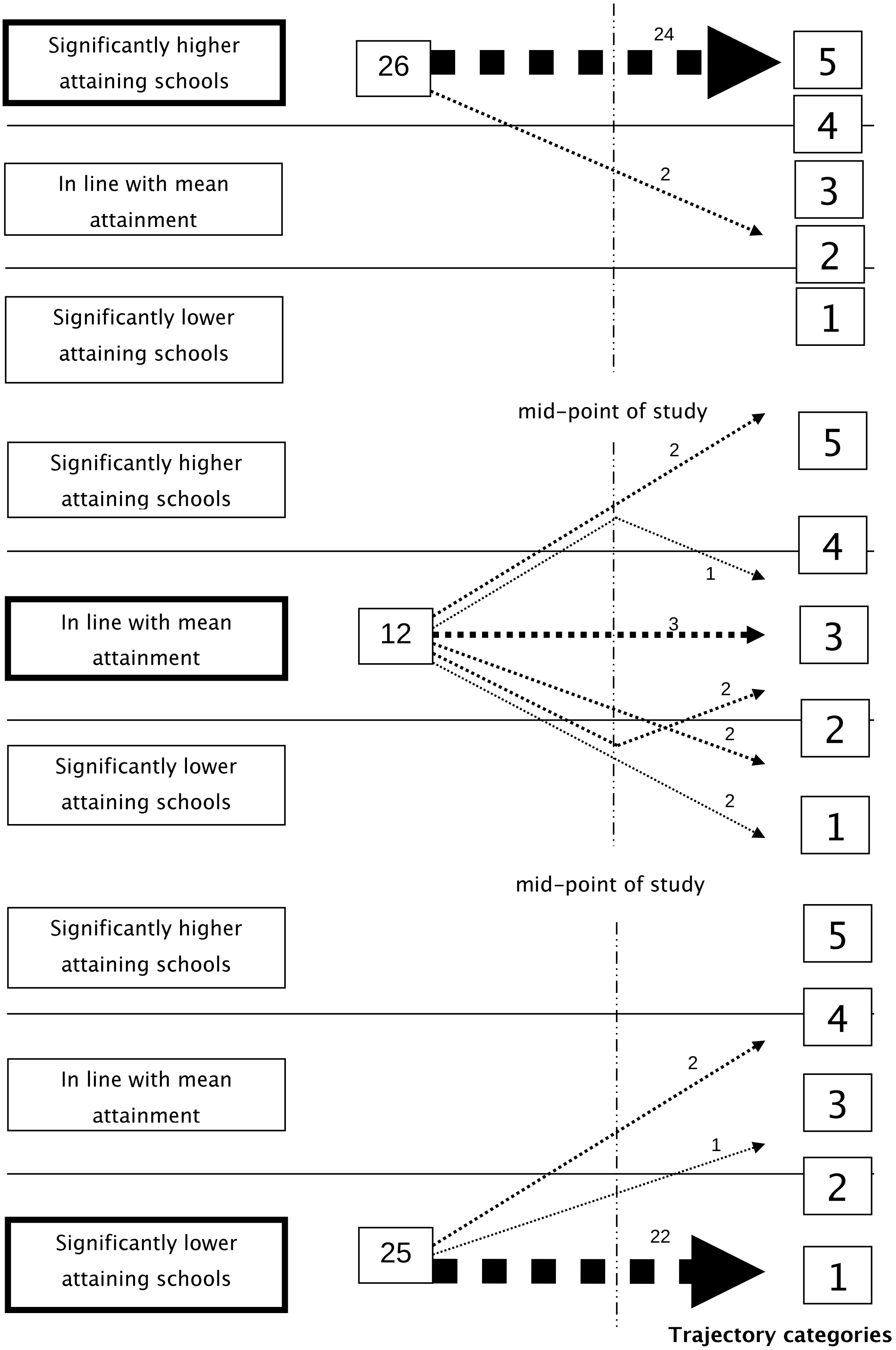

Significantly higher attaining schools
26
24
5
4
In line with mean attainment
2
3
2
Significantly lower attaining schools
1
mid-point of study
Significantly higher attaining schools
2
5
4
1
In line with mean attainment
12
3
3
2
2
2
Significantly lower attaining schools
2
1
mid-point of study
Significantly higher attaining schools
5
4
In line with mean attainment
2
3
1
2
Significantly lower attaining schools
25
22
1
Trajectory categories

Figure 25: Raw attainment 'school improvement trajectories' for the 63 secondary schools in the Lancashire longitudinal study (based on data in Thomas et al, 2007: 295).

3.2 A wider perspective of the outcomes of schooling – non-cognitive measures and school effectiveness

Introduction

As indicated in the quote from Mortimore (1998) in the introduction, and considered in detail in the previous section of this chapter, since the inception of the field the dominant school effectiveness model has been focused on the cognitive outcomes of schooling, in the form of results of tests and assessments of academic achievement. Early school effectiveness researchers were aware of this emphasis (Sammons, 2007; Teddlie and Reynolds, 2000) and, while the dominant focus still remains with cognitive outcomes, a growing body of work has considered the students and school level factors associated with non-cognitive outcomes.

This is an important line of research, not just with initiatives such as SEAL in view. Many school mission statements and national statements of the aims and purpose of education, making state that the aims of schooling are wider than this simplified view.

> *Our aim at Salway Ash School is to:-*
>
> *provide a supportive environment where children, staff, parents, governors and the local community all work together to attain high standards, achieving academic, creative, spiritual, social and emotional fulfilment.*
>
> *(Salway Ash, 2012)*

Even a nation such as Singapore, with a world wider reputation for outstanding academic success from its school system as identified by the McKinsey Report (Mourshed *et al,* 2010; 15 & 19) has a wider perspective of the desired outcomes of education.

- *a confident person*
- *a self-directed learner*
- *an active contributor*
- *a concerned citizen*

> *These outcomes establish a common purpose for educators, drive our policies and programmes, and allow us to determine how well our education system is doing*.
>
> (MoE Singapore, 2009)

While these stated aims were articulated in this form as recently as 2009 there a similar version first published back in 1997.

Of course, once could argue that initiatives such as SEAL, with its stated aim of improving the learning environment, will have an effect on the cognitive outcomes of schooling and so may be part of the school effect that has evaluated using the types of statistical modelling applied in the classic school effectiveness framework and discussed in the earlier section of this review.

The issue with the focus on cognitive outcomes of schooling, affected indirectly by such interventions, is the wider variety of factors that influence such outcomes (again, as indicated in the earlier section of this review). This generates a fair amount of noise in terms of associating an intervention like SEAL directly with cognitive outcomes and increases the chance of measurement error. While the use of cognitive measures as indicators of the success of non-cognitive interventions is practically possible, it is at best but statistically problematic, while on the other hand, the use of non-cognitive measures may serve to reduce the measurement noise but raise questions about the practical significance of the measured outcome data. Nevertheless, as indicated by the SEAL School Case Study in the in Introduction, cognitive measures have strong face validity as the preeminent measures of the outcomes of any programme or intervention in schooling. The high-stakes nature of cognitive outcomes, also discussed in the previous section, keeps the focus on these metrics. In line with the use of school effectiveness data as performance metrics within accountability domain, a key concern is that the development of non-cognitive outcome measures, either regionally or nationally, has potential for unintended consequence by setting up another set of outcomes which students and schools will be required to focus on, to set targets for, and ultimately perform well in.

Non-cognitive measures present some important measurement validity issues such as the problem of broad or conflicting definitions of such outcomes. For example, resilience can be considered to be relationship based, as in overcoming disagreements with in the social aspects of learning, or work based, a capacity to stick with a task in the face of adversity. A key issue therefore is the need to arrive at a commonly accepted construct or operationalisation of the term. This may not be so problematic for the *development* of such instruments, in line with the inductive research tradition, as analyses can help to drive the development of theory to determine a common construct for the non-cognitive outcome. The fact remains, however, that some of the most widely used non-cognitive measures (e.g. Costa and McCrae's (1988) "Big Five") are accused of being atheoretical (Borghans et al, 2008: 984). There are problems when it comes to applying such instruments as measurement tools for educational outcomes in order to determine aspects of school effectiveness, as the assumption persists that the instrument is measuring a common latent construct. This is compounded by the data reduction techniques such as factor analysis used in the psychometric development as the names applied to latent variables that arise from factor analysis can give the impression of measurement precision.

One core assumption made in the school effectiveness field is that the outcomes of schooling can be reliably measured. When it is the academic outcomes of schooling that are under scrutiny, through the results of tests and examinations of students' cognitive learning, we have already seen that this approach is not without its critics, who, while they may not throw out the concept of measurement of academic outcomes per se, certainly contest the validity of determining school effectiveness via such measures. One might suggest that making judgments about school effectiveness on the basis of non-cognitive outcomes could be subject to even stronger criticism but research in this area has not led to wide-scale, high-stakes national testing with schools being ranked by their students' performance. Thus the interest is predominantly in the presence and the size of the school effect for non-cognitive outcomes without, as yet, the accompanying baggage that comes with the same analyses being used in the accountability domain. Whether the politics of the outcomes of schooling will allow things to remain this way may well depend on the perceived importance of non-cognitive outcomes. The change of Government from New Labour to the Conservative/Liberal Democrat coalition suggests that the incorporation of non-cognitive outcomes into school performance measures is not coming any time soon. This can be seen by the marked contrast in recent consultations by the two administrations into school performance measures toward the end of the Labour administration the proposed School Report Cards, which were due to introduced from 2011 (DCSF, 2009; 3), were to include measures of "attainment, pupil progress, wider outcomes, narrowing gaps, parents' views, pupils' views" (ibid; 11). These wider outcomes were to include measures of wellbeing and students' perceptions of school (CSFC 2010; 82). A closer look at the proposals show that quantitative measures of wellbeing that were in view were attendance, persistent absence, permanent exclusions, post-16 progression, pupils provided with at least two hours per week of high quality PE and sport and the uptake of school lunches (DCSF, 2009; 31). The consultation concedes that these are proxy measures for the school's contribution to student wellbeing. An Ofsted (2009) consultation on schools' contributions to wellbeing suggests that the pupil and parental perception measures might include the extent to which the school helps students to manage their feelings and be resilient, to feel safe, and to enjoy school.

By contrast the current coalition government's recent consultation on school performance measures (DfE, 2013) assumed the debate was only about the blend of metrics of the cognitive outcomes of schooling, with no consideration of any measures of non-cognitive outcomes other than the perspective of Ofsted. The accountability indicator consultation was more to do with the balance of the contribution made by the various cognitive metrics, namely raw attainment versus progress measures, and threshold indicators versus measures in which every student makes a contribution.

Early school effectiveness research studies of non-cognitive outcomes

Despite the fact school effectiveness research (SER) studies have focused on the cognitive outcomes of schooling, some of the earliest studies also collected data for non-cognitive outcomes with the intention of determining the magnitude of the association between cognitive and non-cognitive outcomes. This includes the seminal *Fifteen Thousand Hours* study (Rutter et al., 1979) which made tentative efforts to incorporate non-cognitive outcomes, in the form of behavioural outcomes such as the number of delinquent students as a percentage of students in the school, in their analysis.

The conclusions drawn from these early effectiveness studies were inconsistent in terms of the measured association between cognitive outcomes and non-cognitive outcomes of schooling. Brookover et al. (1979) found an inverse relationship between self-concept and cognitive outcomes. Thus the findings of early studies provide a dearth of evidence for a relationship between the cognitive and non-cognitive domains in terms of the school effect. An extensive meta-analysis conducted by Hansford and Hattie (1982) drew together findings from 128 studies involving over 200,000 participants. They were able to utilise (or to calculate) correlation coefficients for each study to determine the relationship between 'self-measures' and cognitive outcomes. The correlations ranged from -0.77 to 0.96 but they found a mean correlation coefficient of 0.21. Another key finding was that certain background factors, familiar as having explanatory power in models of cognitive outcomes, were found to be similarly significant in studies of non-cognitive outcomes, namely socioeconomic status and ethnicity (see also Brookover et al, 1979). Age (grade level) was also found to be a key explanatory variable and other key modifiers focused on the specific outcomes being measures (in both domains) and the psychometric properties of the assessment instruments.

Utilising a school effectiveness framework in their study Mortimore et al. (1988) found no significant association between cognitive outcomes and a range of non-cognitive outcomes including behavioural and attitudinal factors. In other studies, where a significant association is determined there is no indication of the direction of causality. While intuitively we might posit that for non-cognitive factors such as achievement motivation the direction of causality is from the non-cognitive outcome to the cognitive, one might also suggest, with Helmke (1989), that the direction of causality is in the opposite direction, with good academic outcomes leading to positive outcomes in the affective domain.

Knuver and Brandsma's (1993) analysis of a range of non-cognitive outcomes (attitudes to language and mathematics, achievement motivation, academic self-concept and school wellbeing) and cognitive outcomes in language and maths, were correlated against a number of explanatory factors such as socioeconomic status, gender, ethnicity and IQ. The two cognitive outcomes were at least moderately correlated (>0.3) with IQ and SES, whereas only academic self-concept and attitude to mathematics were correlated with IQ, and then only in the older age group under study. When correlating cognitive and non-cognitive outcomes attitudes to language was more strongly correlated to outcomes in language, and similarly for the attitudes to mathematics and outcomes in maths. Academic self-concept was consistently correlated (>0.3) with both outcomes in language and mathematics.

At the school level Knuver and Brandsma (1993) found mixed results for the association between non-cognitive and cognitive outcomes. The school ranks for achievement in language were significantly correlated with the school ranks for 4 of the 5 non-cognitive outcomes, although it is important to qualify this by indicating that the magnitude of the correlations were relatively weak (ranging from 0.16 for the strength of the association with attitude to language, to 0.25 for the association with achievement motivation). For achievement in mathematics the result was different. Only achievement motivation, in terms of school ranks, was significantly correlated with achievement in mathematics and this was the weakest of the significant correlations with a magnitude of 0.14 only. There were exceptional cases in the form of schools that were highly effective in both domains, as there were highly ineffective schools in both domains. Nevertheless, as Mortimore et al (1988) had earlier concluded, evidence of school effectiveness in the cognitive domain is not necessarily concomitant with effectiveness in the non-cognitive domain.

Measuring non-cognitive outcomes of schooling

In reporting one of the early studies of non-cognitive outcomes in a school effectiveness framework, Knuver and Brandsma (1993) discuss the negative nature of the term of 'non-cognitive', which, they assert, leaves plenty of room to cover a wide range of outcomes of education:

> - behavioural aspects (e.g. delinquency, drop-out, vandalism, truancy),
> - social aspects (e.g. learning to cooperate, meaningful interaction with fellow pupils and teachers, social well-being),
> - moral aspects and ethics (e.g. cultural values and norms, intercultural education),
> - aesthetics (e.g. appreciation for the arts),
> - physical education
> - and also affective outcomes like attitudes toward the subjects, learning in general and the school.
>
> (*ibid.*; 190)

SEAL, with its 5 aspects of self-image, managing behaviour, motivation, empathy and social skills overlaps with a number of these dimensions. Knuver and Brandsma consider that some of these outcomes may be regarded more as "pleasant additional effects of education" (*op cit*) rather than as outcomes in their own right and point to affective outcomes such as academic self-concept and wellbeing as examples of this. It is interesting therefore that Gadeyne et al (2006) have indicated that many studies of non-cognitive outcomes have focused predominantly on this type of non-cognitive outcomes in terms of affective outcomes closely allied to academic achievement motivation and self-concept. Nevertheless, Knuver and Brandsma concede that even such 'pleasant effects' may have a positive impact on cognitive outcomes of schooling. Knuver and Brandsma raise the question of to what extent such outcomes should fall within the remit of schools rather than others such as parents.

Huebner at al (1999) point out that many models of psychological wellbeing include multiple wellbeing variables such as self-esteem and other examples of positive affect. When subjective self-reports are made such models require the measurement approaches which can demonstrate construct and discriminant validity to show that the resulting set of multiple affective dimensions contain distinct constructs that the research participants are able to differentiate. As Huebner et al point out "the ability of children to distinguish between global life satisfaction and related constructs, such as self-esteem, needs to be investigated empirically rather than assumed" (1999; 2). To this end they utilised confirmatory factor analysis (CFA) methods to establish the validity of a two factor scale of wellbeing using two previous multi-item scales. These scales focused on 'global' constructs of well-being (global self-esteem and global life satisfaction) rather than those specific to the context of well-being at school. The application of CFA is well established in investigating the validity of instruments consisting of multiple constructs each measured using multi-item scales which are related to an underpinning latent construct (in this case *self-esteem* and *life satisfaction*). CFA provides a more deductive approach to determining validity than exploratory factor analysis (EFA) though some have advocated (Mulaik & Millsap, 2000) that the two can be utilised together in a stepped process, starting with EFA and progressing to CFA, to validation of the measurement model underlying a multidimensional, multi-item scale instrument. In CFA absolute model fit is determined using the chi square statistic. The direction of the hypothesis for model fit is such that a non-significant chi square value indicates a perfectly-fitting model. Though the value of chi square statistic for their model was significant and therefore not a perfect fit, Huebner et al (1999) found that their two factor model was a good enough fit to the data based on a range of indices of absolute and relative model fit. The use of such indices to determine model fit is based on studies examining the effect of sample size on the utility of the chi square statistic as an indicator of fit, such as those conducted by Hu & Bentler (1995; 1999) and Crowley & Fan (1997). The use of fit indices is a point of controversy in the CFA/SEM community, with some purists (Hayduk et al, 2007) arguing that model fit can only be evaluated using the chi square statistic, but this is not the majority view (McDonald & Ho, 2002; Miles & Shevlin, 2007).

Kyriakides et al (2011) used an item response theory approach known as Rasch modelling to produce a revised version of a pre-existing instrument, the Olweus Bully/Victim Questionnaire (OBVQ, Olweus, 1996). The instrument consisted of 40 items to measure a range of aspects of bullying from the perspective of both the bully and the victim. The instrument had been tested for test-retest reliability and internal consistency through Cronbach Alpha higher than 0.80 for the various scales within the instrument. Only very limited analysis of the validity of the instrument had been undertaken and so Kyriakides et al (2011) set out to conduct a thorough test of construct validity using Rasch modelling, a form of logistic regression that produces the odds of transition from one item response to the other (as opposed to the usual graded response model of the standard logistic regression procedure). The analysis provided evidence for the validity of the design of the instrument into scales for students being bullied and students bullying others. Through determining the level of difficulty of each item the Rasch modelling also enabled the ordinal scaling of the original instrument to be modified to give interval level data, which facilitated analysis of the data from the instrument within a multilevel modelling environment prevalent in school effectiveness research frameworks and key to determining the impact of school level interventions on the prevalence of bullying. A one way ANOVA of the data from both the *being bullied* and *bullying others* scales showed that the between school variance in the Rasch student estimates from each scale was significant. This was a necessary precursor to using aggregates of the student level data to give school level measures of bullying that could be utilised in multilevel models to determine the effectiveness of anti-bullying interventions. In one of the contexts of this multinational study (Cyprus and Greece – Kyriakides et al, 2013) the researchers found that a two level model of student and school fit the data better than a three tier model with the intermediate level of the class, suggesting that for reported bullying activity as an outcome the grouping of students into specific classes environment had no significant impact. A null, two level model revealed the between school variance for both the *being bullied* and *bullying others* scales was 10.6% and 9.1% respectively. A range of student level explanatory factors, and corresponding school level aggregates, were added including SES, ethnicity and gender, none of which added any significant explanatory power, but, unsurprisingly, a prior measure on the OBVQ was a significant predictor of later bullying activity, both at the student and the school level. The addition of these factors reduced the proportion of the explained variance at the school level for the two bullying activity scales by only a small amount down to 10.1% and 7.8%%

This in itself is an interesting study as it applies a theoretical approach derived from school effectiveness research that indicates the extent to which school level factors can impact on a variety of student outcomes of schooling, both cognitive and non-cognitive; the Dynamic Model of Education Effectiveness (Creemers and Kyriakides, 2006; 2008). The Dynamic Model was used to design interventions to reduce bullying that were based on substantial interventions focused at the whole-school level, both in terms of policy changes and improvement to the school learning environment, achieved through engagement with the whole school community (Kyriakides et al, 2013). SEAL also advocates a universal approach to its adoption (Weare, 2004; Weare & Gray, 2003).

These school-level policy and learning environment factors, together with a measure of the capacity of each school to be self-evaluating, were evaluated using a teacher questionnaire consisting of multi-item scales with teacher responses given via Likert scales. CFA was used to test the multidimensionality of the research instruments for school-level factors, and so establish the construct validity of the measurement model. Like the CFA conducted by

Huebner et al (1999), Kyriakides et al (2013) found that their CFA produced a significant chi square value, suggesting the model did not fit well on this most stringent test, but that a range of absolute and approximate fit indices including the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) were within acceptable ranges according to thresholds reported by Hu and Bentler (1999). As confirmation, a single factor model, combining all the scales to a single latent construct was shown to be a substantially poorer fit to the data. When these school level factors were added to multilevel models of the students responses to the two OBVQ scales of on bullying activity they each added significant explanatory power for both the outcomes measures on the being bullied and the bullying others scales with positive polices, environments and capacity for self-evaluation being negatively associated with the prevalence of bullying. The resulting MLM showed 8.7% of the explained variance was between schools for the *being bullied* scale and 6.7% for the *bullying others* scale. The anti-bullying interventions based on the Dynamic Model of Educational Effectiveness were shown to be significantly effective at reducing the level of bullying activity reported via both OBVQ scales and also reduced the proportion of the explained variance at the school level still further to just under 6% which would indicate a positive outcome in terms of bullying being less of a factor to influence school choice, where the ability to exercise such choice exists.

Examining change over time

With measures of affective non-cognitive outcomes in the form of self-evaluations, and even peer- or teacher-evaluations, there is a key consideration as to whether increases in measures are always the most favourable outcome. Is there a point at which, particularly for self-evaluations, a higher score is a *less* desirable outcome? Linked to the discussion of bullying as an outcome of schooling, Baumeister et al (1996) had concluded that "highly favourable self-appraisals [of self-esteem] are the ones most likely to lead to violence" (pp. 7-8). They linked this to the issue of a negative evaluation by others resulting in a threat to the ego of the person with high self-esteem and the resulting discrepancy between the internal and the external appraisal of the person producing an effect that they term "threatened egotism" (p. 12). The resulting conflict would leave the individual with a choice, either to reject the external appraisal and thus experience negative emotions toward the person making the low appraisal which may result in aggression or violence, or to reject an over inflated self-appraisal which may result in negative emotions toward the self which in turn could lead to social withdrawal.

Salmivalli et al (1999) discuss this issue in a study involving both peer- and self-evaluations of self-esteem as predictors of participation in bullying situations by teenagers. The researchers utilised items from a pre-existing instrument, the long established Rosenberg Self-Esteem Scale (1965) selecting just four items from the full ten item scale, which is puzzling on the surface as the items are relatively short and one would imagine that having the necessary literacy skills was not an issue since the participants in the study were teenagers, but the Rosenberg self-esteem scale is scored using a Guttman scale (Rosenberg, 1979) and the three of the items selected by Salmivalli et al (1999) are scored as one item in the Guttman scaling, while the fourth is scored as an individual item. The internal consistency of the reduced scale, evaluated by Cronbach alpha was 0.64. The other two scales (peer-evaluation, 5 items, and defensive egoism, 3 items) appear to be new to the study. Exploratory Factor Analysis (EFA) of the 12 item instrument suggested that the three factors were indeed distinct with each item loading being only above 0.30 for its designated factor. The fact that the items would have been completed by two different sets of respondents (self-evaluations and peer-evaluation) perhaps raises questions about this approach to establishing construct validity. The data from the three scales were used to separate the adolescents in clusters. Those young people, predominantly males, with high self-evaluation of their self-esteem, high defensive egoism but low than peer evaluation of self-esteem (what the authors refer to as *defensive self-esteem*) was associated with bullying behaviour in terms of bullying others and assisting a bully. Whereas those with genuine high self-esteem (as evaluated by both self and peers) were associated with defending those being bullied. This is in line with Baumeister et al's (1996) view that a conflict between internal and external appraisals of self-esteem may lead to conflict that resolves itself in violent behaviour.

Large scale longitudinal school effectiveness studies

Introduction to the studies

Two of the most comprehensive effectiveness studies of both cognitive and non-cognitive outcomes of schooling have been undertaken by researchers working on longitudinal studies of school effectiveness. One of these is the LOSO project, Longitudinal Onderzoek in het Secundair Onderwijs (Longitudinal Research in Secondary Education), undertaken in the Flanders region of Belgium. The LOSO project began in September 1990 (De Fraine et al., 2005) as a large scale longitudinal study of schooling in the Flanders region of Belgium (see Van

Damme et al., 2002, for a summary of the LOSO project up to that time). The project was extremely ambitious in scope and, from the outset, collected data for both cognitive and non-cognitive outcomes of schooling. The other large-scale longitudinal study of educational outcomes is the Effective Pre-School, Primary & Secondary Education (EPPSE) project conducted in England (Sylva et al 1999). The EPPSE study, currently ongoing, tracks the progress and development of a sample of over 3,000 children from pre-school through to post-compulsory settings from 6 different local authorities located in 5 different geographical regions across England. Thus far the project has reported findings on the progress of the cohort at Year 2 (ages 6-7), Years 5/6 (ages 9-11) and Year 9 (ages 13-14), which is very similar to the span of this study, and is continuing to Year 11 (age 15-16) and beyond. As for the LOSO project, data was collected for both cognitive and non-cognitive outcomes of schooling from the outset of the study.

Development of measurement scales of non-cognitive outcomes of schooling in longitudinal studies

Prior to the LOSO project members of the research team had developed a wellbeing questionnaire instrument covering eight factors: interest in learning tasks, relationship with teachers, wellbeing at school, attentiveness in the classroom, motivation towards learning tasks, attitude to homework, academic self-concept and social integration in the class (Van Damme et al., 2002). Like many such questionnaires, each factor is measured using a multi-item scale (in this case consisting of 4-10 items per factor). The scales demonstrated good psychometric properties with Cronbach alpha values ranging from α = 0.63 to 0.88 (Van Damme, 1997) and later reported as 0.80 to 0.89 based on data from later within the LOSO study (Van Damme et al., 2002).

Gadeyne et al (2006) point to issues of measurement error being at least partly the reason why the size of the school effect for non-cognitive outcomes may be appreciably smaller than that for cognitive outcomes. They suggest this is possibly due to the reluctance to use multiple measures of outcomes through parent and teacher perspectives and observations, as well as students self-reports, but recognise the time and cost implications of such comprehensive approaches to data collection. They also point out that many previous studies have focused on a narrow set of non-cognitive outcomes clustered around the psychosocial factors such as achievement motivation and academic self-concept while relatively few studies have broadened the focus to variables such as those related to behavioural or social outcomes. To this end Gadeyne et al (2006) designed a study to collect both cognitive (reading, spelling and mathematics) outcomes and non-cognitive (behaviour problems) in kindergarten and first-grade children in primary schools in a rural locality of Flanders. The non-cognitive measures were teacher ratings of children's behavioural activity based on a pre-existing instrument known as the Child Behaviour Checklist which consists of eight scales some of which were combined to give four measures in all: *internalizing problem behaviour*, *externalizing problem behaviour*, *attention problems*, and *social problems* (*ibid*, 68). Details of internal and test-retest reliabilities are provided by the researchers and a range of validity checks are referred to as being demonstrated, but with no measures of validity provided in the methodology.

Gadeyne et al (2006) conducted a two-level MLM analysis with students at level one and classes at level 2. The total number of classes involved in the study was 22. This figure is less than the preferred number of level two units required to minimise the presence of bias in estimates of the standard errors of variance at the highest level. Maas and Hox (2005) suggest that standard errors for variances at level two in a two-level model are underestimated if the number of level 2 units is substantially less than 100. With a sample size of 30 units at level two estimates are biased by 15%, which drops to 7.3% for 50 units at level 2. However, point estimates for variables in the model are correctly calculated, even for as few as 10 units at level two. Therefore the significance of the level 2 class effect in this study needs to be viewed tentatively. In null models for the three cognitive outcomes the class-level variance for spelling and mathematics was found to 18% and 24% respectively. Two of the four non-cognitive outcomes (attention problems and internalising problems) were of a similar order of magnitude (22% and 28% respectively). Each of these level 2 variances were significant at the 99% level which provides something of a buffer in terms of the potential bias in the standard errors of level 2 variances reported by Maas and Hox (2005). The other variances for spelling outcomes and for social problems and externalising behaviour problems were either non-significant (reading) or significant at the 95% level and so more susceptible to Type II error. Although the proportion of the variance at the higher level in the model for 2 of the non-cognitive outcomes are in line with those for cognitive outcomes, it is important to note that in this study the highest level was the class, rather than the school and school level variances would potentially be lower, in line with other studies reporting smaller school effects for non-cognitive outcomes.

Gadeyne et al (2006) then introduced a number of background variables including a range of teacher ratings of the children's behaviours. None of the teacher ratings of student behaviours, when added individually to the null model, had significant explanatory power for cognitive outcomes, but they did demonstrate significant explanatory power for the non-cognitive behavioural outcomes (based on student self-ratings). The addition of all the teacher-rated factors to the model explained 74-92% of the class-level variance in problem behaviour (ibid: 74). When combined, the teacher-rated factors explained a smaller proportion of the class level variance in the cognitive outcomes of spelling and mathematics, but the improvement in model fit was not significant due, most likely, to the reduction in parsimony of the model through the addition of so many additional factors.

Finally, gain scores (in terms of progress from kindergarten to Grade 1) for the cognitive outcomes were calculated at class level, and added to the MLMs, “as a measure of teacher academic effectiveness” (ibid: 74). Gain scores for reading and spelling had no significant explanatory power for the class-level variance in the non-cognitive outcomes, but maths gain scores did demonstrate significant explanatory power for three of the four non-cognitive outcomes explaining 16 to 49% of the class-level variance in these outcomes (ibid.: 74). The relationship was positive with students making good progress in maths tended to demonstrate reduced behaviour problems in the higher grade class. Of note is the fact that the class level prior attainment measure for maths at kindergarten level was *not* a significant predictor of behavioural issues in Grade 1.

In the EPPSE study researchers focused on two sets of non-cognitive outcomes. The researchers utilised pre-existing Strengths and Difficulties Questionnaire (SDQ - Goodman, 1997) which yields measures of social-behavioural outcomes. Four of the SDQ factors were utilised in the EPPSE study: hyperactivity, self-regulation and pro- and anti-social behaviour. Importantly, in view of the critical observation of Gadeyne et al (2006), the study utilised teacher assessments of children’s social behaviours from the SDQ (Sammons et al, 2007b: i), rather than student self-assessments which are also available via the SDQ.

Similar data was collected from the EPPSE cohort at the end of Year 2 (age 7), Year 5 (age 10) and Year 9 (age 14) utilising the same four factors from the SDQ. Verifying the construct validity of the measurement model proved somewhat problematic. For example, in the analysis at the end of Year 9 (Sammons et al, 2011b), the construct validity of the four factor model was established using CFA techniques and though the resulting chi-square test statistic was significant (chi sq = 2990.576, df=431, p<0.001) the researchers considered that the goodness of fit of the final model was adequate (with NFI = 0.91 and RMSEA = 0.043), and “well within the conventional range of acceptability” (*ibid*: 104), although they do indicate that a fit of 0.95 for NFI would be considered to be a sign of superior model fit. There is no information given about any iterative steps in the model building process but the report does indicate that pairs of measurement error terms for individual items were allowed to covary wherever the software utilised to conduct the CFA indicated that the addition of a covariance term would result in significantly improved model fit (*ibid*: 103). No theoretical justification for the addition of covariance terms was provided. Inspection of a pictorial representation of the final CFA model shows that a total of 26 covariances between pairs of measurement error terms of individual items were added across the total 4 factor, 32 item model. 5 of these covariance terms were between items loading on different factors. Only three individual items out of the 32 were not involved in a covariance pair (*ibid*: 10). Such *post-hoc* modification of the measurement model to establish adequate model fit suggests the SDQ data was challenging to work with in terms of its psychometric properties.

The EPPSE study also employed student self-report measures of several factors, referred to as dispositions, from surveys during the longitudinal study; again at the end of Years 2, 5 and 9. The student self-report items were reported in each case using a four factor model: alienation (Year 2)/anxiety and isolation (Year 5/9); academic self-image; behavioural self-image and enjoyment of school (Sammons et al, 2008a). In the case of the Year 2 survey the researchers identified their model by first conducting a principal components analysis (PCA) using varimax rotation (explaining 43% of the total variance), followed by CFA. Once again, extensive *post-hoc* modification of the measurement model was required including discarding a number of cross loading items and factors with low Cronbach alpha measures, and utilising model fit modifications suggested by the software used to conduct the CFA. Despite these modifications the resulting four factors in the final model were found to have relatively low estimates of scale reliability with Cronbach alphas of 0.52 (3 items); 0.57 (3 items); 0.62 (3 items) and 0.69 (6 items). In the case of the Year 5 survey it was felt that the initial EFA/PCA yielded results that were unsuitable for model development and so items from a larger set were allocated to factors by the researchers and carried forward to a CFA, which after several rounds of modification, were deemed to have yielded an acceptably fitting factor model. In each case model fit was based on goodness of fit statistics rather than the more demanding test of a non-significant chi square statistic. In the Year 5 survey the Cronbach alphas for the four factors derived from the CFA were 0.62 (4 items); 0.74 (4 items); 0.62 (4 items) and 0.76 (7 items).

For the multi-item scales used to estimate student self-response to the disposition factors at the end of Year 9 the EFA/PCA followed by CFA procedure was followed again (Sammons et al, 2011c: 48-59). The data from a total of 58 items (utilising 3 and 4 point Likert scale response types) were carried forward to an EFA with varimax and oblique rotation of factors. The EFA yielded a total of 15 factors with Eigenvalues greater than 1.00 which accounted for 63-64% of the total variance in the dataset. Those items with a value of 0.40 or greater for the standardised factor loading were retained and any items loading on two or more factors were dropped from the model. After the item reduction steps the 6 factors with Cronbach alpha internal consistency scores greater than 0.60 were then carried forward to a CFA with each item constrained to a single factor. No indication is given as to whether or not covariances between error terms for specific items were allowed as there is no full specification of the final factor model. The final model was deemed to have acceptable fit (chi-square = 1656.12, df=449, RMSEA = 0.042, GFI = 0.93, CFI=0.94, n=1503). The resulting model reported a wider range of competences were collected than in previous waves of data collection, but the factors enjoyment of school, academic self-concept (now divided between maths and English) and anxiety were common across all three waves (Year 2, Year 5 and Year 9). Additional dispositions measured were referred to as citizenship values and popularity (Sammons et al, 2011c). The extent of model modification required to produce acceptably fitting measurement models for non-cognitive outcomes data indicates one of the extra challenges in working in this area.

Analysis of trends in non-cognitive outcomes over time

In a review of studies of non-cognitive outcomes a team of researchers working on the LOSO study (De Fraine et al, 2005) consider evidence indicating the attitudes to school become more negative during the early years of secondary education and that this is also the case with student wellbeing as an outcome. They point to psychological changes during puberty and changes in the school environment as explanation for the decline in these non-cognitive outcomes. Thus a strong negative association with age is to be expected when working with outcomes in the affective domain. De Fraine et al (2005) suggest that cross-sectional studies of students of a range of ages provide inadequate evidence of changes in wellbeing over time and so designed a longitudinal study of wellbeing to identify genuine trends in wellbeing over time and, as a secondary question, the extent to which these are associated with a school effect. They went on to analyse non-cognitive outcomes from the LOSO dataset of students in

Flanders. Data from a sample of nearly 3,800 students from 53 secondary schools (from the overall LOSO cohort of 6,000 students) were collected, and only included students who remained in their school for the duration of the period of interest and who were not required to repeat a year in the Flemish system. The wellbeing questionnaire instrument was administered four times in the first, second, fourth and sixth years of secondary schooling. Attrition due to respondent mismatch over the course of the longitudinal study resulted in a 54% response rate for all four survey occasions. The data were analysed using a multivariate method with time modelled using dummy variables, and also in a multilevel growth curve framework as measurement occasions (level 1) were nested within students (level 2) with schools modelled as the highest level (level 3). As well as identifying longitudinal trends in the wellbeing data, the multilevel growth analysis allowed the school effect to be determined (this evidence will be considered in the following section examining the presence of a school effect in non-cognitive outcomes).

In studies involving longitudinal data a key advantage employing a multilevel growth curve analysis is that it does not require individual participants to have a complete set of wellbeing measurements from all survey occasions. Any participant with at least one measurement can be incorporated into the analysis so avoiding bias due to missing data that isn’t missing-at-random and also the problems of attrition in longitudinal surveys (Snijders and Bosker, 1999: 181).

The multivariate regression model applied to the data utilising dummy variables for measurement occasion revealed a significant negative trend in wellbeing over time. This was confirmed by multilevel growth curve modelling of the data which also showed that a cubic growth (decay) model fit the data better than a quadratic model, based on the measured deviance (Akaike information criterion, or AIC) of each model from the data.

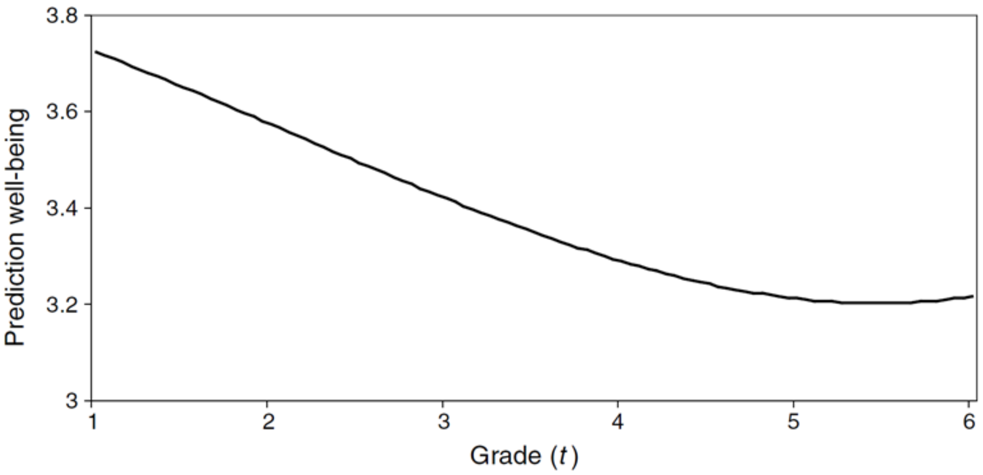


Figure 26: Cubic multilevel growth curve for wellbeing based on data from the LOSO study (De Fraine et al., 2005: 307)

The cubic model (Figure 26) suggests a modest upturn in wellbeing toward Grade 6, which is absent from the quadratic growth curve model that was applied to the same data, and the researchers indicate that, though the cubic model is a more reliable fit to the data, it doesn't lend itself readily to a simple interpretation in terms of a causal factor for the modest upturn at the end of the period of secondary schooling. As age was modelled coarsely by school grade it may, one might postulate, be an artefact of the coarse measure of time as opposed to using, say, age in months as the time variable. Multilevel growth curve models are able to deal with the range of variability in the time of measurement that this would introduce (Raudenbush & Bryk, 2002). Despite the fact that the AIC measure of model fit includes a component for model parsimony, it may be that the quadratic model is the more practically parsimonious of the two, despite having poorer fit to the data.

In a later study that was also part of the LOSO project Van de gaer et al (2009) analysed data using multilevel latent growth curve models across a 6 year period for a sample over 2,600 students from 50 secondary schools participating in the LOSO study. The focus was on two non-cognitive outcomes, namely *student motivation* and *academic self-concept*, with a view to determining the school effect on the growth observed in each outcome and the extent to which school effects for the two non-cognitive outcomes are associated. Measures, in the form of self-reports, of each outcome were taken on four occasions (grades, 7,8,10 and 12). Students who repeated a grade in the Flemish school system, or who changed schools or study tracks were removed from the sample to minimise the influence of grade retention or mobility on the findings. These groups of children might be considered to be particularly vulnerable, due the degree of disruption that grade retention or mobility could bring about and the corresponding impact on both cognitive and affective factors. It may have been better, therefore, to add factors to the model to adjust for the occurrence of grade-retention and mobility in the original sample. The researchers concede that this approach limits the generalisability of their findings to students taking a consistent route through secondary schooling ion Flanders. As with the previous study, the advantage of using a growth analysis allowed students to be included with an incomplete set of measures, and so reduce the loss of participants due to attrition over the course of the six year period. The growth curve analysis was conducted within a structural equation modelling (SEM) framework with latent variables included for the intercept (Grade 7 as time zero), linear and quadratic growth parameters. These were estimated for each outcome separately and also for a combined model incorporating both outcomes, allowing covariances between the growth parameters for each outcome to be estimated as an indication of the degree to which elements of the growth curve models are associated. Once again, model fit was based on absolute and relative fit indices rather than the more stringent criterion of a non-significant estimate of the chi square statistic for each model. The estimates for indices of model fit were close to but sometimes just outside the pre-selected cut-offs (≥0.95 for CFI and TLI [Tucker-Lewis Index], and ≤0.05 for RMSEA). The difference in chi square was used to estimate improvements between nested models (i.e. linear growth only and liner plus quadratic terms).

The chi square difference test showed that, for both outcomes, the quadratic models fit the data significantly better than the models containing only the linear growth terms. Both models produced a negative point estimate for the linear growth term indicating the fall in non-cognitive outcomes with age (grade) as students progress through secondary education, that has been observed in other studies (such as De Fraine et al, 2005). Though the quadratic models were significantly better fits to the data the point estimates for the coefficients of the quadratic terms were non-significant. Similarly only the variances for the intercept and linear term were significant at the student level, whereas the school-level variances for all three terms were significant. When the variance was partitioned 17.2% of the variance of the linear growth term in the model for motivation was located at the school-level. For academic self-concept it was as high as 50%. The figures below, for select schools from the overall sample, illustrate graphically why the proportion of the variance at the school level differed for the two non-cognitive outcomes.

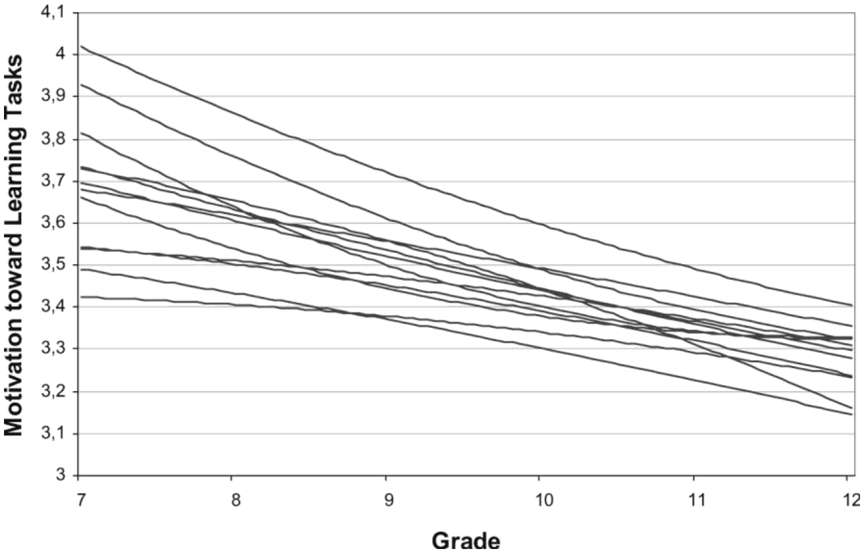


Figure 27: School-level growth curves of the motivation outcome for a selection of schools (Van de gaer et al., 2009: 244)

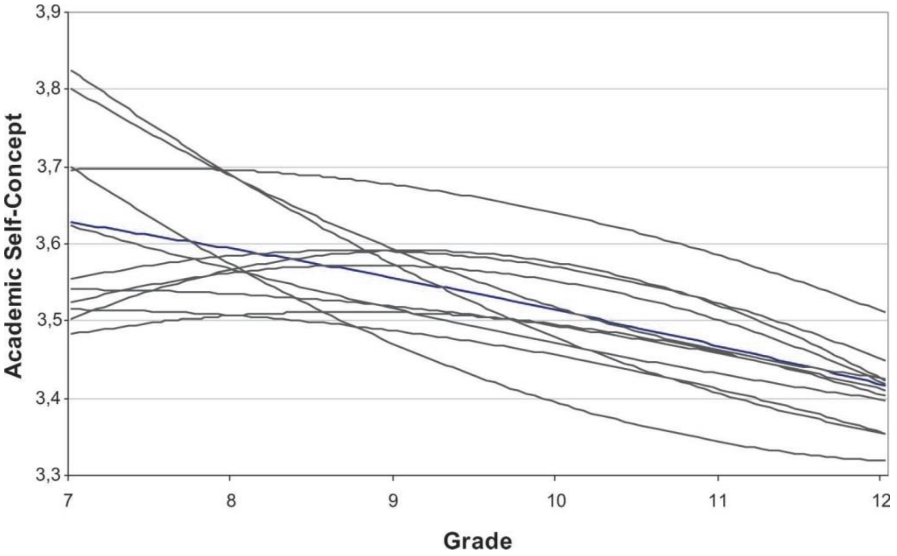


Figure 28: School-level growth curves of the academic self-concept outcome for a selection of schools (Van de gaer et al., 2009: 244)

In Figure 27 it is clear that the school-level growth curves for motivation tend to converge due to smaller variation between schools by the end of secondary schooling than there was at the beginning. Most of the schools appear to have similar shaped growth curves (or rather decline) over the course of Grades 7-12. For academic self-concept (Figure 28) the picture is quite different with some schools showing a decelerating declining trend while others demonstrate an accelerating decline with some schools even showing an increase in academic self-concept across Grades 7-9 before declining across the remaining grades. The common feature shown in both figures is that no school has a higher mean motivation or academic self-concept when comparing Grade 7 and Grade 12.

Van de gaer et al (2009) also ran a multilevel growth curve model combining both non-cognitive outcomes. This more complex model required some of the parameters for the quadratic growth terms (variance and covariances) to be fixed at the student and school levels in order for the models to converge. The covariances showed that both outcomes were significantly associated both in terms of the intercept values and the growth parameters. An identical pattern was found for covariances at the school level though these were not always significant at the 95% level which the researchers state is due to the low number of schools in the sample. This meaning that students tend to start secondary education rating themselves high in both outcomes or low in both outcomes, or schools tend to have similarly high mean values for both outcomes. The rate of decline in each outcome is similar for each student or school. They also found that the addition of two student level factors (gender and intelligence, both considered time invariant, argued cogently by the authors in the case of intelligence) had little effect on the covariances at the student level, but covariances at the school level were almost halved in value.

In the EPPSE study, for the social-behavioural outcomes of schooling from Goodman's SDQ, by the end of Year 5 (age 10) gender was found to be significantly associated with greater mean ratings of hyperactivity and anti-social behaviour for boys, who also received lower mean ratings of self-regulation and pro-social behaviours (Sammons et al, 2007b: 12-13). Gender also had the greatest effect sizes for hyperactivity and pro-social behaviours. By the end of Year 9 (age 14) these significantly negative differences in the ratings for boys compared to girls were still observed and remain significant even after adjusting for other student and family level factors such as socio-economic background and special educational needs status.

For the self-reports of student disposition outcomes the EPPSE researchers found that all four self-perception measures , namely alienation/anxiety and isolation, academic self-image; behavioural self-image and enjoyment of school, yielded more positive outcomes for the students in Year 2 than they did for Year 5 (Sammons et al, 2008a). This trend continued through to the end of Year 9 (Sammons et al, 2011c). This is in line with other research looking at similar non-cognitive outcomes over time described in this review and from the growth models of student self-ratings derived from the LOSO study (Van de gaer et al, 2009). Gender was associated with the disposition outcomes, with girls reporting significantly more positive mean scores for enjoyment of school and behavioural self-image but also higher mean self-ratings for the anxiety and isolation factor. There were no significant differences observed in the academic self-image factor by gender (Sammons et al, 2008a). For the Year 5 outcomes the researchers found that student level contextual factors, including some family background variables, have poorer explanatory power for these non-cognitive self-ratings than they do for academic outcomes (Sammons et al, 2007a), and for the teacher ratings of student social behaviours from the SDQ survey (Sammons et al, 2007b). By the end of Year 9 (age 14) there were still some significant gender differences observed in the disposition outcomes, with boys having significantly higher academic self-concepts for maths and self-rating of their popularity than girls, whereas girls reported higher level of anxiety and citizenship values. The gender differences for enjoyment of school and English academic self-concept were not significant at this stage (Sammons et al 2011c, 28).

Non-cognitive factors as the outcomes of schooling – is there a school effect?

Opdenakker and Van Damme (2000) utilised a multilevel (3 level) analysis of mathematics, language (Dutch) and wellbeing outcomes found a marked difference between the proportion of variance observed at the class and school levels for the two cognitive outcomes compared to the analysis for wellbeing as an outcome. Applying null models with no other dependent factors included in the model the outcome were as follows:

	Variance components /%		
	Maths achievement	Dutch achievement	Wellbeing*
school	19.6	32.5	6.5
class	23.2	23.2	4.0
student	57.2	44.3	89.5

Table 9: Comparison of the variance components for null models of three outcome variables (adapted from Opdenakker and Van Damme, 2000: 174-5)

* the wellbeing scale recorded here was one of 8 multi-item scales for non-cognitive outcomes reported in the study and is the scale which demonstrated the highest variance above the level of the student. The other seven ranged from 9.7% to only 5.4% of the total variance combined at class and school levels.

It is important to note that these models do not include any adjustment for a prior attainment measure in the analysis and so are not value added models which may well result in inflated values of the variance components for the two cognitive outcomes, which seem high in comparison with normally reported measures of the school effect for cognitive outcomes of around 10-15%. The researchers did have access to measures of numerical or verbal intelligence for the same students and after these were added to the models the variance components for maths and Dutch achievement at the school and class level fell to around half of their values in the null model which still left them higher than the variance component for the wellbeing factor in the unadjusted null model. By contrast the addition of a measure of achievement motivation as an affective explanatory variable for the non-cognitive outcomes in the study reduced the variance components at the school and class level to no less than two thirds of their values in the null models, with some barely reduced at all, and so this, what might be considered a pseudo prior attainment measure, had a much smaller explanatory power than similar prior attainment measures tend to have for cognitive outcomes. This may well be due to the range of non-cognitive factors included in the study here suggesting greater variance across the range of non-cognitive measures for an individual student which may have consequences for the reliability of non-cognitive measures constructed from several multi-item scales as these authors went on to do in a late study (Van Landeghem et al., 2002). The researchers also examined the presence of any significant evidence of a random slope in the models for the class and school level which would be a sign of differential effectiveness across the range of explanatory variables of numerical and verbal intelligence and achievement motivation. Achievement in maths and Dutch both showed evidence of differential effectiveness at the school level (through the presence of a significant random slope across the range of numerical or verbal intelligence respectively), but only maths achievement was differentially effective at the class within schools level. Of the 8 non-cognitive scales none showed the effective of differential effectiveness across the range of achievement motivation at the school level but four showed differential effectiveness at the class within schools level (wellbeing, relationships to teachers, interest in learning tasks and academic self–concept).

Landeghem et al (2000) reviewed a range of previous studies that included non-cognitive outcomes similar to those included in their study, including some of the individual scales from their previous work. The raw effects (based on null multilevel models for just two [students/school] or three [student/class/school] levels). Showed that the variance portioned at levels above that of the student is generally lower for non-cognitive outcomes than for typical 2 and 3 level school effectiveness studies conducted using multilevel analyse, though the proportion of variance at the class and or school level are nonetheless not insignificant, both in the statistical and practical sense.

Grp	**Non-cognitive outcome (details of study)**	**Variance component /%**			**Notes:** based on Van Landeghem et al. (2002)
		student	**class**	**school**	
1	Motivation (Grisay, 1996)	94.0	3.6	2.4	
	Interest in learning tasks (Van Damme et al., 2000)	93.3	3.6	3.1	*Part of the scale for the 'environment' outcome*
2	Image de soi scolaire (Grisay, 1996)	94.9	4.9	0.2	academic self-concept
	Zelfbeeld (Knuver, 1993)	99.0	0.0	1.0	academic self-concept
	Academic self-concept (Van Damme et al., 2000)	93.5	4.3	2.1	*The basis of the scale for the 'self' outcome*
3	Schoolwelbevinden (Elchardus et al., 1998)	94.9	---	5.1	wellbeing at school
	Wellbeing at the school (Van Damme et al., 2000)	87.8	8.2	4.1	*Part of the scale for the 'environment' outcome*
4	Relationship with Teachers (Van Damme et al., 2000)	89.5	6.4	4.1	*Part of the scale for the 'environment' outcome*
5	Schoolbeleving (Knuver, 1993)	90.0	6.0	4.0	School experience – covers both wellbeing and relationships to teachers
	Schoolbeleving (experience) (Hofman et al., 1999)	78.0	---	22.0	

Table 10: A comparison of raw effects from a number of European effectiveness studies of non-cognitive outcomes of schooling (adapted from van Landeghem, 2002: 446)

As can be seen from Table 10 above, the majority of these early studies of non-cognitive outcomes produced results in the range between 5 and 12% for the total variance components at the higher levels, above that of the student, (not including the studies with highest and lowest variance components at the higher levels).

Gray (2004) also conducted a review of 30 years of research in research in the British secondary schools context into "a range of affective, social and other non-cognitive outcomes" (ibid., p.185). His conclusion is that the school effect as indicated by non-cognitive outcomes is not as large as it is for cognitive outcomes, and the factors which typically provide explanatory power to traditional SE models do not have the same explanatory power to explain the variance in non-cognitive outcomes. The selection of studies for Gray's review required that the studies included at least one non-cognitive outcome and that there was also some form of variance partitioning so that the between school variance could be estimated. Gray (2004) reported that in studies including student attitudes to school as an outcome (whether students expressed a self-reported liking for school) there was a range of 5-9% of the variance at the level of the school. When prior measures of liking were adjusted for (in the way that prior attainment is modelled in school effectiveness studies) then the variance at the school level tended to drop by about half.

A study by Thomas et al (2000) of student attitudes in Scottish schools utilised four non-cognitive factors. Working with a total of 42 items across a four factor scale, they found that only 3 items showed a between school variance of 3% or more after adjustment for prior measures around 3 years previously. After adjusting for prior measures of each scale, and the addition of a range of independent factors with significant explanatory power, the between school variance dropped to just 1% for three of the four factors (including pupil engagement) and to 2% for the fourth factor, namely teacher support.

Van Landeghem et al (2002) conducted research on non-cognitive outcomes using school effectiveness designs as part of the LOSO project. The research team developed their earlier work on scale development and simpler multilevel models conducted on individual scales by examining four more complex, non-cognitive variables as the outcomes of schooling in the first 2 years of secondary education. The outcome variables were constructed as combinations of the multi-item scales developed in earlier work, and the modelling utilised the classic school effectiveness design applying multilevel modelling to facilitate partitioning of the variance in each outcome between and within schools. The outcomes were; the degree to which the student feels at home in the school environment ('environment'), the extent to which the students does his/her best in their school work ('work'), the academic self-concept ('self') and social integration in the class ('peers').

Where it differed from the approach employed by Thomas et al (2000) was that the multilevel analysis did not include prior measures of the specific non-cognitive outcome in the model, although there were explanatory variables such as achievement motivation and immunity to stress included in the models. The analysis also included other explanatory variables such as age (to account for students repeating a year), initial cognitive ability, a measure of family socioeconomic status, sex, and whether Dutch or another language is spoken at home. These factors were aggregated at first grade class, second grade class and school level, and variables indicating class size were also included.

Analysis of three level (student, class, school) null models (not including any other independent variables) showed modest amounts of variance explained at the class and school levels and that the class level explained more of the variance than the school level for the outcomes ‘self’ and especially for ‘environment’, and the school effect was larger for ‘work’ and only marginally so for ‘peers’. The proportion of variance explained at the class and school level in each case tended to be smaller than those for observed in similar models for cognitive outcomes (Van Landeghem et al, 2002: 447) and this is in line with previous research on non-cognitive outcomes (p. 446). They go on to discuss that temporary circumstances such as quarrels and changes of mood may give rise to more variance and greater measurement error for non-cognitive outcomes, which could be examined by deducing the test-retest reliability of the measures through a simple repeat test correlational study. The research team found that the introduction of the set of 7 background variables, indicated above, at the student level reduced the variance explained at the class and school level still further, depending on the specific non-cognitive outcome, with all seven background factors having significant explanatory power for the ‘self’ variable and sex is the only factor that explains significant variance in all four outcomes (with positive gains in outcomes for girls in all but the academic self-concept outcome). Group composition factors, made up of means of each of the seven background factors, reduced the unexplained variance at each level still further.

Konu et al (2002a) conducted a large scale survey of students from 458 Finnish secondary schools located across the country. They utilised a modified form of a pre-existing research instrument (the Beck depression inventory) to measure the subjective wellbeing of approximately 87,000 students aged 14-16, together with 56 dichotomous background factors at the student and school level divided into what the researchers termed background and school context factors. The analysis utilised multivariate linear regression rather than multilevel modelling which may result in the estimation of standard errors having a downward bias due to not taking into account the effect of clustering at the school level and so some factors being judged to be significant when, in reality, they are not, due to the estimation bias. The lack of MLM also resulted in a limitation in the analysis which meant it was not possible to partition the variance in the models and so determine a school effect for general wellbeing.

They found that their final regression model, which retained 36 factors, explained around 22% (boys) to 25% (girls) of the variance in the wellbeing indicator. The researchers also state that subsidiary models containing just one set of explanatory factors reveal that the combination of school context factors explain more variance in the wellbeing indicator than student background factors although I would suggest this may be an artefact of the higher number of dichotomous factors allocated to the school context set (35 out of the total of 56). There was a strong conclusion that student's grade (year group), and their socio-economic status had a lower effect on general subjective wellbeing than expected.

Interestingly, the same dataset was analysed in a multilevel modelling context by a modified team of researchers (Konu et al, 2002b). Their dataset was well suited to analysis using MLM with 458 schools and each school having no fewer than 70 pupils as participants in the survey research. 50 of the dichotomised background factors outlined above were used in this multilevel analysis at the student level and 48 were aggregated as the (percentage of students in the school giving the positive outcome) and centred to produce school level factors for the multilevel models. Of these only 35 student level and 10 school level factors had sufficient and significant explanatory power to be retained in the final models. Other factors were also introduced to the final MLMs including age (grade), gender, family structure and level of guardian's education. Through the use of MLM Konu et al (2002b) concluded that only 1% of the variance in the outcomes of the student scores for the General Wellbeing Indicator was located between schools. As the researchers point out, though this proportion of the variance is small, the difference in mean wellbeing score between the highest and lowest schools can still be substantial, which is analogous to the situation with models of cognitive/academic outcomes of schooling. In the analysis conducted by Konu et al (2002b) there was a 2 point difference in scores, on the 13 point scale for wellbeing, between the schools with highest and lowest means (p. 197). As Konu et al themselves concede, the use of a general wellbeing indicator, rather than one specific to the wellbeing in the school environment, may limit the proportion of variance in the dependent variable observed at the school level. The researchers speculate that the low level of between school variance may be due to the homogeneity of Finnish secondary schools, and also the fact that wellbeing, unlike academic outcomes, isn't a *direct* area of focus for schools and so is more of a subsidiary aim. It is interesting to speculate whether an initiative such as SEAL might shift the perspective in this regard, with the development of social and emotional wellbeing taking a more central role, though perhaps still not quite as central as academic outcomes. A longitudinal multi-level study, beyond the scope of the research described here, might reveal evidence of such a shift in the centrality of developing social and emotional skills, though this would clearly require a sustained and developed approach to the implementation of SEAL which, as it is non-statutory, will need to fight alongside the introduction of other statutory and compulsory initiatives which are vying for the attention of school leaders and teachers.

In their longitudinal study of the wellbeing of secondary school students using multilevel growth curve analysis conducted by De Fraine et al, the researchers concluded that the "influence of the school upon its students' well-being is only minor" (2005; 312). The multivariate model applied to the data, with age/grade modelled using dummy variables, showed school level variance for each separate measurement occasion was significant but small (3.8-7.4% of the combined student and school variance). The multilevel growth curve model (with measurement at level 1; students at level 2; schools at level 3) when modelled with a cubic polynomial growth function, showed significant variance between schools for the initial intercept in the model (4.8% of the total student and school level variance) but of the three growth factors only the linear term in the model was significant (2.4% of the total student and school level variance) with the quadratic and cubic growth factors in the model having zero variance at the school level, suggesting that between school differences in growth trajectories of students' well-being are negligible. Both the multivariate and multilevel models were for null models with no dependent variables added. Addition of explanatory variables would most likely serve to reduce the percentage of variance at the school level still further.

In the EPPSE study some modest but nonetheless significant school effects were also observed, both for the teacher ratings of students' social-behavioural outcomes, and for the self-reported student dispositions. The EPPSE study also examined the impact of school effectiveness (in terms of cognitive outcomes) and quality of previous and current schools attended on these non-cognitive outcomes as determining the longitudinal impact, particularly of pre-school education, was a key aim of the study.

For the social-behavioural outcomes of students derived from the SDQ, two-level null models (students nested within schools) of each of the four outcomes at the end of Year 5 (10 years old) were analysed. These yielded proportions of the variance in the non-cognitive outcomes at the school level between 4-12%, with self-regulation exhibiting the lowest school effect and pro-social the highest (Sammons et al, 2007b: 36). These were reported as being similar in magnitude to school level effects observed among the same students when they were at the end of Year 1 (5 years old). In line with other studies described in this review , these estimates of the school effect for non-cognitive outcomes are much smaller than the school effects of

19-21% estimated in null models for the same sample of Year 5 students in the cognitive outcomes of reading and mathematics (Sammons et al, 2007a).

The collection of teacher ratings of the social-behavioural outcomes during Year 1 as a measure of prior attainment for each of the outcomes enabled a multilevel value added model of progress to be estimated. In these models the proportion of variance at the school level was reduced by adjusting for prior ratings of each of the two negative measures: hyperactivity reducing from 7.2% to 5.9%, and anti-social behaviour from 4.9% to 3.3%. By contrast, the proportion of the variance at the school level increased slightly from 12.1% to 13.7% for pro-social behaviour, and almost doubled from 3.9% to 7.5% for self-regulation (Sammons et al, 2007b: 36-37) suggesting that the impact of schools' efforts to promote these behaviours across the primary phase becomes more pronounced when adjustment is made for the ratings of children's pro-social and self-regulation closer to entry to the primary phase of schooling.

When the quality of pre-school and the effectiveness of the primary school (in terms of academic outcomes) are added to models of progress in non-cognitive outcomes it is clear that increasing quality of pre-school and increasing effectiveness of primary school are associated in an additive way with the development of the positive attributes of self-regulation and pro-social competence, whereas the effects on hyperactivity and anti-social behaviours are more equivocal. Increased primary school academic effectiveness however, does have a marked association with the reduction of ratings of hyperactivity and anti-social behaviours for children who did not attend pre-school. One might postulate, therefore, that attending any quality of pre-school helps to buffer the possible differential effects of attending primary schools of differing effectiveness in these negative behavioural outcomes.

For social-behavioural outcomes data collected at the end of Year 9 (aged 14) null models for the teacher ratings of each of the four social-behavioural outcome factors were analysed to estimate the proportion of variance at the school level before adjusting for any other explanatory factors (Sammons 2011 et al, 2011b). For the two positive behavioural factors of self-regulation and pro-social behaviour the proportion of the variance at the school level was found to be 7.1% and 7.8% respectively, and for hyperactivity and anti-social behaviour the school level variance proportions from the null models were 7.3% and 6.1% respectively. This was a narrower range of school effects than the 4-12% observed in the null models for the outcomes at the end of Year 5. In common with all longitudinal studies the sample suffered from attrition and particularly so by this stage, 9-10 years after the initial wave of data collection. The respondents from the original cohort were also spread across a very wide range of secondary schools (just over 520) and were taught by teachers with little prior awareness or association with the EPPSE study. Multiple imputation methods had been used in the analysis of the end of Year 9 data across all outcome measures to deal with issues of unit non-response. When null models were run on the imputed data set the proportion of the variance observed at the school level were even lower (by 0.8-1.1%) for all but the self-regulation outcome, which was almost the same as for the non-imputed data set.

VA type MLMs adjusted for prior levels of social-behavioural outcomes from the end of Year 6, indicated development in these factors across KS3 (11-14 years old). In the VA type models the school effects for self-regulation dropped to 6.5% whereas the effect for pro-social behaviour remained almost unchanged at 7.7%. For hyperactivity the school effect from the VA model was estimated at 6.5% and for anti-social behaviour the school effect dropped to 3.5% (Sammons et al, 2011b: 139-142). Once again, these school effects for non-cognitive outcomes are appreciably smaller than the effects for cognitive outcomes in the same study (Sammons et al, 2011a) in which proportion of the variance at the school level was estimated between 18-24% for the various cognitive outcomes examined by the study.

As at previous time points, the researchers added school quality factors for secondary school attended into the models for the four factors from the SDQ teacher ratings of students' social-behavioural outcomes, together with quality factors for the pre-school and primary school attended. High quality pre-school education (compared to either low quality or no pre-school experience) was found to have a modest but nonetheless significant association with improved ratings of students' social-behavioural outcomes from the SDQ even at the end of Year 9. This pre-school quality effect appeared to be moderated by the quality of the home leaning environment, so that even low quality pre-school experience, for students with low or average home leaning environments was associated with higher ratings of pro-social behaviour and self-regulation. The quality of primary school attended had no significant association with ratings of social-behavioural outcomes after adjusting for student and family background, in contrast to analysis of academic outcomes for the same students (Sammons et al, 2011a) where primary school quality was found to have a significant effect on cognitive outcomes at the end of Year 9.

For secondary school quality, the researchers utilised two measures: a 4 year average CVA score, and also the most recent school inspection data from Ofsted reports. After adjusting for student and family background, the average CVA of the secondary school offered no significant explanatory power for the SDQ teacher ratings of students when combined with student level demographic and home background factors. The Ofsted ratings for schools rated at the lowest level (inadequate) for behaviour of learners was associated with poorer social-behavioural outcomes. Interestingly, interaction effects of secondary school quality (measured by Ofsted ratings) when combined with student background demographics suggested that attending a good or outstanding secondary school brought more benefit to students from advantaged backgrounds, in terms of their perceived social-behavioural outcomes, than for those from disadvantaged backgrounds. This was the opposite of the interaction effect that had been observed for attending primary school (Sammons et al, 2011b).

The research team do note the issues with inter-rater reliability of secondary school teachers' ratings of student behaviour (Sammons at al, 2011b: 103). Some of these issues relate to varying frames of reference in terms of school norms of student behaviour, which would also be pertinent to teachers providing ratings of students' behaviour in primary schools. For example, where behaviour of a certain type might be viewed differently by teachers in, say, a school with Ofsted ratings of good or outstanding compared to how the same types of behaviour might be rated by teachers in a school considered by Ofsted to be inadequate. The researchers found that students in inadequate schools were rated as having higher levels of self-regulation, on average, compared to teachers in satisfactory or good schools. The nature of teaching in secondary schools, with students being taught by a number of different teachers each day and across the week, also presents challenges for the reliability of teacher ratings. The researchers also indicate that care needs to be taken in interpreting school effects for the EPPSE sample as more than half (60%) of the schools had only one child from the sample in attendance at the school, with the average number attending each school being 2.6 children (Sammons et al, 2007b: 36).

For the student self-reported depositions of anxiety &isolation, academic self-image, behavioural self-image and enjoyment of school at the end of Year 5, two-level null models (students nested in schools) yielded a proportion of the total variance located at the school level of between 3 and 7% for all the variables except enjoyment of school which, perhaps unsurprisingly, was the variable with the highest school effect of with over 11% (Sammons et al, 2008a: 13). The addition of a range of student and home contextual factors to the MLM resulted in either negligible or modest reductions of the variance at both the school and student level and as a result, the proportion of the variance at the school level was either virtually unchanged from the null model (in the case of anxiety and isolation with 3.4% of the total variance and academic self-image with 7.3% of the total) or reduced by around 15% for enjoyment of school to 9.4% of the total variance, and reduced by 20% for behavioural self-image to 3.8% of the total variance (Sammons et al, 2008a: 13-14).

Further multilevel models were run with the addition of the year 2 self-ratings as prior measures of non-cognitive attainment in each disposition to form value added type models, and additional contextual factors were also added to produce CVA equivalent models. On the whole these resulted in further reductions of the variance at both the school and student level, especially in the case of the CVA equivalent model. In the CVA type models the proportion of the variance at the school level was no longer significant for anxiety and isolation and ranged from 3.5% for behavioural self-image through 6.6% for academic self-image and up to 9.5% for enjoyment of school. In the VA type models, only adjusted for Yr2 self-ratings from the related survey, the school effects were higher than in the CVA models, as one might expect, (except for academic self-image which was the same) ranging from 3.6% through to 10.9%.

For the self-rating dispositions data collected at the end of Year 9 (age 14) two-level null models for each disposition outcome were estimated and the proportion of the variance at the school level was found to be non-significant for four of the Year 9 dispositions (maths academic self-concept, anxiety, citizenship values and popularity). Only English academic self-concept and enjoyment of school had significant proportions of variance at the school level with 5.8% and 3.8% respectively. When a range of student and family background factors as explanatory factors were added to the MLM (i.e. adjusting for context only and not prior levels of non-cognitive outcomes) only English academic self-concept retained a significant proportion of the variance at the school level with 3.4% (Sammons et al 2011c, 24-25). In line with findings from these null and context only MLMs, simple KS2-3 value added MLMs, adjusting only for prior disposition measures at the end of Year 6, yielded significant proportions of variance at the school level for English academic self-concept (5.1%) and enjoyment of school (4.3%) (Sammons et al, 2011c: 44). When KS2-3 CVA type models were produced by adding in further student and family background contextual measures none of the dispositions yielded significant school effects for development of outcomes across KS3 (Sammons et al, 2011c: 45).

Core measures of pre-school quality utilised in the EPPSE study showed no significant association with students' self-ratings of the disposition factors. This was in contrast to academic outcomes and the teacher ratings of students' social-behavioural outcomes of the SDQ. Specific measures of pre-school academic effectiveness did show significant associations between highly effective pre-schools and the dispositions citizenship values and anxiety. As for the social-behavioural outcomes, the effectiveness of the primary school attended (based on school CVA score) made no significant difference on disposition outcomes, except that effectiveness in science was significantly associated with citizenship values (Sammons et al, 2011c: 37). Likewise, as for the social-behavioural outcomes, the effectiveness of secondary school attended (based on school CVA score) was not found to be associated with any of the disposition outcomes, but the Ofsted inspection judgments did provide some significant associations, particularly with the disposition enjoyment of school, which was found to be significantly associated to those Ofsted judgements related to learner achievement, learners' progress and the development of workplace and other skills, for schools that were rated outstanding compared to schools rated inadequate. The anxiety and maths self-concept dispositions were also found to be significantly associated with specific Ofsted outcomes. These positive associations may need to be treated with care however, as a number of other Ofsted judgments were found to have negative associations, with lower levels of anxiety and an increase of self-reported citizenship values for schools rated inadequate by Ofsted across six of the inspection judgments (Sammons et al, 2011c: 38-39).

Once again, the researchers indicate that the relatively high number of schools with very few EPPSE cohort children attending them causes issues for the reliability of these estimates. When two-level null model analyses were conducted on a subset of 66 schools with higher representation from the study cohort (mean =24) the proportion of the variance at the school level was greater and significant for all dispositions except anxiety, ranging from 11.2% for enjoyment of school through 6.2% for maths academic self-concept, 5.3% for English academic self-concept 4.7% for popularity, to 2.9% for citizenship values which suggests that the non-significant school effects in the KS2-3VA and CVA type models described above might be a result of the model specification issues created by having so few level 1 units (students) nested in many of the level 2 units (schools). Bell et al (2010) in a simulation study showed that a high proportion of level 2 singleton units (i.e. schools with only one participating student in the data) had little effect on the point-estimates of level 2 factors but did cause level 2 standard errors to be biased resulting in larger confidence intervals and increased Type 1 error rates, particularly where the number of level 2 units is low (around N=50). That said, the bias was negligible for simulations including around 500 level 2 units, which was the case in the secondary school phase of the EPPSE study described here.

These results from the EPPSE study provide further evidence that the school effects for non-cognitive outcomes measured by student self-ratings are generally weaker than those observed for academic/cognitive outcomes, except perhaps in the case of the factor enjoyment of school, which is almost in line with some of the more modest estimates for school effects on cognitive outcomes. Interestingly, the EPPSE study data shows that the same applies to school effects on teacher rated non-cognitive outcomes in the form of social-behavioural measures via the SDQ. The finding that the effects of high quality preschool can last through to age 14 for both cognitive and non-cognitive outcomes was a key finding for the project and is in line with a *post hoc* analysis, conducted by Heckman et al (2010), of the results from an experimental research design to determine the effects of pre-school programmes (the Perry Preschool study). The analysis demonstrated that, although the positive boost on IQ scores as a result of a two year pre-school programme diminishes after a few years there is a lasting significant effect on cognitive achievement (state standardised tests) and on non-cognitive skills (particularly traits which they termed personal behaviour and socio-emotional state). The pre-school programme focused on the development of social skills, organisation and planning as well as cognitive development. The gains in non-cognitive traits for the treatment group had particularly strong explanatory power on the variance on lifetime outcomes. The improvement in personal behaviour traits were associated with a reduction in crime for males and females, while improvements in the two non-cognitive domains and in cognitive outcomes for females were significantly associated with improvements in a variety of life time outcomes including high-school graduation rates and employment rates.

Summary

This chapter has provided a thorough overview of how school effectiveness research designs can be employed to measure the relative progress of students, and potentially the performance of classes and schools, in both cognitive and non-cognitive outcomes. It has indicated that the longstanding legacy of the school effectiveness data being made available to schools in increasingly sophisticated forms is not without its problems, especially in terms of capturing the group and school effects in a meaningful way that leads to a substantive interpretation that is useful for informing school improvement. Critical voices who argue that school effectiveness approaches add no interpretive value have been challenged, while at the same time considering the case for care must be taken in the use and interpretation of such measures.

The application of school effectiveness approaches to measuring non-cognitive outcomes has also been shown to be valid, and to produce meaningful information. At the same time it is acknowledged that the magnitude of school and class level effects are appreciably smaller for many of the non-cognitive outcomes studied in past research, and the methodological challenges even greater than for cognitive outcomes. With this in mind, we move to the research design and methodology for the study that is the focus of this thesis.

Chapter 4: Research Methodology

This chapter details the research design and methods that will be employed in the study. It begins by considering the post-positivistic research paradigm to which school effectiveness research belongs and the types of research designs that are common when working in this paradigm. Research questions are then identified and the survey design that will be employed is outlined, including a description of the main research instruments and the nature of the sample of schools participating in the study. The analytical approaches employed for determining the survey measurement model and then analysing the resulting non-cognitive outcomes are describes and justified. The chapter concludes by identifying the research questions for the focus on Family SEAL, and provides a similar consideration for the methodological and analytical approaches to be employed there.

4.1 The research paradigm of school effectiveness

Cohen et al. (2000) describe the nature of educational inquiry in terms of three lenses through which the practice of research might be viewed. The first encapsulates approaches which are related to a positivist perspective of the world, utilising scientific methods of inquiry, the second relates to an interpretivistic worldview that makes use of naturalisitic inquiry, and the third and final lens incorporates views from a critical theoretical perspectives (p3).

The positivist adopts a realist epistemology with a conviction that research is able to uncover an existing reality (Muijs, 2004). The positivist researcher values methods adopting objective perspectives in which the she acts as the external, detached observer of reality. The detachment is important to avoid the introduction of any bias introduced by the presence of the researcher. When such an approach is applied it is possible to develop theories and laws through strict observation and measurement and develop an understanding of reality by determining the relationships that exist between the gathered data. Such an approach is commonly termed the ‘scientific method’ in which these relationships may be proposed, tested and either verified or rejected through the on-going collection of data.

The scientific method lends itself to the adoption of either a deductive or inductive approach. In the deductive approach the research design flows from an *a-priori* theoretical framework and the aim is to apply the generalised view of theory to the specific context in which the research is based and confirm the applicability of the generalised theory to a specific context (Cohen et al., 2000). In an inductive approach the scientist acts as observer, seeking to gather as much data as possible in order to build up the necessary weight of evidence which allows patterns in the data to be identified within the specific context of the research which can then be generalised and facilitating *post-hoc* construction of a generalised theoretical framework. Cohen et al. (2000: 4-5) cite Mouly (1978) who contends that the common practice of science follows an ebb and flow of inductive and deductive approaches which produces:

> *…a back-and-forth movement in which the investigator first operates inductively from observations to hypotheses, and then deductively from these hypotheses to their implications, in order to check their validity from the standpoint of compatibility with accepted knowledge... This dual approach is the essence of the modern scientific method.*

This application of the scientific method to research in social settings leads to the classic approach of the dispassionate, disconnected researcher, designing and developing experimental techniques which gather data in order to provide evidence to develop or test theory. Correspondingly this objective perspective requires a deterministic view of human nature that facilitates measurement of latent constructs such as social and emotional competence, motivation or wellbeing through nomothetic methods (Burrell & Morgan, 1979) which through the application of measurement procedures, seek to find general rules, associations and, where possible, causal relationships between measured factors.

The ideal vehicle for positivistic research is the carefully designed experiment in which all possible variables are identified and then carefully controlled to facilitate the development (inductive), confirmation or rejection (deductive) of experimental hypotheses. Where experiments are conducted on complex systems with an extensive variables set (or in which all possible variables are hard to identify) the optimal design is the controlled experiment. Controlled experiments are particularly suited to researching effects in living systems due to the complexity of live subjects and the concomitant challenge of controlling all possible dependent and extraneous variables. The natural variation between living subjects requires that careful consideration be given to the selection of subjects for the experiment in order to avoid any selection bias. It may be possible to engage the whole population in the area of investigation and so negate the opportunity for selection bias in the design, but this may not necessary (nor expedient) if careful sampling of the population can be employed to ensure that the potential for selection bias is removed by having a representative subset of the population. Once the subjects for the study have been selected, either as the whole population, or as a representative sample, then allocation either to the control or the experimental groups is required. Random allocation of subjects to the control and experimental groups is normally undertaken in order to further avoid any opportunity for selection bias. Thus the staple experimental research design of the positivistic paradigm, when working in a social setting, is the randomised control trial (RCT). The RCT is sometimes referred to as the 'gold standard' experimental approach to research design, especially by those focused on determining the efficacy of some kind of treatment or intervention, such as a drug trial or the implementation of a new educational programme. The accolade of 'gold standard' has been critiqued within medical disciplines (Kaptchuk, 2001; Slade and Priebe, 2001) and Simon(2001) suggests that in certain contexts RCTs may be more appropriately considered to be a 'silver standard', especially where RCTs focus on a narrow patient group or exclude important segments of the population, which may create issues for generalising their results.

In social science settings such a detached, objective perspective to research as required by the positivist paradigm, can be difficult to establish. Genuine experiments in which all extraneous variables are controlled can be very difficult to design. Some argue that the notion of measurement in a social context is problematic, though not to the extent that it would require the whole research approach built on the foundations positivism to be deconstructed. Such a view is often referred to as *post-positivism*, which acknowledges the critique of positivism, particularly around the problems of measurement, without rejecting the notion of realism itself (Muijs, 2004). In the post-positivistic domain one may concede that any measurement of a social phenomenon is inherently flawed, but that it can, especially if robustly designed and rigorously validated, be close enough to be a useful representation of social reality to be of practical use. As Muijs (2004: 5-6) puts it

> *Rather than focusing on certainty and absolute truth post-positivist social science focuses on confidence – how much can we rely on our findings? how well do they predict certain outcomes?*

Some argue that this is what makes RCTs perfectly suited to research in the social sciences since the use of controls, in the right experimental design with randomly allocated subjects, serves to minimise the error that could result from a huge numbers of extraneous variable that are challenging to identify and even more challenging to control. Goldacre (2010) points out that RCTs have a longstanding heritage in education research and Forselunda et al. (2007), in a review of education research literature that sets out to determine when random allocation to treatment and control groups was first used in education research, located a claim that random allocation designs had been invented by education researchers such as E.L. Thorndike early in the 1900s (p373). Nevertheless, a range of philosophical, practical and pragmatic critiques of RCTs grew over time (Cook, 2007) leading to what Cook (*op cit*: 338) describes as "an anti-positivist and pro-constructivist turn" in education research. Hammersley (1997; 144) suggests that "much educational research in the first two-thirds of the 20th century was devoted to scientific investigation of effective teaching" and that the failure of this endeavour to establish a conclusive body of knowledge on the practice of teaching resulted in a shift away from positivist research to more interpretivistic methods which Hammersley (1997) dates back to the 1970s.

The importance of relevance

In recent times there has been a resurgence of belief in the importance of research as a knowledge base for informing educational practice and teaching as to be viewed as an evidence-based profession, encapsulated in Hargreaves' (1996) speech to the Teacher Training Agency. Hargreaves compared teaching and medicine as "profoundly people-centred professions" (p1) but with very different views of the "kind of science and so the kind of research, involved in each profession" (p1). As a result Hargreaves argues that there is a divide between theory and practice with the former accorded a low status beyond the initial education and training of teachers so that, in stark contrast to the work of medical practitioners, teachers engage in their professional practice, and are effective to a certain extent, despite being largely in ignorance of in any body of knowledge upon which their work might be based and, importantly, unconcerned with this state of affairs. The teaching profession has, Hargreaves claims "decoupled promotion from both practitioner expertise and knowledge of research" (p4) which, when combined with a tendency for senior leaders in schools to become detached from their previous experience of regular classroom practice results in leaders disconnected from both theory and practice.

Hargreaves contrasts the cumulative nature of medical research with a decidedly non-cumulative approach to education research which seeks to identify educational game changers or paradigm shifts rather than the steady accumulation of evidence to develop practice knowledge. Hammersley (1997), in response to Hargreaves' challenge, takes issue with the call for a simplistic recourse to a natural science model of cumulative knowledge building to be applied in social science contexts such as education. While he argues that the debate around the distinctiveness of social phenomena can and has been overplayed, Hammersley points to two key areas of difficulty, namely, the measurement issues in the context of social phenomena and validation issues to establish causal relationships, as key reasons why Hargreaves' call for research to provide and evidence base for practice is complex and problematic.

School effectiveness research, with its aim to determine the nature and magnitude of the contribution schooling makes to educational outcomes, sits squarely within the post-positivistic tradition. It seeks to gather evidence from across a section of society through measurement techniques and to generalise from specific cases at different levels of the education system (students, classes, schools etc.) to the wider population in order to determine a replicable set of common factors that operate at these various levels which result in raised or diminished outcomes over fixed stages of schooling. This is done, while recognising the complexity of schools as organisations and the students and teachers working within them, and so SER seeks to rise to the challenge of *understanding* the social reality of schools by developing models of schools and of schooling in order to establish a set of principles or practices by which highly effective schools may be recognised and characterised. SER also seeks to meet the challenge of *applying* this understanding of schools by providing a framework in which certain principles, policies and practices might be applied to improve the effectiveness of schools so that outcomes, especially but not exclusively, for students might be more equitably distributed.

4.2 Research Design

Both experimental (or quasi-experimental) and non-experimental research designs fall within the remit of the post-positivistic epistemology of SER. The relatively young field of research has focused predominantly on non-experimental methods in the form of large scale survey designs, in order to establish its understanding of the effects of schooling and latterly to build theoretical models of school effects (such as The Dynamic Model of School Effectiveness of Creemers and Kyriakides, 2008). Such survey designs allow our understanding of the nature of the reality of schooling to be built up and modelled in the form of relationships and associations between interplaying actors and factors within schools. A key limitation is that non-experimental designs make it difficult or impossible to determine the direction of causality in the relationships between such factors (Muijs, 2004). This study will employ a survey design in an effort to develop our understanding of the nature of non-cognitive outcomes in relation to the SEAL programme, as implemented in the schools of one local authority in England.

4.2.1 **Research questions**

With universal SEAL (**US)** in view the approach taken will be a large scale survey design, utilising a pre-existing questionnaire instrument, but developing the analysis of the data through a careful validation of the underlying multidimensional latent structure of the questionnaire. The aim of this stage of the research will be to establish a valid and reliable measurement model for some of the non-cognitive outcomes of schools within the participating the local authority, that are related to the SEAL programme. The measures derived from this step will then be carried forward to developing statistical models of these effects, and factors that help explain the variation in non-cognitive outcomes, using multilevel linear regression techniques that have been a staple method applied within SER, as outlined in the review in Chapter 3 of this thesis.

This design will help to answer the following research questions:

USRQ1: What is the nature and contribution of student and school level factors that explain the variation in SEAL related non-cognitive outcomes?

USRQ2: Do potential school- and class-level effects exist in non-cognitive outcomes related to SEAL?

USRQ3: How do the school- and class-level effects observed for these non-cognitive outcomes compare with school- (and class-) level effects observed for cognitive outcomes?

As an example of a more targeted approach to implementing aspects of the SEAL programme, the study of Family SEAL (**FS**) will utilise pre-existing questionnaire instruments to collect the parent and teacher ratings of students' emotional literacy in the five aspects of SEAL. A basic one group pre-test post-test analysis of the parent and teacher ratings will be adopted. Students participating in the Family SEAL programmes will be divided into two groups for comparison in terms of teacher's pre intervention perceptions as to the level of concern in terms of their social and emotional competence. The emotional literacy questionnaire data will be supplemented with some simple open ended questions to allow parents to report the benefits they perceive they and their children have received through engaging with Family SEAL.

This will facilitate collecting data to answer the following three research questions:

FSRQ1: Does Family SEAL result in improvements in the social and emotional skill level of participating children as rated by teachers and parents?

FSRQ2: Are the improvements in RQ1 greater for children identified by the class teacher (or school SEAL coordinator) as a cause for concern in terms of their social, emotional and behavioural development?

FSRQ3: What wider benefits do parents report after participation in the series of Family SEAL workshops in terms of:

- benefits for their children,
- benefits for themselves,
- benefits in their relationship with their children?

4.3 Utilising a pre-existing survey to develop a set of student self-rating scales for SEAL related skills and attitudes

4.3.1 Selection and implementation of the survey instrument

As discussed in the introduction, the aim of the research working in partnership with the Local Authority, was to assist school support staff in utilising currently available tools to evaluate impact, rather than generating new tools. Discussion with the LA's Behaviour and Attendance National Strategies Consultants revealed that two pre-existing surveys had been utilised to provide baseline assessments in a small number of the LAs primary schools. These were the *Emotional Literacy Pupil Checklist* (Faupel 2003) and the Key Stage 1 and Key Stage 2 versions of the *About Me and My School* questionnaire (Hallam *et al.* 2006). In the construction of the Emotional Literacy Pupil Checklist the authors had considered the considered the development of sub-scales based around the Goleman's (1996) five-dimensional model of Emotional Intelligence. The proposed sub-scales were self-awareness, self-regulation (of emotions), motivation, empathy and social skills. Due to the shared theoretical underpinning with SEAL the survey could have provided a useful tool for providing the opportunity to add a student voice dimension to the SEAL evaluation dataset. Reliability analysis of the proposed sub-scales, however, based on Cronbach's alpha values (0.34-0.61) for a sample of pupil responses suggested that it would be inappropriate to provide a sub-scale scores and so the Pupil Checklist only provides an overall emotional literacy score (Faupel 2003: 33).

The other surveys with available to the study with pre-existing data, the KS1 and 2 versions of *About Me and My School* , were originally designed as part of the evaluation of the Primary Social, Emotional and Behavioural Skills (SEBS) Pilot (Hallam *et al.* 2006) which had a heavy but not exclusive emphasis on what was the forerunner to Primary SEAL. In this Local Authority these surveys had been used by a small, self-selecting sample of primary schools to provide a baseline measure of social and emotional skills prior to implementing SEAL. The KS2 version of the survey consists of 40 statements rated on a 5 point Likert scale (see Appendix 1). The KS1 version consists of a subset of 25 statements from the 40 above (some of them slightly simplified) to which students respond simply with Yes, No or Don't Know. It was felt that a three point rating scale did not allow sufficient discrimination to apply factor analysis data reduction techniques as described below.

A few of the schools that had already used the *About Me and My School* survey had produced statement by statement analyses showing the distribution of responses given by students

(Figure 29). Some schools went on to produce break-downs by gender and by year group and/or teaching group. Such analyses had produced useful data for planning purposes to inform priorities for initial implementation of SEAL, but the utility of the data, both at the level of the individual student and for groups of students, was diluted by the large number of individual responses. For this reason it was decided to employ confirmatory factor analysis (CFA) of the KS2 questionnaire data to investigate its potential to produce a student self-rating 'score' for each of the five aspects of SEAL with a view to providing a tool for student self-assessment as they develop SEAL related skills. The resulting summary model derived from the CFA should simplify interpretation of the survey data and thus assist staff seeking to use the data to inform implementation and development of SEAL in their schools.

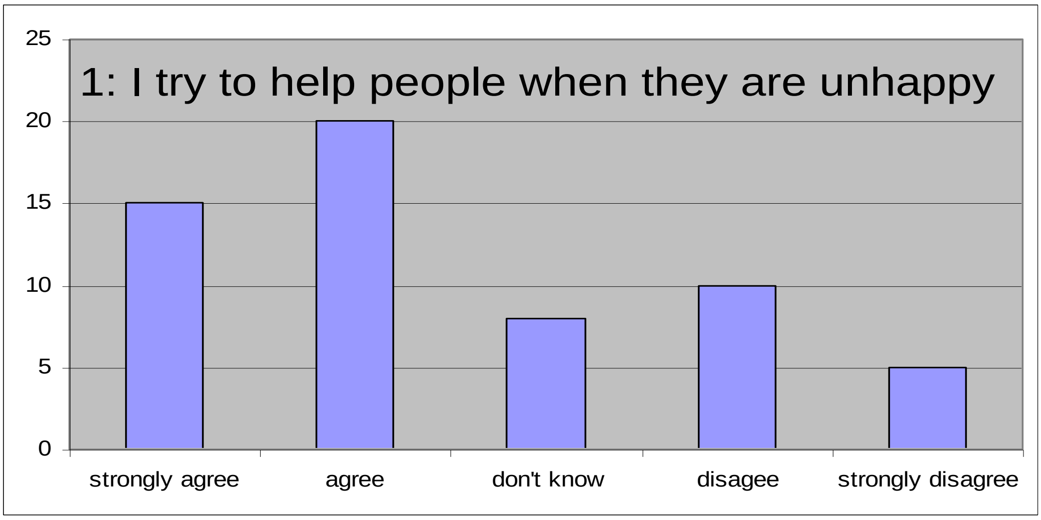


Figure 29: Typical histogram representation of responses to items from the KS2 "About Me and My School" survey

An initial round of data collection with 4-5 schools involved the survey being administered on paper and the responses transcribed into a spreadsheet. This approach was superseded by transferring the questionnaire to the web-based tool Survey Monkey so that the questionnaire could be administered online to groups of students. This obviated the need to transcribe the data and so reduced the risk of errors in transcription of the paper responses.

The nature of the research opportunity did not allow for multiple rounds of data collection using the questionnaire instrument, nor for any equivalent of a prior attainment measure of the non-cognitive dimensions to be collected, apart from only a very small subset of schools that repeated the survey at the beginning and end of an academic year. This set of schools were too few in number to allow multilevel modelling to be conducted on that limited sample, so the set of dependent variables that can be applied to the model would only contain contextual factors but no prior measures of each non-cognitive factor, unlike the traditional school effectiveness approach. This is not ideal but nonetheless provides scope to see if non-cognitive outcomes related to an initiative like SEAL demonstrate a potential school effect and can be utilised in situations in which pre-measures are not available. Lenkeit (2013) has a useful summary of the differences between models containing only contextual factors and those that also have prior attainment and multilevel growth models of progress.

As well as the analysis conducted for this study described below. All the student level results were returned to each participating school with some supporting documents to aid interpretation of the data.

4.3.2 Recruitment of participating schools

The schools that completed the survey did so over the period of about 18 months (from summer 2006 to winter 2007-08) and represented a self-selecting sample of schools from across the local authority that were implementing, or were about to implement the SEAL programme (either primary or secondary). Schools were recruited by the Local Authority National Strategy Behaviour and Attendance Consultants. In total 55 schools participated including a mix of first/junior, primary, middle and secondary schools and also two special schools for children with a high level of special education needs. Schools decided which students in their school should complete the survey, with the majority of schools selecting students from either a whole year group or whole key stage and about a third to one half surveyed the whole school. A summary descriptive analysis of the contextual data for the participating schools in provided in Chapter 5 (Section 5.2).

SEAL implementation characteristics in the participating schools

As a non-statutory initiative the SEAL programme allows schools a great deal of autonomy in the way that they implement SEAL. As described in Chapter 2 a key principle of SEAL is its universal nature which maintains that there are gains to be had by all students (and even by all members of the school community) through engagement with the SEAL programme. The schools engaging with SEAL in the LA were doing so in a phased way. This was, at least in part, a means of best utilising the limited support resources available from the team of Behaviour and Attendance (B&A) Consultants . Thus, by the time of the 2006-2008 data collection period for this study, there would have been a mix of early adopter schools together with those just embarking on implementation of SEAL, the latter being especially the case for schools in the secondary phase, as Secondary SEAL was only introduced by the Department for Children Schools and Families (DCSF) during the period of data collection. Unlike some programmes designed to address the development of social and emotional competence (e.g. Webster Stratton, 2004), there is no compulsory curriculum for SEAL and so implementation is both varied and fluid. In keeping with the National Strategies 3-wave model outlined in Chapter 2, resources are provided for schools to implement SEAL predominantly at two of the three waves (for all students (Wave 1) through universal provision, and to targeted groups of students (Wave 2), as opposed to students on a 1:1 basis). Materials are also provided for staff development and awareness raising. For primary schools the Primary SEAL programme (DfE S 2006) provided materials for all students across school years 1-6 (ages 5-11), whereas for Secondary SEAL (DCSF no date) introduction of SEAL materials was phased for school years 7 through to 9 (ages 11-14) although some schools did utilise the SEAL approach with older students in school years 10 and 11 (ages 14-16). Many schools in the primary phase utilised SEAL in discrete lessons with SEAL studied explicitly in these lessons. In some primary schools SEAL themes were often introduced and concluded through assemblies (using supplied resources) and sometimes promoted via student praise and rewards systems. For other schools SEAL was more implicit and taught as part of their wider personal social and health education (PHSE) programme. This was particularly the case in secondary schools, in which SEAL was likely to be incorporated within the existing PSHE programme through sessions taught either by pastoral tutors or designated PSHE teachers. As SEAL is related to the development of competencies to support learning through traditional cognitive aspects of the curriculum, a limited number of schools made this link explicit by, for example, inclusion of a SEAL based objective or learning outcome alongside the subject-based intended learning outcomes.

As well as the phased implementation of SEAL across the LA’s schools, some schools utilised a phased implementation within their school by only using SEAL with a specific year group. This would be true for primary phase schools and secondary phase schools. The summary descriptives for schools in Section 5.2 indicate the breadth of coverage of the SEAL Survey across the set of participating schools to give some indication of the proportion of schools taking such a phased approach.

Thus, the status and universality of SEAL among students and staff would vary from school to school. This clearly presents issues of programme fidelity and this study makes no attempt to determine programme fidelity. If an appreciable school effect can be observed then utilising appropriate measures of programme fidelity may well be necessary in any future evaluation of impact of SEAL teaching and learning on non-cognitive outcomes. Anecdotal evidence gleaned via attendance at regional SEAL development days led by the LA's B&A Consultants suggest that the extent and intensity of implementation of SEAL depends largely on the energy, enthusiasm and influence of the school SEAL coordinator and corresponding support and awareness of the school senior leadership, especially that of the Headteacher.

Determining whether there is a significant school effect, both in statistical and practical terms, will be important in considering whether a universal survey of student self-reported competences has the potential to evaluate and therefore inform the implementation and development at the class and school level.

In order to ensure the data collected for this study was available to the participating students and schools, as well as the analysis conducted for this study described below, all the student level results were returned to each participating school with some supporting documents to aid interpretation of the data.

4.4 Confirmatory Factor Analysis Methodology

The fundamental nature of confirmatory factor analysis (CFA) is its hypothesis-driven nature (Brown 2006) in which data is approached from a firm *a priori* framework based on evidence and theory. This is in contrast to exploratory factor analysis (EFA) where the data is approached with no such prerequisite theoretical structure in mind. The strength of CFA over EFA is the theory-led approach where a theoretical framework is confirmed through data analysis rather than relying on serendipitous discoveries from the sample under study which may or may not be representative of the population. As a result, CFA is the common method of choice for the development of scales by examination of the latent structure of test instruments – the determination of *construct validity*. CFA is also able to assess the construct validity of instruments in a robust way because of its facility to adjust for measurement error in the data set (*ibid*).

In order to keep in tune with the theory-driven nature of CFA each of the 40 items in the "About Me and My School" survey were assigned to one of the 5 aspects of SEAL based on Goleman's (1996) five-dimensional model of emotional intelligence. This allocation (Table 11) was carried out independently of the framework of latent constructs associated with the survey by the original research team (Hallam et al*,* 2006) as at that stage in the research, the original source of the survey had not been identified. This served to facilitate model comparison to ascertain whether the SEAL based model proposed in this paper proved to be a better fit than the original model employed by Primary SEBS evaluation research team.

Item	Wording of item/statement	SA	MF	Mot	Emp	SSk	Att
Q1	I try to help people when they are unhappy.				✓		
Q2	I often forget what I should be doing.		✓				
Q3	I know what things I'm good at.	✓					
Q4	I often lose my temper.		✓				
Q5	I get annoyed when other people make mistakes.				✓		
Q6	I can usually describe how I am feeling.	✓					
Q7	I get upset if I don't do something well.			✓			
Q8	I find it difficult to make new friends.					✓	
Q9	I know when people are starting to get upset.				✓		
Q10	If I find something difficult I still try to do it.			✓			
Q11	I'm easily hurt by what others say about me.	✓					
Q12	I calm down quickly after I have got angry or upset.		✓				
Q13	Other children let me play with them.					✓	
Q14	I laugh at other children when they get something wrong.				✓		
Q15	I am usually calm.		✓				
Q16	I have lots of friends at school.					✓	
Q17	I find it easy to pay attention in class.		✓	✓			
Q18	I worry about the things I can't do well.	✓		✓			
Q19	I like my class.						✓
Q20	I work quietly in my class.		✓				
Q21	I want to do well in my work.			✓			
Q22	I sometimes leave the room without permission.		✓				
Q23	I get on well with my teachers.						✓
Q24	I sulk or argue when I am told off.		✓				
Q25	I can ask a question and wait for an answer.		✓				
Q26	I can take turns.		✓				
Q27	I listen well in class.		✓				
Q28	I am happy being me.	✓					
Q29	I am good at some things.	✓					
Q30	I can work without my teacher's help.		✓				
Q31	I get up and wander around the classroom.		✓				
Q32	Playtime is fun.						✓
Q33	Our teachers are fair in the way they treat us.						✓
Q34	It is easy to work in my class.			✓		✓	
Q35	I can talk to my teacher about anything.						✓
Q36	I am sometimes picked on or bullied by other children.	✓				✓	
Q37	I can tell the teacher if anyone is unkind to me.						✓
Q38	I sometimes bully or pick on other children.				✓		

Q39	I like coming to school.						✓
Q40	I like lunchtime.						✓
	Total number of associated items (shared items)	**7(2)**	**13(1)**	**6(2)**	**5**	**5(2)**	**8**

Table 11: Assignment of survey items to the aspects of SEAL

SA = self-awareness, MF = managing feelings, Mot = motivation, Emp = empathy, SSk = Social Skills, Att = attitudes to school and teachers

Measurement models for the latent constructs associated with the survey were constructed and analysed using Amos 7.0 (Arbuckle 2006). The chosen estimation method was maximum likelihood (ML) estimation. This method assumes a multivariate normal distribution for the observed variables in data set and also assumes that the observed variables are continuous (Byrne 2001: 70). The use of ML estimation with non-normal data can lead to the calculation of biased standard errors and an inflated chi-squared statistic for the model (Brown 2006). Extreme non-normality can result in incorrect parameter estimates (factor loadings) (*ibid*: 75-76). The data under study here are categorical and descriptive statistics showed that a number of the observed variables demonstrated substantial skewness (range -2.80 to 2.03 mean 0.62 [absolute values 0.01 to 2.80 with a mean of 1.10]) and kurtosis (range -1.28 to 8.75 mean 0.96 [absolute values 0.00 to 8.75 with a mean of 1.65]). Mardia's test of multivariate kurtosis gave a value of 337.2 with a critical ratio of 43.9 showing statistically significant evidence that the assumption of multivariate normality is not justified. Estimation methods have been developed that do not rely on the assumption of multivariate normality, such as the asymptotic distribution-free (ADF) estimator (Browne 1984) which is available in Amos. ADF estimation requires a very large sample size, a limit to the number of observed variables (<25) and also relies on the very strong assumption that a continuous, multivariate normally distributed latent variable underlies each categorical observed variable (Byrne 2001: 71). Chou, Bentler and Sartorra (1991) have argued that it is preferable to treat the categorical variables as continuous and to perform a correction on the chi-squared test statistic known as the Satorra-Bentler scaled statistic (Satorra & Bentler 1988; 1994). Amos does not have the facility to calculate the scaled test statistic. Therefore some *post hoc* modification to remove from the model the items with the highest kurtosis and/or skewness will be carried out and model-fit reassessed.

Model fit will be assessed in the first instance by the chi-square (χ^2) test statistic, for which a non-significant ($p>0.05$) χ^2 value is the indication of a well-fitting model as the difference between the model under test and the actual covariance pattern in the data is non-significant. The χ^2 test is a very sensitive and with large samples has the power to detect very small deviations between the proposed model and the actual covariance structure in the data. As a result, a set of ‘goodness of fit’ indices have been developed. Although measures are not without their critics, extensive simulation studies have enabled appropriate cut-off values to be proposed (Hu & Bentler 1999) for assessing model fit. The use of fit indices to assess model fit is a source of controversy within the structural equation modelling community that utilised CFA as a first step to establishing a measurement model. Some SEM purists (see Hayduk et al, 2007) have argued strongly that model fit can only be evaluated using the chi square statistic, but this is not the majority view (McDonald & Ho, 2002; Miles & Shevlin, 2007) and the use of fit indices in studies with large sample sizes is widespread and was utilised in studies considered in the literature review in Chapter 3 (Huebner at al, 1999; Kyriakides et al., 2013).

In this study the cut-off values for good fit based on absolute fit indices will be considered to be:

- values of 0.05 or less for the root mean square error of approximation (RMSEA) index,
- confirmation of the value of the RMSEA being 0.05 or less by a non-significant ($p>0.05$) value of the parameter pCLOSE which is calculated from the 95% confidence interval for the RMSEA estimate
- values of 0.08 or less for the standard root mean square residual (RMSR),

and for the incremental fit indices:

- values of 0.95 or above for the comparative fit index (CFI) and the Tucker-Lewis index (TLI)
- the value of the goodness of fit index (GFI) and the normed fit index (NFI) will also be reported although these indices are now considered to be poorer incremental fit measures than CFI and TLI. (Hu & Bentler 1999)

When a model is estimated by Amos, the software produces modification indices (Arbuckle 2005: 111) to show changes to the model that would improve model fit. A subset of these indices suggest allowing error terms of items to covary (rather than the default position which assumes that the error terms are independent). Allowing pairs of responses to covary implies that there may be some common, underlying factor influencing both responses that isn’t explained by the assignment to an underlying latent construct. Such covariance may be due to common measurement error (for example, in a set of negatively worded items). *Post hoc* use of such modification indices goes against the *a priori* nature of CFA and is therefore controversial, sometimes being referred to as data-dredging, so only those responses that could be justified with theoretical underpinning were allowed to covary in this way such as items clearly measuring a related construct but assigned to different latent variables.

Once a measurement model for the data derived from the questionnaire has been established the resulting outcomes will be carried forward to analysis using multilevel linear regression techniques.

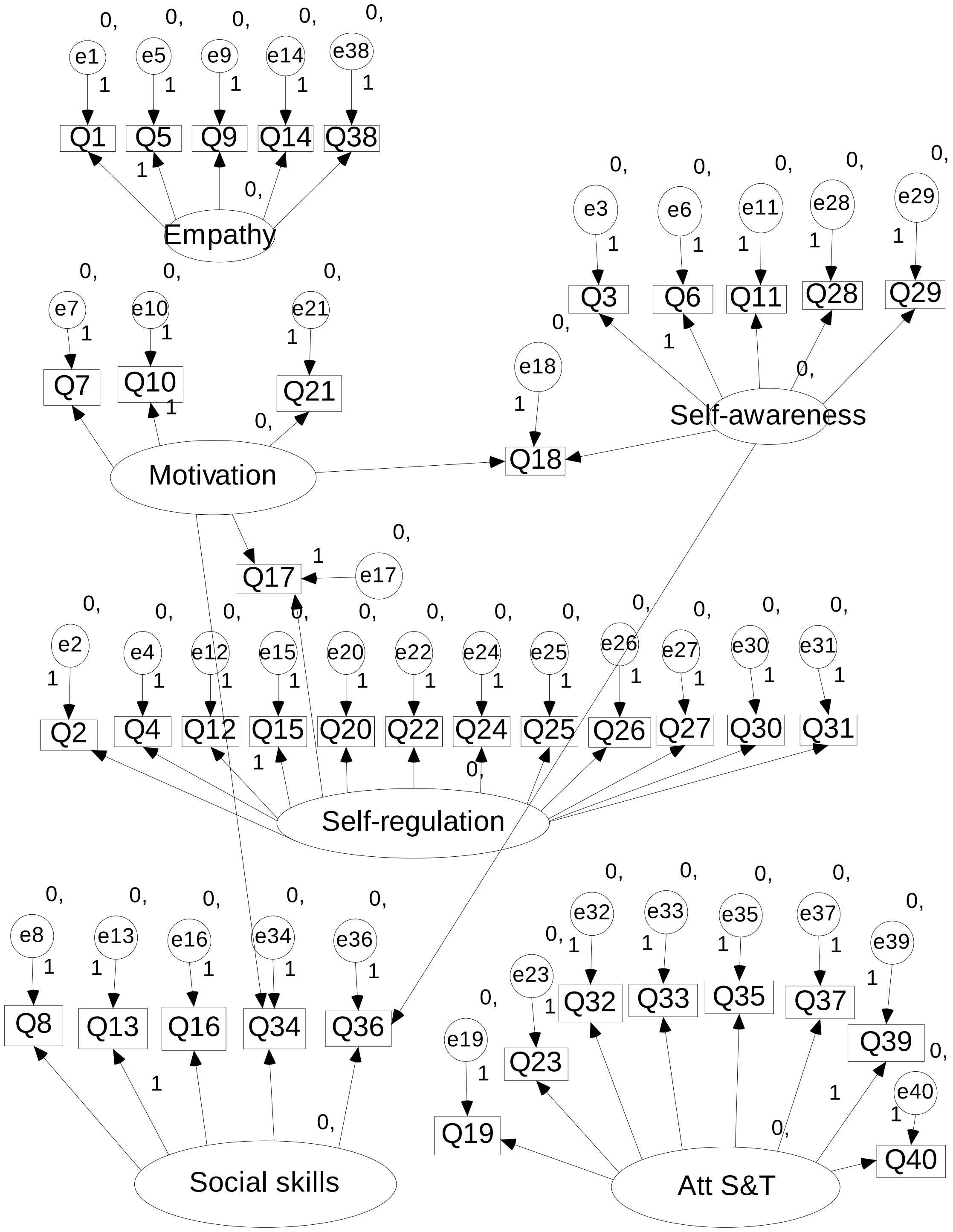


Figure 30: – Initial six dimensional CFA model based on assignment of survey items to the 5 aspects of SEAL plus the additional dimension of Attitudes to School and Teachers

4.5 MLM methodology

The factor loadings derived from the CFA will be used to provide a weighting for the contribution that each questionnaire item makes to its associated dimension, also as derived from the CFA (as a latent construct measured as a multi item scale). The student responses to each item, based on the five-point Likert scale, were multiplied by the appropriate factor loading after reversing the direction of the scale for any negatively worded items.

It was possible to calculate a minimum and maximum value for each dimension in the measurement model and these were used to scale the student scores for each dimension onto a common percentage scale so that a score of 100% represented the most positive combination of responses possible to the items associated with each dimension, and a score of 0% the most negative combination of responses. The scaling to a percentage score also aided interpretability when the student level data was returned to each school.

The resulting scores for each of the SEAL Survey dimensions derived from the measurement model were analysed using multilevel linear regression techniques (Snijders & Bosker, 1999) using the programme MLwiN 2.02 (Rasbash et al., 2005a).

The survey data is suitable for multilevel modelling (MLM) through analysis by multilevel linear regression techniques as the data are nested as follows (Figure 31):

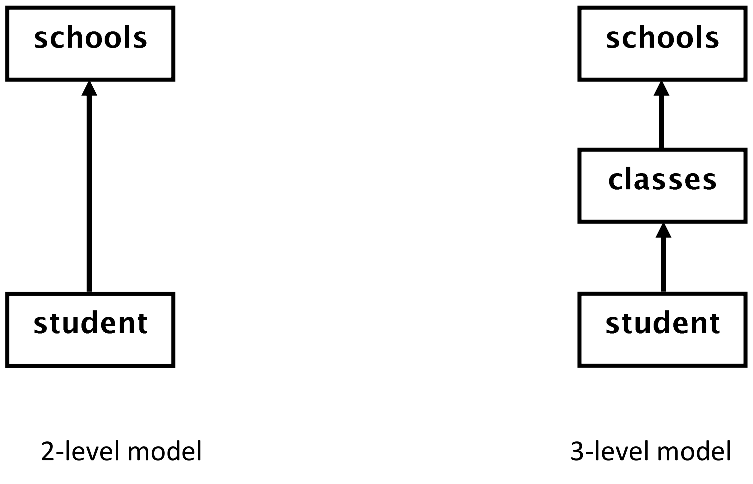


Figure 31: 2- and 3-level MLMs

In the first step of model building null models were produced for each of the survey dimensions validated through CFA. The coefficient of the intercept was added as a fixed effect, and the variance at each level in the multilevel model estimated as the random effects of the model .

For a 2-level model the null model represents:

$$y_{ij} \sim \mathrm{N}(XB, \Omega)$$

fixed part: $$y_{ij} = \beta_{0_{ij}}$$

random part: $$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$$

$$[u_{0j}] \sim \mathrm{N}(0, \Omega_u) \colon \Omega_u = [\sigma^2_{u0}]$$

$$[e_{0ij}] \sim \mathrm{N}(0, \Omega_e) \colon \Omega_e = [\sigma^2_{e0}]$$

For a 3-level model the null model represents:

$$y_{ijk} \sim \mathrm{N}(XB, \Omega)$$

fixed part: $$y_{ijk} = \beta_{0_{ijk}}$$

random part: $$\beta_{0ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk}$$

$$[v_{0k}] \ \sim \mathrm{N}(0, \Omega_v) \colon \Omega_v = [\sigma^2_{v0}]$$

$$[u_{0jk}] \sim \mathrm{N}(0, \Omega_u) \colon \Omega_u = [\sigma^2_{u0}]$$

$$[e_{0ijk}] \sim \mathrm{N}(0, \Omega_e) \colon \Omega_e = [\sigma^2_{e0}]$$

Models were estimated using the method of iterative generalised least squares, or IGLS (Goldstein, 2003).

Estimation of the null models allows the variance to be partitioned between each level in the random part of the model using the following formulae:

$$\frac{\sigma_{u0}^2}{\sigma_{u0}^2+\sigma_{e0}^2}$$

for the proportion of the variance at the school level in a 2-level model, and,

$$\frac{\sigma_{u0}^2}{\sigma_{v0}^2+\sigma_{u0}^2+\sigma_{e0}^2} \quad \text{and} \quad \frac{\sigma_{v0}^2}{\sigma_{v0}^2+\sigma_{u0}^2+\sigma_{e0}^2}$$

for the proportion of the variance at the class level and school level respectively in a 3-level model. Strictly speaking, these formulae represent the intra-class correlation for each model, but are often interpreted as the proportion of the total variance that occurs between members of the same higher level group in the model (Goldstein, 2003; Rasbash et al, 2005b).

Adding student level variables

As discussed in the previous chapter, the approach adopted in school effectiveness research is to add a number of independent or explanatory variables to models in order to move closer to determining the school effect. Generally additional factors should be measures of factors that are beyond the control of the school, or in the case of prior attainment measures, measures of competence or skill levels reached by the student prior to the period of interest. In the models for the SEAL related non-cognitive outcomes derived from the SEAL Survey student level factors will be added to the null-model including gender and age (by the coarse measure of school year or grade), as these have been shown to be key variables for cognitive outcomes as per the review in Chapter 3, and also based on previous studies of non-cognitive outcomes (Knuver and Brandsma, 1993; Konu et al, 2002b; Van de gaer et al, 2009) although some countries in the study of bulling as an outcome conducted by Kyriakides et al (2013) found such variables to be non-significant in their MLMs. After the addition of student level factors to each model in MLwiN the new models were estimated using the IGLS method. The fit of the model were examined using the deviance (-2*log likelihood) statistic reported by MLwiN (Rasbash et al, 2005b).

For a 2-level model with explanatory factors only at the student level the model is:

$$y_{ij}{\sim}\mathrm{N}(XB,\Omega)$$

fixed part: $y_{ij} = \beta_{0ij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} \ldots$

random part: $\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$

$$\left[u_{0j}\right]{\sim}\mathrm{N}(0,\Omega_u)\colon \Omega_u = [\sigma_{u0}^2]$$

$$[e_{0ij}]{\sim}\mathrm{N}(0,\Omega_e)\colon \Omega_e = [\sigma_{e0}^2]$$

For a 3-level model with explanatory factors only at the student level the model is:

$$y_{ijk}{\sim}\mathrm{N}(XB,\Omega)$$

fixed part: $y_{ijk} = \beta_{0ijk} + \beta_1 x_{1ijk} + \beta_2 x_{2ijk} \ldots$

random part: $\beta_{0ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk}$

$$[v_{0k}]\ {\sim}\mathrm{N}(0,\Omega_v)\colon \Omega_v = [\sigma_{v0}^2]$$

$$\left[u_{0jk}\right]{\sim}\mathrm{N}(0,\Omega_u)\colon \Omega_u = [\sigma_{u0}^2]$$

$$\left[e_{0ijk}\right]{\sim}\mathrm{N}(0,\Omega_e)\colon \Omega_e = [\sigma_{e0}^2]$$

As it was not possible to collect prior measures for each non-cognitive dimension the responses to the items “I’m sometimes bullied or picked on by other children” and “I sometimes bully or pick on other children” will be added as a type of student level proxy for prior wellbeing by measuring the student self- report of past association with bullying and antisocial behaviour, either as a victim or as the perpetrator. Opdenakker and Van Damme (2000) had used prior achievement motivation as a proxy for prior wellbeing in their analysis of student wellbeing.

Adding school level variables

The 2-level and 3-level models containing student level explanatory variables will be carried forward for the addition of school level variables. These will include the mean and spread (standard deviation) of both the *victim* and *bully* proxy variables discussed above as well as other school level factors. These will include measures of socio-economic status (SES), shown to be significant in SER models of cognitive outcomes and also some models of non-cognitive outcomes as discussed in Chapter 3. Finally, further set of school level factors, drawn from the Fischer Family Trust (FFT) database, based on the raw attainment and contextual value added outcomes at the school level for the years leading up to the administration of the survey, and also the trend in these school level attainment and progress measures over the same period, will be added. The choice of attainment and progress measures will be based on the measures that are high stakes for the students and the school, and published in national "league tables" of school performance. This will make it possible to determine whether association between such high-stakes school level measures of cognitive outcomes and non-cognitive outcomes may exist and facilitate an estimation of the proportion of variance in non-cognitive outcomes that they may explain.

For a 2-level model with explanatory factors at both the student and school levels the model is:

$$y_{ij} \sim \mathrm{N}(XB, \Omega)$$

fixed part: $y_{ij} = \beta_{0ij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + \beta_4 x_{4j} \ldots$

random part: $\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$

$$[u_{0j}] \sim \mathrm{N}(0, \Omega_u) \colon \Omega_u = [\sigma_{u0}^2]$$

$$[e_{0ij}] \sim \mathrm{N}(0, \Omega_e) \colon \Omega_e = [\sigma_{e0}^2]$$

For a 3-level model with explanatory factors at the levels of student (level 1) and school (level 3) the model is:

$$y_{ijk} \sim \mathrm{N}(XB, \Omega)$$

fixed part: $y_{ij} = \beta_{0ij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + \beta_4 x_{4j} \ldots$

random part: $\beta_{0ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk}$

$$[v_{0k}] \sim \mathrm{N}(0, \Omega_v): \Omega_v = [\sigma_{v0}^2]$$

$$[u_{0jk}] \sim \mathrm{N}(0, \Omega_u): \Omega_u = [\sigma_{u0}^2]$$

$$[e_{0ijk}] \sim \mathrm{N}(0, \Omega_e): \Omega_e = [\sigma_{e0}^2]$$

4.6 Family SEAL Methodology

The remainder of this section explores the approach that the Local Authority took during the implementation and evaluation of Family SEAL in a pilot group of seven schools.

4.6.1 Context and organisation

The pilot was a joint Local Authority project led by staff from the Extended Schools Service together with the SEAL Consultant from the Local Authority Primary Strategy Team. The initial focus was to engage schools which were already implementing SEAL at a school and curriculum level and which were also, for the purpose of securing funding, schools in challenging circumstances and therefore likely to benefit from intervention and support to engage with parents more fully. These schools were located in areas of relatively high social and economic deprivation as measured by a range of Local Authority indicators including the percentage of children entitled to free school meals. This funding enabled the Local Authority to support the pilot programme schools by providing them with an external facilitator who was experienced in working with parents. The external facilitators were recruited from Local Authority Educational Psychology Service, the Child and Adolescent Mental Health Service (CAMHS). The Educational Psychology Service and CAMHS both provide services to young people and their families. The Educational Psychology service works exclusively with young people in school settings to assess and give advice on individual children's learning, behavioural, emotional or social problems to parents, teachers and other professionals working with the children. CAMHS works as part of the National Health Service (NHS) in England and Wales has a broader remit considering the mental health needs of young people both in and outside of school. A detailed description of CAMHS services in England is available from the NHS Health Advisory Service (1995). Each external facilitator had experience in leading workshop-based approaches with parents. The schools identified at least one internal facilitator, such as the school SEAL Co-ordinator, a member of the teaching staff or a Teaching Assistant, with an interest in working with parents and also with strong communication skills.

From the outset the school-based facilitators appreciated the opportunity of receiving support from their external facilitator, as they felt less skilled in working with parents in workshop settings. Some school-based facilitators expressed their relief at having an 'expert' to work alongside them. At the end of the pilot all the facilitators were asked to report on the experience of collaborating on the project, particularly in the area of skill-sharing, in order to inform the development of future Family SEAL projects.

Of the schools originally approached to participate in the pilot, five felt in a position to take on Family SEAL and engage with the project. A further two schools, which were not deemed as being in challenging circumstances but had staff already experienced in working with parents in groups, also expressed an interest in running Family SEAL as part of the pilot and so representatives from these schools were also invited to attend the initial training session. Both additional schools were confident that they would be able to run the Family SEAL workshops without the support of an external facilitator and so two school-based facilitators were selected to lead the programme in these schools.

In order to support the schools during the project each school was paired with at least one other pilot school so that they could discuss issues that arose. In addition, each school was visited by the Local Authority's Primary SEAL Consultant in order to provide additional support, to share best practice from other schools in the project, to monitor fidelity of delivery, and to gain specific insight into the success of the project. The Primary SEAL consultant was not a mental health or Educational Psychology practitioner but had previously worked for primary school Behaviour Support Services both in this Local Authority and in an Inner City Local Authority with experience of supporting schools in implementing the wider Primary SEAL programme since 2005.

Allied with the recruitment of schools in challenging circumstances the original intention was for schools to target children that they would consider to be causing concern in respect to their social and emotional development and particularly those children from ‘hard to reach’ families who were not engaged with school. Such a targeted approach would have been in line with the ‘Wave 2’ intervention approach as outlined the National Strategies ‘waves of intervention’ model (see Figure 1 above). In practice schools were reluctant or found it impossible to restrict their focus to these children and families due to the level of interest shown by other families and also the difficulties in exclusively engaging ‘hard to reach’ families. Attendance at the introductory presentation was the result of a general invitation either to all the parents for a particular class or age group, or, in the case of very small schools, to all parents of children in the school. Participation in the series of workshops was then arranged for those parents who signed up on a voluntary basis. Thus the Family SEAL in the Local Authority pilot project was not strictly a *targeted* intervention but rather *universally* offered provision with voluntary take-up. This places it somewhere between waves one and two in the National Strategies model (Figure 1) as it is a group intervention but with participation open to all rather than a targeted group of pupils.

4.6.2 **Implementation and training**

The seven participating schools were invited to a half-day training event in the Autumn term, 2007. This was also the opportunity for internal facilitators in 5 of the 7 schools to be partnered with their external facilitator for the first time in order to discuss logistics, planning dates and times as well as the approach. Whilst the external facilitators brought key experience in working with parents, they were unfamiliar with SEAL. Also, both sets of facilitators required familiarisation with Family SEAL and so the training provided a key opportunity to share expertise and to develop and inform the approach to collaboration between external and internal facilitators.

The training introduced facilitators to the content of SEAL within the context of supporting emotional health and well-being. It also focused on developing partnership work with parents by highlighting core principles of mutual respect, collaboration, recognition, and by adopting a non-expert model in which both facilitators and participants are learning and sharing together. The training also explored the Family SEAL resource itself, the introductory presentation and the seven subsequent workshops based on the themes of Primary SEAL (see Table 1 above). Logistical issues such as space, time for planning and resource development were also considered and time was given for the facilitator pairs to commence with their initial planning and to start their partnership of working together. Finally, as measuring the impact of the pilot was central to the project, there was a focus on the instruments that would be used in the evaluation of the Family SEAL pilot.

4.6.3 Measuring the impact of the Family SEAL pilot project: the views of teachers and parents

As referred to in the introduction, some wide-reaching claims for the potential impact of SEAL have been made. For such a small-scale, short-duration pilot project as this it was important to have an appropriate focus for evaluating the impact of the project. Humphrey (2008, Humphrey *et al.*, 2008) makes the distinction between *proximal* and *distal* variables in the evaluation of school-based social and emotional learning (SEL) programmes. He lists typical *proximal*[10] variables of SEL programmes as measures of pupil or staff social and emotional skills, measures of school climate, and contrasts these with *distal* variables such as improved behaviour of pupils, a reduction in pupil exclusions, improved pupil attendance and increases in pupil attainment (Humphrey, 2008). For the purpose of evaluating the impact of this pilot project we have restricted the focus to measures of proximal variables only, and more specifically, to teacher and parent ratings of the social and emotional skills of the participating children collected before and after engaging with Family SEAL. This focused, quantitative view of impact was supplemented with qualitative evidence from parental evaluation questionnaires collected at the final Family SEAL workshop in the series. Ethical approval was received from the University Research Governance Office prior to the pilot study schools issuing their general invitation to parents to participate in Family SEAL. After parents were recruited for the programme parents were issued with information about the pilot research project and given an opportunity to discuss their participation in the research with their child and pose questions to the facilitators before completing a consent form. No parents recruited to the programme objected to participating in the pilot research.

4.6.4 Research Questions

FSRQ1: Does Family SEAL result in improvements in the social and emotional skill level of participating children as rated by teachers and parents?

[10] For the purpose of this study *proximal* outcomes are considered to be those most closely related to the nature of the intervention, namely the improvement of pupils' social and emotional competence, whilst *distal* outcomes are those associated with changes in the potential for learning for both the individual and the class that are further removed from the core aims of the programme.

FSRQ2: Are the improvements in RQ1 greater for children identified by the class teacher (or school SEAL coordinator) as a cause for concern in terms of their social, emotional and behavioural development?

FSRQ3: What wider benefits do parents report after participation in the series of Family SEAL workshops in terms of:

- benefits for their children,
- benefits for themselves,
- benefits in their relationship with their children?

4.6.5 **Research Design**

Parent and Teacher Emotional Literacy Checklists (Faupel, 2003) were used to evaluate the social and emotional skills of children participating in the Family SEAL pilot project. The checklists were issued to the parent or carer participating in the Family SEAL workshops and also to the class teacher of each child. Each checklist consists of a series of items linked to one of five dimensions (self-awareness, self-regulation, motivation, empathy and social skills – Figure 4). As this model is drawn from the work of Goleman (1996) the dimensions are very closely allied to the five social and emotional aspects of learning that form the basis of SEAL and so made the instrument a useful tool for reporting evidence of impact back to SEAL coordinators in a format they were familiar with and could be used to inform ongoing SEAL development in their school.

Since this evaluation of Family SEAL described here was designed and conducted the Emotional Literacy Checklists (Faupel, 2003) have also been used in an evaluation of small-group SEAL interventions (Humphrey *et al.*, 2008). In that study as in this one, a small proportion of participating children were younger than the specified age range for the instrument (ages 7-11). Humphrey et al. found that this did not significantly skew the data set (*ibid*: 34).

Faupel (2003) reports that scale reliability testing was conducted for the Parent and Teacher Emotional Literacy Checklists, derived from analysis of a pilot sample of 732 students from 28 schools. Cronbach alpha values for the five dimensional subscales range from α= 0.70 to 0.88 for the teacher checklist and α= 0.58-0.75 for the parent checklist. Faupel also developed a student self-rating checklist but this demonstrated poorer scale reliability (α= 0.34-0.61) and so the authors concluded that only an overall emotional literacy score (α= 0.76) could be generated from the pupil checklist. The content validity of the subscales for the teacher and parent checklists was confirmed by principal components analysis, combined with oblimin rotation of the identified principal components due to the high correlation observed between the components (ibid: 34-38). The lack of suitably reliable sub-scale scores from the pupil checklist made the Emotional Literacy Pupil Checklist less suitable for the purpose required here, namely to incorporate and element of the pupil voice to the evaluation dataset. However, the teacher and parent checklists have been used as part of this study for Family SEAL.

The same Primary school that had provided pilot data for the CFA model developed for the student self-rating SEAL skills survey had also utilised the Emotional Literacy Checklists to provide teacher and parental ratings of the social and emotional skills of selected students across the primary age range. Some students in each of the classes under study were identified a causing social and emotional concerns by their class teachers. These children were matched with randomly selected 'control' children to a maximum of 3 concern and 3 control students from each class.

One parent/carer of each student and the class teacher were asked to fill in the appropriate Emotional Literacy Checklists. For the purpose of analysis the subscale scores were converted to percentages as the number of items associated with each subscale was different for each version of the checklist.

Each of the 5 dimensions in the emotional literacy model is measured by 4 items in the case of the teacher checklist, and 5 items in the case of the parent checklist. Participants respond using a 4 point Likert scale. For the purpose of analysis these responses were converted into a numerical score of 4 for the most positive response through to 1 for the least positive response. The item scores are then totaled to give an overall score for each of the five dimensions and this score was then converted to a percentage/ratio of the maximum possible score to allow direct comparison of parent and teacher ratings of the children's social and emotional skills (FSRQs1&2).

Prior to commencing Family SEAL class teachers or SEAL Coordinators were asked to identify children who they considered to be causing concern in their social or emotional development. This information allowed comparisons of the changes in emotional literacy levels to be made between the children causing 'concern' and 'non-concern' children and so determine the potential for impact from Family SEAL as both a *universal* and *targeted* intervention (FSRQ2). The classification of children as social or emotional cause for concern was left as a matter of the class teacher's professional judgment but was informed by 'criteria for assessing emotional and behavioural development' issued by the Qualifications and Curriculum Authority (QCA - Table 12). The QCA is a public body charged by the Government with oversight of the development of the national curriculum and the administration of national assessment procedures.

Learning behaviour	**Conduct behaviour**	**Emotional behaviour**
1 Is attentive and has an interest in schoolwork	6 Behaves respectfully towards staff	11 Has empathy
2 Has good learning organization	7 Shows respect to other pupils	12 Is socially aware
3 Is an effective communicator	8 Only interrupts and seeks attention appropriately	13 Is happy
4 Works efficiently in a group	9 Is physically peaceable	14 Is confident
5 Seeks help where necessary	10 Respects property	15 Is emotionally stable and shows good self-control

Table 12: QCA Criteria for assessing emotional and behavioural development (QCA, 2001: 3)

At the end of the series of Family SEAL workshops parents were asked to complete a simple semi-structured evaluation questionnaire which allowed them to express what both they and their child had gained from being involved in Family SEAL (FSRQ3). This instrument consisted of 5 questions:

1: What do you feel you gained from being involved in Family SEAL?

2: What did your child gain from being involved in Family SEAL?

3: Which activities did you get the most from?

4: Is there anything about Family SEAL that you would like to see changed?

5: What will you do to try and make the most of what you have learned from being involved in Family SEAL?

Many of the Family SEAL facilitators expressed particular concern that this questionnaire be kept simple and open-ended in order to avoid causing anxiety for parents with limited literacy skills. Although session-by-session evaluation forms are provided in the Family SEAL materials

(DfES, 2006) most facilitators felt that making requests for written feedback at the end of each session would prove too onerous for parents and hinder the efforts of facilitators to establish relationships with the parents in the group. It was therefore decided that facilitators would not be requested to use the session-by-session evaluations forms.

4.7 Summary

This chapter discusses the research design employed in this study. It makes the case for locating the school effectiveness approach within the post-positivistic research paradigm and outlines the survey design employed for both universal and Family SEAL elements of the research. Research questions are identified to focus on measuring the non-cognitive outcomes related to the SEAL programme. The selection of participating schools and students are described, relying predominantly on self-selection by schools and then schools making purposive decisions about which students should engage. For Family SEAL there was also an element of self-selection among the families volunteering to take part.

The chapter went on the justify the application of confirmatory factor analysis (CFA) to determine an suitable measurement model for the SEAL Survey and the use of two- and three-level multilevel models (MLMs) to analyse the data on the non-cognitive outcomes derived from the CFA. The MLMs allow both school and class level effects to be estimated as well as testing whether specific student and school level context factors are significantly associated with the non-cognitive outcomes. This leads to the following chapter in which the results of the measurement model determination by CFA are presented, together with the MLM analyses of non-cognitive outcomes derived from the dimensions of the measurement model. The results of the Family SEAL pre- and post-intervention surveys will be included with the discussion of results in Chapter 6.

Chapter 5: Results

This chapter is divided into two substantive sections. In the first, the determination of the measurement model for the SEAL Survey, established via confirmatory factor analysis (CFA), will be presented. This section illustrates some of the challenges faced in working with data for non-cognitive outcomes particularly with data derived from self-report data, similar to those reported in Chapter 3 (section 3.2). It includes the results of *a post-hoc* exploratory determination of a measurement model from the full SEAL Survey dataset via principal component analysis (PCA) which is compared with the model development conducted via CFA.

In the second section the results of a series of two- and three-level multilevel models (MLMs) for each of the seven non-cognitive dimensions derived from the CFA are presented.

5.1 Results of CFA Analysis of the SEAL Survey

5.1.1 Analysis of the initial measurement model

Fit measure	6 dimensional SEAL baseline model	Original survey model*	Original survey model* (equal loadings)
Chi-square	1422.3	1417.2	1554.6.3
Degrees of freedom	721	650	682
Ch-sq/df	1.973	2.180	2.279
P	<0.001	<0.001	<0.001
CFI	0.690	0.650	0.602
GFI	0.763	0.746	0.708
TLI	0.665	0.622	0.590
NFI	0.533	0.511	0.463
RMSEA	0.065	0.072	0.075
pCLOSE	<0.001	<0.001	<0.001
RMSR	0.116	0.134	0.142
Information criteria (in order of increasing penalty for complexity)			
AIC	1620.3	1599.2	1672.6
BCC	1663.9	1636.9	1697.9
CAIC	2058.8	2002.2	1933.9
BIC	1959.8	1911.2	1874.9

Table 13: Model fit parameters for the SEAL baseline survey model and the models proposed by the original researchers (Hallam et al. 2006)

* In the original model two latent constructs were omitted ("awareness of own emotions" and "awareness of emotions in others" as these constructs were measured by responses to a single item. The CFA would not converge when such single measure latent constructs are included unless the error variance for the observed variable is set to a fixed value. There was no robust way of estimating the error variance associated with the survey items associated with this two single indicator constructs and so therefore the decision was taken to omit them from the model.

Global model fit comparison measures -

AIC = Akaike information criterion, BCC = Browne-Cudeck criterion, CAIC = consistent Akaike information criterion, BIC = Bayes information criterion. In every case, the lower the value of the criterion the better the global fit of the model.

Both the 6 dimensional SEAL based model and the original survey model represent extremely poor fit. Arguably, according to chi-square per degree of freedom and a variety of fit model indices, the 6 dimensional SEAL model is marginally better than the original model for the survey, but according to information criterion measures which allow model comparisons for non-nested models, the original model appears marginally better, although this conclusion needs to be set against the omission of the two latent constructs that are measured by a single survey item. When the factor weightings for the original model are all set to 1 (or -1 for negatively worded items) to force equal loading of items to each latent construct, which better reflects the application of the original survey model, the model fit is shown to be poorer still by all comparative measures except the two information criteria have the highest sensitivity for model parsimony (namely CAIC and BIC).

Since the evidence is mixed in terms of which model is best fitting the decision of which model might be best to carry forward for re-specification can be predominantly based on utility. The SEAL based model is clearly the more useful of the two models in terms of its ability to inform the implementation and development of SEAL in schools. Therefore, as a result of these analyses and the utility argument, it was decided to pursue re-specification of the 6 dimensional, SEAL based model to try and arrive at an acceptably fitting model.

Communalities (R^2 values) of the 40 survey items were inspected (see table x) and showed that several of the items had very low levels of variance explained by the model.

Item	Communality	Item	Communality	Item	Communality
Q6	0.028	Q11	0.202	Q39	0.327
Q7	0.046	Q36	0.203	Q33	0.328
Q9	0.049	Q18	0.204	Q34	0.336
Q32	0.094	Q22	0.21	Q28	0.346
Q37	0.098	Q31	0.214	Q19	0.386
Q5	0.098	Q26	0.222	Q25	0.388
Q3	0.106	Q10	0.228	Q16	0.425
Q40	0.131	Q4	0.236	Q20	0.436
Q1	0.14	Q8	0.244	Q23	0.437
Q21	0.147	Q24	0.246	Q14	0.478
Q30	0.151	Q2	0.249	Q27	0.619
Q12	0.163	Q15	0.264	Q13	0.644
		Q29	0.277	Q17	0.66
		Q38	0.277		
		Q35	0.278		

Table 14: Communalities (percentage variance explained) for the 40 survey items in ascending order of value.

With the exception of item Q5 as this was considered to be a core item for the self-awareness latent construct, the items with communalities below 0.1 (which represents items with standardised regression weights below 0.313) were deleted from the model.

Item	Associated domain	Standardised regression weight
17	MF	1.148
17	Mot	-0.440
18	SA	-0.475
18	Mot	0.036
34	SSk	0.173
34	Mot	0.501
36	SA	-0.256
36	SSk	-0.253

Table 15: standardised regression weights of cross-loaded items in the baseline SEAL model

The standardised regression weights for all cross-loaded statements were examined (Table 15) and cross-loadings deleted where there was a clear difference between the standardised regression weights for the cross-loadings. The modified model (designated *Mod1*) was analysed and the fit compared with the original SEAL based model (see Table 16). It is clear that this is a better fitting model by all measures but the fit would still be considered poor based on Hu and Bentler’s (1999) thresholds for fit indices.

Fit measure	**6 dimensional SEAL baseline model**	***Mod1* model**	***Mod 2* model**	***Mod 3* model 8 dimensions**
Chi-square	1422.3	1016.7	662.4	712.8
Degrees of freedom	721	544	361	406
Chi-sq/df	1.973	1.867	1.835	1.756
P	<0.001	<0.001	<0.001	<0.001
CFI	0.690	0.762	0.817	0.827
GFI	0.763	0.799	0.837	0.838
TLI	0.665	0.740	0.795	0.801
NFI	0.533	0.606	0.678	0.681
RMSEA	0.65	0.062	0.61	0.058
pCLOSE	<0.001	0.001	0.009	0.037
RMSR	0.116	0.106	0.112	0.101
Information criteria (in order of increasing penalty for complexity)				
AIC	1620.3	1187.7	810.4	892.8
BCC	1663.9	1220.1	832.9	922.3
CAIC	2058.8	1568.6	1138.2	1291.4
BIC	1959.8	1482.6	1064.2	1201.4

Table 16: CFA Model fit comparisons - baseline SEAL model and first three post -hoc modifications

5.1.2 The issues of non-normality

As discussed above, the non-normal nature of the data causes problems for estimation in CFA, particularly under ML estimation as used here. The items were ranked in order of kurtosis (which is more problematic for ML estimation than skewness) see table x, and the most problematic statements deleted from the model. These were the items considered to have a kurtosis >3.0. Although this choice was somewhat arbitrary these items also represent those with the greatest absolute value of skewness and, unsurprisingly, the most extreme mean values, suggesting that they fail to discriminate adequately between students. The resulting model, designated *Mod 2*, was analysed and the model-fit inspected (see Table 17).

Item	mean	sd	skew	Abs val of skew	se of skew	kurtosis	Abs val of kurtosis	se of kurtosis
Q21	4.654	0.767	-2.801	2.801*	0.161	8.747	8.747*	0.321
Q29	4.474	0.893	-2.170	2.170*	0.161	4.992	4.992*	0.321
Q26	4.482	0.826	-2.001	2.001*	0.161	4.662	4.662*	0.321
Q3	4.452	0.963	-2.027	2.027*	0.161	3.935	3.935*	0.321
Q38^{+}	1.548	1.034	2.032	2.032*	0.161	3.408	3.408*	0.321
Q40	4.382	1.062	-1.926	1.926*	0.161	3.158	3.158*	0.321
Q19	4.417	0.992	-1.897	1.897*	0.161	3.114	3.114*	0.321
Q1	4.202	0.809	-1.290	1.290*	0.161	2.835	2.835*	0.321
Q10	4.316	0.927	-1.605	1.605*	0.161	2.724	2.724*	0.321
Q22^{+}	1.623	1.082	1.850	1.850*	0.161	2.583	2.583*	0.321
Q16	4.386	1.028	-1.786	1.786*	0.161	2.579	2.579*	0.321
Q28	4.316	1.152	-1.723	1.723*	0.161	1.957	1.957*	0.321
Q14^{+}	1.662	1.109	1.619	1.619*	0.161	1.565	1.565*	0.321
Q33	4.149	1.084	-1.388	1.388*	0.161	1.493	1.493*	0.321
Q23	4.254	0.946	-1.284	1.284*	0.161	1.447	1.447*	0.321
Q32	4.254	1.093	-1.438	1.438*	0.161	1.287	1.287*	0.321
Q11^{+}	3.211	1.442	-0.205	0.205	0.161	-1.277	1.277*	0.321
Q8^{+}	2.575	1.501	0.454	0.454*	0.161	-1.244	1.244*	0.321
Q18^{+}	3.114	1.400	-0.137	0.137	0.161	-1.223	1.223*	0.321
Q7^{+}	2.825	1.359	0.099	0.099	0.161	-1.203	1.203*	0.321
Q36^{+}	2.610	1.408	0.327	0.327*	0.161	-1.171	1.171*	0.321
Q4^{+}	2.763	1.378	0.230	0.230	0.161	-1.152	1.152*	0.321
Q2^{+}	2.877	1.308	0.098	0.098	0.161	-1.062	1.062*	0.321
Q31^{+}	2.355	1.430	0.687	0.687*	0.161	-0.912	0.912*	0.321
Q12	3.583	1.394	-0.611	0.611*	0.161	-0.881	0.881*	0.321
Q20	3.425	1.227	-0.411	0.411*	0.161	-0.703	0.703*	0.321
Q35	3.513	1.254	-0.490	0.490*	0.161	-0.700	0.700*	0.321
Q37	4.083	1.175	-1.231	1.231*	0.161	0.623	0.623	0.321
Q39	3.750	1.371	-0.844	0.844*	0.161	-0.513	0.513	0.321
Q34	3.605	1.173	-0.528	0.528*	0.161	-0.464	0.464	0.321
Q15	3.904	1.191	-1.076	1.076*	0.161	0.425	0.425	0.321
Q30	3.636	1.203	-0.645	0.645*	0.161	-0.321	0.321	0.321
Q5^{+}	1.899	1.147	1.100	1.100*	0.161	0.319	0.319	0.321
Q9	4.048	1.079	-0.966	0.966*	0.161	0.286	0.286	0.321
Q6	3.789	1.102	-0.629	0.629*	0.161	-0.270	0.270	0.321
Q17	3.610	1.177	-0.691	0.691*	0.161	-0.244	0.244	0.321
Q25	3.904	1.231	-0.902	0.902*	0.161	-0.224	0.224	0.321
Q13	3.882	1.191	-0.891	0.891*	0.161	-0.110	0.110	0.321
Q27	3.868	1.202	-0.941	0.941*	0.161	0.044	0.044	0.321
Q24^{+}	1.969	1.243	1.058	1.058*	0.161	0.000	0.000	0.321
mean			-0.624	1.102		0.963	1.646	

$^{+}$ marks items set in a negative context. * significant ($p<0.05$)

Table 17: Descriptive statistics (mean, standard deviation, skewness and kurtosis) for data set

At this stage a substantial number of statements, 12 in total, had been deleted from the model the allocation of statements was reviewed.

Self-awareness

Q11 I'm easily hurt by what others say about me.

Q18 I worry about the things I can't do well.

Q28 I am happy being me.

Q36* I am sometimes picked on or bullied by other children.

Managing Feelings

Q2 I often forget what I should be doing.

Q4 I often lose my temper.

Q12 I calm down quickly after I have got angry or upset.

Q15 I am usually calm.

Q17 I find it easy to pay attention in class.

Q20 I work quietly in my class.

Q22 I sometimes leave the room without permission.

Q24 I sulk or argue when I am told off.

Q25 I can ask a question and wait for an answer.

Q27 I listen well in class.

Q30 I can work without my teacher's help.

Q31 I get up and wander around the classroom.

Motivation

Q10 If I find something difficult I still try to do it.

Q34 It is easy to work in my class.

Empathy

Q1 I try to help people when they are unhappy.

Q5 I get annoyed when other people make mistakes.

Q14 I laugh at other children when they get something wrong.

Social skills

Q8 I find it difficult to make new friends.

Q13 Other children let me play with them.

Q16 I have lots of friends at school.

Q36* I am sometimes picked on or bullied by other children.

Attitudes to teachers

Q19 I like my class.

Q23 I get on well with my teachers.

Q33 Our teachers are fair in the way they treat us.

Q35 I can talk to my teacher about anything.

Q39 I like coming to school.

* Item cross-loaded across two dimensions

Statements deleted from model

Q3 I know what things I’m good at.

Q6 I can usually describe how I am feeling.

Q7 I get upset if I don’t do something well.

Q9 I know when people are starting to get upset.

Q16 I have lots of friends at school.

Q21 I want to do well in my work.

Q26 I can take turns.

Q29 I am good at some things.

Q32 Playtime is fun.

Q37 I can tell the teacher if anyone is unkind to me.

Q38 I sometimes bully or pick on other children.

Q40 I like lunchtime.

Inspection of the items assigned to each dimension of the model suggested that the construct self-awareness was more akin to resilience, with the exception of the item Q28 – "I am happy being me". The item also had the lowest standardised regression weight of the four items assigned to this construct. Therefore item Q28 was removed from this dimension and the dimension was reassigned the name resilience and item Q7 "I get upset if I don't do something well" added to see if it weighted appreciably on the new construct.

The managing feelings construct had many items loading on it. It was considered that some of the items related to managing feelings of anger and frustration whilst the remainder were related to the students' self-management of their behaviours in the classroom. The Primary SEAL programme has a particular focus on the management of feelings of anger and frustration so it was felt that a dimension that highlighted this particular aspect of the students' responses to the survey would provide focused and informative data to inform programme implementation and development. Thus, this construct was divided with the items related to feelings of anger and frustration associated with "managing feelings" and the remainder associated with a new construct called "managing behaviour".

The cross loading item Q36 was also adjusted and the loading to social skills deleted

A new construct of self-awareness was considered using items Q28 "I am happy being me" and Q29 "I am good at some things" and item Q30 – "I can work without my teachers help" was reassigned to this construct as its regression weighting was the lowest (the only item loading <0.4) of the factors loading on the new managing behaviour construct.

The resulting model, designated *Mod 3*, was analysed and the model-fit inspected (see Table 16). Despite the addition of two extra dimensions/latent constructs and the reinstatement of some survey items this model demonstrated better, although still not acceptable fit, when compare to any other model developed thus far, based on chi-sq per degree of freedom and the full range of fit indices. The information criteria all showed poorer fit than the Mod 2 model but this is not surprising since the complexity of the model has increased with the addition of the extra constructs. It was therefore decided to pursue further respecification of this model and thus *Mod 3* became a new baseline for further model comparison.

At this point the modification indices available in Amos were inspected (sometimes known as Lagrange multipliers) as described in the methodology section above, were consulted. The standard default in Amos is to provide modification indices (MIs) above a value of 4.0 which represents a significant chi-sq change for a change of one degree of freedom. As chi-square is sensitive to sample size the value of the MIs are related to sample size. At this stage only those MIs with a value larger than 10 were consulted.

Several covariances were justified in terms of common wording of the items or through theoretical justification as outline in Table 18.

Added covariance	**Justification for acceptance**	**Value of covariance**
Q7 <--> Q18	Common wording of items Q7 I get upset if I don't do something well. Q18 I worry about the things I can't do well.	**0.33**
Q22 <--> Q31	Common outcome in behaviour Q22 I sometimes leave the room without permission. Q31 I get up and wander around the classroom.	**0.23**
Q18 <--> Q31	Theoretical link – attention span problems when frustrated with work Q18 I worry about the things I can't do well. Q31 I get up and wander around the classroom.	**0.26**
Q7 <--> Q31	Theoretical link – attention span problems when frustrated with work Q7 I get upset if I don't do something well. Q31 I get up and wander around the classroom.	**0.21**
Q35 <--> Social Skills	Theoretical link – Social Skills required to take teacher into confidence Q35 I can talk to my teacher about anything.	**0.20**
Added correlation	**Justification for acceptance**	**Correlation value**
Q13 --> Q19 correlation	Theoretical causal link – feelings towards class affected by willingness of peers to engage with child Q13 Other children let me play with them. Q19 I like my class.	**0.32**

Table 18: Covariance terms added to CFA model based on inspection of covariance modification indices

These modifications can be re-examined by running the model with another sample of data. They may not be generalisable to the wider sample that will be drawn from the LAs schools. Therefore careful attention will be paid to the significance of these added covariance terms when a larger, more representative sample of student responses is analysed using the CFA model.

At this stage the two dimensions showing the lowest factor weights were empathy and resilience so the models these dimensions were deleted from the model to investigate the effect on model fit (see Table 19). The *Mod 4* minus resilience showed substantial improvement in fit on every measure (although once again still below the acceptable fit criteria). The improvement in fit for *Mod 4* minus empathy was not substantial and the *Mod4* minus both resilience and empathy showed mixed fit results in comparison with the *Mod 4* – resilience only model (except in the information criteria statistics where the increased parsimony of the model with both constructs removed shows more substantial improvement in fit).

Fit measure	***Mod 3* model 8 dimensions**	***Mod 4* model covariances**	***Mod 4* minus resilience**	***Mod 4* minus empathy**	***Mod 4* minus both resilience and empathy**
Chi-square	712.8	616.6	455.2	518.7	378.7
Degrees of freedom	406	400	300	323	234
Chi-sq/df	1.756	1.541	1.517	1.606	1.618
p-value	<0.001	<0.001	<0.001	<0.001	<0.001
CFI	0.827	0.878	0.898	0.882	0.898
GFI	0.838	0.859	0.875	0.868	0.884
TLI	0.801	0.858	0.880	0.862	0.879
NFI	0.681	0.724	0.756	0.745	0.776
RMSEA	0.058	0.049*	0.048*	0.052	0.052
pCLOSE	0.037	0.592*	0.656*	0.362*	0.345*
RMSR	0.101	0.093	0.083	0.096	0.085
Information criteria (in order of increasing penalty for complexity)					
AIC	892.8	808.6	611.2	684.7	510.7
BCC	922.3	840.1	633.1	709.0	527.0
CAIC	1291.4	1233.8	956.7	1052.3	803.0
BIC	1201.4	1137.8	878.7	969.3	737.0

* Indicates statistic has crossed threshold for acceptable model-fit

Table 19: Model fit statistics for models introducing covariance terms and deletion of empathy and resilience latent variables

5.1.3 Testing of models on larger sample

A larger sample of students was gathered (n=1904 drawn from 11 different schools ranging from first schools, middle, primary and secondary from across the Local Authority and school years 1-8) and the *Mod 4* model analysed using this larger sample. A larger sample gives more statistical power to investigation the misspecification of the model and consideration of points of strain. This can be seen by the large increase in the chi-square value for the *Mod 4* model (from 616.6 for n=228 to 2453.7 for n=1904).

The covariance terms associated with possible theoretical links (shown in Table 18) were found to be non-significant with the larger sample so these were deleted. This would suggest that the significant effects signified by the coavariance terms in the data set from the single pilot school are not generalisable to the more representative sample of schools.

The residual covariance and standardised residual covariance matrices were investigated. Residual covariances are the difference between the sample covariances in the data and the implied covariances derived from the assumption that the model is correct. They therefore provide a measure of the departure of the model from the real data. In the standardised residual covariance matrix the residual covariances are divided by an estimate of their standard errors (Jöreskog & Sörbom 1984). Significant departures (p<0.05) of the model from the data have values greater than two (Arbuckle 2006). The variables with a pattern of high values of standardised residual covariances (in this case as high as 4.0-7.0) were deleted from the model. As these variables also included two of the statements associated with empathy this latent construct was deleted from the model. The new model was designated as *Mod 5*. The model still had a significant chi-square value (p<0.001) suggesting significant ill fit between the model and the observed covariance structure in the data set, but examination of the model fit statistics suggested that suggested that an acceptably fitting model had been deduced using all four goodness of fit indices identified above.

Fit measure	***Mod 4* model with covariances**	***Mod 5* model**	***Mod 5* without covariances on item Q24**	***Mod 6* Final Model**
Chi-square	2453.7	486.3	617.3	507.8
Degrees of freedom	400	146	148	146
Chi-sq/df	6.134	3.331	4.171	3.478
p-value	<0.001	<0.001	<0.001	<0.001
CFI	0.862	0.962*	0.947	0.958*
GFI	0.912	0.975	0.968	0.974
TLI	0.840	0.950*	0.932	0.945
NFI	0.840	0.946	0.932	0.942
RMSEA	0.052	0.035*	0.041*	0.036*
pCLOSE	0.052*	1.000*	1.000*	1.000*
RMSR	0.066*	0.036*	0.043*	0.038*
Information criteria (in order of increasing penalty for complexity)				
AIC	2645.7	614.3	741.3	635.8
BCC	2649.0	615.7	742.7	637.2
CAIC	3274.7	1033.6	1147.5	1055.1
BIC	3178.7	969.6	1085.5	991.1

* Indicates statistic has crossed threshold for acceptable model-fit (Hu & Bentler 1999)

Table 20: Model fit statistics for the final set of CFA models

An adjustment was made to the motivation dimension to exchange item Q34 “It is easy to work in my class”, with Q30 “I can work without my teacher’s help” as it was felt this, together with the statement Q10 “If I find something difficult, I still try to do it” would be a better indicator of motivation in terms of ability to work independently. The item "It is easy to work in my class" may draw out responses that draw more from the peer-effects in the social environment of the classroom rather than the more intrinsic view of motivation that is in line with Goleman's five-dimensional model. This resulted in the final measurement model designated as *Mod 6*. Despite the poorer model fit based on the goodness of fit indices (the value of TLI is just under the suggested threshold value of 0.95 although the other three values are still within the range for good fit) this adjustment will be retained as it places more weight on interpretability of the final model over the slightly diminished model fit.

5.1.4 Renaming the latent constructs in the CFA model

The titles of the latent constructs needed to be modified in the light of the fact that a number of items had been deleted from the original model.

The items linked to self-awareness are both focused on self-image rather than awareness of the students' own emotions which lies at the heart of the SEAL dimension of self-awareness so the title *self-image* may be more appropriate, or 'My feelings about myself'.

The items linked to *managing feelings* are focused on the management of feelings of anger and frustration. As highlighted above, this is in accord with the emphasis placed within the SEAL programme on the management such feelings within the wider context of developing one's own emotions. As a result, the title seems appropriate but it may be worth adding a sub-clause so that the title reads 'Managing feelings: anger and frustration'.

The *managing behaviour* dimension is a collection of items related specifically to class work rather than general behaviour around the school so the title could be helpfully modified to 'managing my behaviour in the classroom'.

The items linked to motivation are also focused on classwork but within the context of working independently so the dimension has been renamed *independence*.

The items linked to resilience are a blend of resilience in the context of both work and relationships so the general title *resilience* was considered to be appropriate although it will be important to make it clear that the balance of items and their factor weights is in favour of resilience in the context of school work.

The two items associated with the social skills dimension were narrower in focus than the overall aspect of social skills and so it was decided to re-identify this dimension as *friendships*.

The seventh dimension of attitudes to school and teachers broadly retains its original designation but, as three of the four associated items are focused on attitudes to teachers the title was adjusted to become *attitudes to teachers and school*.

Both the *Mod 5* and *Mod 6* models therefore represent acceptable fit against the criteria set out for goodness of fit indices (Hu & Bentler 1999). The Mod 6 model is chosen over Mod 5 due to the more informative association of items to the dimension *independence* (formally referred to as motivation). This model is summarised below in Figure 32.

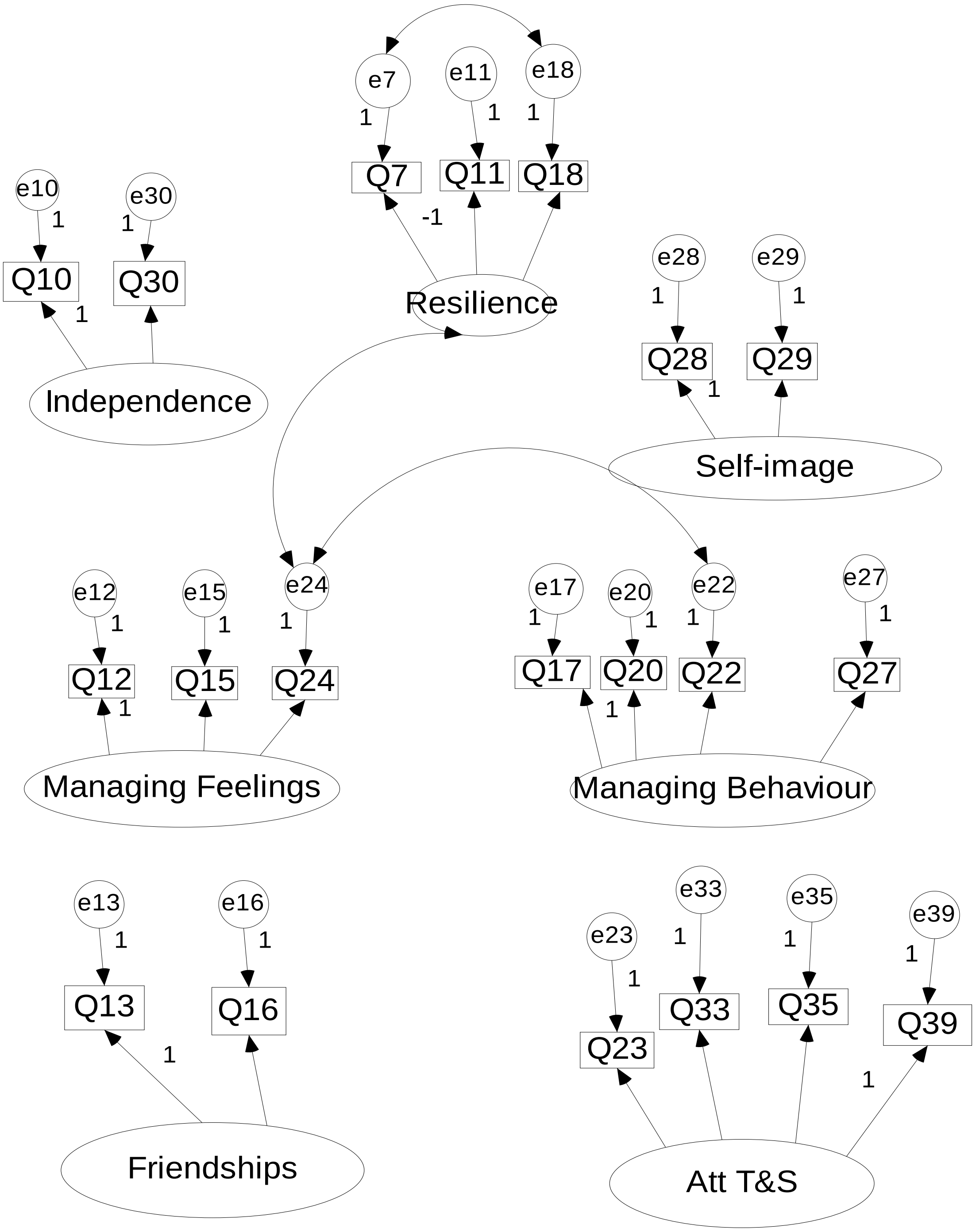


Figure 32: Final CFA model for the SEAL skills and attitudes student self-rating survey

5.1.5 Scale Reliability and problems with Cronbach's Alpha

Cronbach's coefficient alpha tends to misestimate scale reliability due to assumptions which constrain the values associated with measurement errors and factor loadings to be equal. Such conditions are rarely actualised in real data sets (Brown 2006: 338). Brown suggests an alternative scale reliability metric specifically developed in the context of covariance structure analyses that are the heart of the CFA measurement model (Raykov 2001; 2004). The generally accepted threshold value for a scale reliability is the same as that for Chronbach's Alpha, namely 0.7 .

Raykov's scale reliability is calculated as the proportion of true score variance to the total observed variance in the measure as follows:

Scale reliability = true score variance/total observed variance

$$\rho_Y = \frac{Var(T)}{Var(Y)} = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum \theta_{ii} + 2\sum \theta_{ij}}$$

Where

ρ_Y = scale reliability coefficient

λ_i = unstandardised factor loading

Θ_{ii} = unstandardised measurement error variance

Θ_{ij} = unstandardised measurement error covariance

SEAL model dimension	survey items associated in the model	standardised regression weights	Raykov's scale reliability for the dimension
Self-image	Q28 I am happy being me.	0.608	0.744
	Q29 I am good at some things.	0.553	
Managing Feelings	Q12 I calm down quickly after I have got angry or upset.	0.593	0.679
	Q15 I am usually calm.	0.689	
	Q24 I sulk or argue when I am told off.	-0.406	
Managing Behaviour	Q17 I find it easy to pay attention in class.	0.752	0.797
	Q20 I work quietly in my class.	0.642	
	Q22 I sometimes leave the room without permission.	-0.400	
	Q27 I listen well in class.	0.808	
Independence	Q10 If I find something difficult I still try to do it.	0.547	0.677
	Q30 I can work without my teacher's help.	0.355	
Resilience	Q7 I get upset if I don't do something well.	-0.497	0.751
	Q11 I'm easily hurt by what others say about me.	-0.603	
	Q18 I worry about the things I can't do well.	-0.527	
Friendships	Q13 Other children let me play with them.	0.689	0.841
	Q16 I have lots of friends at school.	0.764	
Attitudes to Teachers and School	Q23 I get on well with my teachers.	0.729	0.824
	Q33 Our teachers are fair in the way they treat us.	0.711	
	Q35 I can talk to my teacher about anything.	0.518	
	Q39 I like coming to school.	0.608	

Covariance terms	Justification	covariance	Correlation
e18<-->e7	Closely related contexts in survey items involving students' emotional response to not 'doing well'.	0.195	0.201
e22<-->e24	Related context of defiance towards teacher/authority/rules.	0.186	0.183
e24<-->resilience	Relationship between sulking/arguing when told off and emotional resilience.	-0.142	-0.241

Table 21: Standardised regression weights, covariance parameters and scale reliability values for the final CFA model

All the above standardised regression weights and covariance terms are significant ($p<0.001$).

5.1.6 Producing percentage scores for the student self-ratings for each of the seven dimensions from the survey model

Inspection of the standardised regression weights suggest that it would be inappropriate to equally weight the student responses to the survey items when calculating a score for each of the seven dimensions so the standardised weights will be used as factor loadings in calculating scores. This is processed as follows:

- Convert response on 5 point Likert scale to an ordinal numerical scale 1=strongly agree through to 5=strongly disagree
- Multiply the resulting ordinal response to each survey item by the corresponding standardised regression weight for the item.
- Add item scores together for the items associated with each dimension to give the overall score for each dimension.
- Determine range of possible scores for each of the seven dimensions using the combination of the most positive and least positive responses.
- Express each score as a percentage based on the position along the range of possible scores.

5.1.7 Post script – comparative exploratory analysis of SEAL Survey dimensions

The extensive iterative modification of the initial CFA model during analysis in order to determine an acceptably fitting model suggests that the data is problematic for analysis by CFA. Such extensive modification goes against the *a priori* nature of CFA. On recommendation of the examiners an exploratory approach to identification of scale dimensions in the SEAL Survey has been performed as a *post-hoc* comparison to see how the dimensional structure of the SEAL Survey determined via an exploratory approach compares to the confirmatory factor analysis described here.

Data reduction utilising principal components analysis (PCA), an approach analogous to exploratory factor analysis, was conducted on the full dataset. PCA was also utilised to determine the dimensions in the similar self-report data gathered as part of the EPPSE longitudinal study described in Section 3.2 above (Sammons et al, 2008a; Sammons et al, 2011c).

A number of checks on the raw data were performed. There were some significant correlations between a number of the 40 items in the survey, although none of these were particularly high and the determinant for the coefficients had a value of 0.00015 which is greater than the threshold of 0.00001, suggesting multicollinearity is not an issue for the dataset (Field, 2005).

After extraction of components by PCA the Kaiser Meyer Olkin (KMO) measure of sampling adequacy was 0.932 which is well above the minimum suggested threshold of 0.5, and the Bartlett's test of sphericity was significant (approx chi-square = 139646.2, df=780, p<0.001). The anti-image correlation matrix yielded KMO values for each individual item greater than 0.5 (in fact greater than or equal to 0.786 for all items). These tests indicate that a factor analysis approach is suitable for these data (Field, 2005: 648-652).

The scree plot for the PCA (Fig 33) suggests a three or possibly a five factor solution is appropriate which would account for only 34.9% (in the case of the extraction of 3 component) or 42.3% (for the 5 component solution) of the total variance in the data. Kaiser's criterion, including components with Eigenvalues greater than 1, would suggest extraction of 8 components might be appropriate, which would explain 50.6% of the total variance.

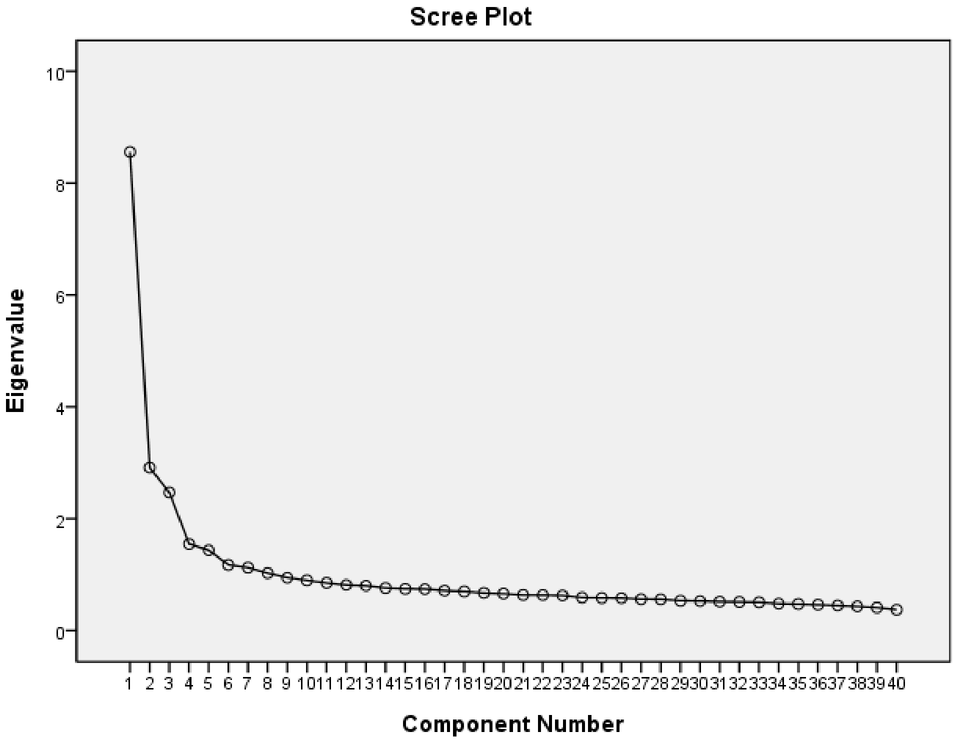


Figure 33: Scree plot for the PCA of the full SEAL Survey dataset

Before rotation was performed component 1 had 35 of the 40 items loading on it with weights of 0.300 or greater while components 2 and 3 each had 9 items loading on them. Components 4, 5 and 6 each had three items loading on them and, finally, components 7 and 8 had 2 items.

From the 40 by 40 reproduced correlations matrix there were 177 (22%) non-redundant residuals with values greater than 0.05 which is less than 50% and therefore suggests that the data is suitable for reduction by PCA/EFA (Field, 2005: 656).

Oblimin rotation of components, with Kaiser normalisation, was performed on the extracted components. The solution to the rotation fully converged. This yielded the pattern matrix of loadings for each of the eight extracted components (Table 22). The identities and loadings of the eight components were interpreted, comparing the extracted components to the scales derived through conducting the confirmatory factor analysis (CFA) of the complete dataset. Cronbach alpha values were calculated for each of the 8 components to provide an estimate of the scale reliability for each component, after reversing the ordinal response scale for items Q4, Q13 and Q16 as they loaded together with items with scales running in the other direction.

Four dimensions derived from the CFA had reasonable to good matches with components extracted from the PCA. The final CFA dimensions *attitudes to teachers and school*, *friendships* and *resilience* matched all of their items with one of the components derived from the PCA, though the first two of these dimensions from the CFA matched components in which additional items from the PCA were also loading. The *managing feelings* dimension from the CFA matched two out of three of its items with a component from the PCA. The remaining 3 dimensions from the CFA, namely *independence*, *managing behaviour* and *self-image*, did not have a good match with any of the PCA components. Five of the eight components from the PCA had Cronbach alpha scale reliabilities below the commonly accepted threshold of 0.7 including two of the four components not corresponding well to any of the CFA dimensions, one of which had a Cronbach alpha of 0.473.

The results of the PCA, coupled with the modifications made during the CFA process, suggest the nature and quality of the self-report data obtained from the SEAL Survey does not lend itself to clear cut factor models. Multiple models might represent themselves as possible solutions through the process of data reduction by factor analysis. The implications of this for the results and conclusions that can be drawn from this study will be taken up in more detail in the final chapter of this thesis (Chapter 7).

Item	1	2	3	4	5	6	7	8	Possible identity
Q34	0.391								1- Includes all four items from the final *attitudes to teachers and school* scale plus other items allocated to that scale in the original CFA model.
Q39	0.546								
Q19	0.562								
Q23	0.61								
Q37	0.708								
Q33	0.749								
Q35	0.811								
Q16		-0.669							2- Includes both final items from the *friendships* scale plus items allocated to the original Social Skills scale in CFA.
Q13		-0.641							
Q8		0.693							
Q36		0.694							
Q21			-0.346						3- Anti-social behaviour scale. This component does not match easily to any of the scales derived from the CFA, although it does share some items from the empathy and self-regulation scales in the original CFA model.
Q5			0.353			0.41			
Q2			0.359						
Q24			0.445						
Q14			0.608						
Q38			0.648						
Q31			0.649						
Q22			0.708						
Q17				0.391			[0.305]		4- Academic self-concept scale. Component also doesn't match well to any scale derived from the CFA. Shares some items from the *managing behaviour* scale in original CFA.
Q3				0.48	[-0.303]				
Q29				0.6					
Q30				0.624					
Q27	[0.301]			0.37					
Q9					-0.639				5- Potential emotional awareness scale. Also doesn't match easily to any of the scales derived from the CFA.
Q1					-0.509				
Q6					-0.506				
Q11		[0.315]				0.455			6- Match with the *resilience* item derived from the final CFA.
Q18						0.706			
Q7						0.744			
Q4							-0.632		7- Includes 2 of the 3 items from the *managing feelings* scale plus items allocated to original CFA Social Skills scale.
Q15							0.736		
Q12							0.759		
Q32								0.797	8- Perception of free time at school.
Q40								0.85	
Q25 & Q28		No loadings >0.300							Items not loading well on any component. Includes one item from final *self-image* scale Q28.
Q20							[0.324]		
Q26					[-0.308]				
Q10				0.315	-0.331				Low, cross-loaded item.
	0.810	0.665	0.750	0.698	0.473	0.614	0.677	0.705	Cronbach α scale reliability

Table 22: Adjusted pattern matrix for PCA of full SEAL Survey dataset

5.2 Results of Multilevel modelling of the SEAL Survey outcomes

5.2.1 Description of the dataset

The full data set contained data from 8836 students. These students were clustered within 390 different classes which in turn were clustered within 55 different schools.

Phase	School type	Schools	Classes		Students	
Primary	First	8	29	204	637	4944
	Primary/Junior	27	116		2803	
	Yrs 5&6 in 9-13 middle	10	59		1504	
Secondary	Yrs 7&8 in 9-13 middle		51	186	1430	3892
	11-16/18 secondary	9	130		2302	
	13-18 secondary	1	5		160	
Total		55	390		8836	

Table 23: The distribution schools, classes and students between phases

This clustered nature of the data, with students nested within classes nested within schools, makes it highly appropriate to analyse the data using multilevel linear regression techniques (also known as hierarchical linear modelling, HLM, or multilevel modelling, MLM). Analysis was conducted using MLwiN version 2.02 (Rasbash et al, 2005a). MLwiN requires the data to be sorted in a way that reflects the nesting or clustering, and so the data were arranged by ordering the responses first by school then by class within school. Within classes the data were sorted by last and then first name but this wasn’t essential for analysis using MLM.

A summary of the participating schools, by school type, phase of education and the school year groups surveyed is provided below in Table 24. Contextual data for each of the participating schools is also provided in the table below using percentile ranks data obtained from the Fischer Family Trust (FFT) database. The use of percentile ranks makes it possible to locate the position of the mean level of deprivation of students attending the school within the national dataset. The percentile ranks run so that high ranks indicate greater levels of deprivation. The RankFSM measure is based on the proportion of children attending the school who are entitled to fee school meals. The RankGDF measure is based on the mean Acorn ratings of the home neighbourhoods of children attending the school (Acorn is described in Chapter 3 above). The RankSE is based on the FFT socioeconomic factor, also described in Chapter 3 above, which combines both FSM and Acorn measures of deprivation.

School	date surveyed	phase	years surveyed	RankFSM	RankGDF	RankSE
1	Sep-08	secondary	8	33	43	38
2	May-08	middle	7	14	23	18
3	Dec-07	primary	3-6	54	86	70
4	Jul-08	primary	3-6	32	71	51
5	Feb-08	middle	5-8	40	43	41
6	Feb-08	primary	1-6	9	50	29
7	Feb-08	First	3-4	38	17	27
8	Nov-07	primary	3-6	21	24	22
9	Feb-08	primary	3-5	10	4	7
10	Mar-08	First	4	7	15	11
11	Sep-08	middle	6-8	30	17	23
12	Dec-07	first	1-4	48	61	54
13	Oct-07	middle	5-8	33	46	39
14	Jul-08	primary	1-6	36	12	24
15	Jun-08	middle	5-8	15	17	16
16	Jun-08	First	3-4	35	46	40
17	Nov-08	secondary	9	43	27	35
18	Oct-08	secondary	7-9	17	25	21
19	Feb-08	secondary	7-11	36	28	32
20	Oct-07	First	3-4	7	20	13
21	Dec-07	primary	3,4 & 6	11	6	8
22	Sep-08	first	3-4	10	18	14
23	Dec-07	primary	1,3,5 & 6	missing		
24	Nov-07	middle	5-6	17	26	21
25	Nov-08	primary	4-6	37	26	31
26	Oct-08	primary	3-4	10	23	16
27	Dec-07	primary	3-6	17	5	11
28	Feb-08	junior	3-6	21	53	37
29	Dec-08	first	3-4	42	33	37
30	Jun-08	primary	2-5	15	10	12
31	Jun-08	primary	3-6	11	22	16
32	Sep-07	primary	5-6	9	27	18
33	Apr-08	middle	5-8	40	34	37
34	Dec-08	secondary	9	40	34	37
35	Jul-08	secondary	7-8	39	32	35
36	Oct-08	primary	3-6	20	20	20
37	Dec-07	primary	3-6	77	87	82
38	Jul-07	primary	4	38	54	46
39	Oct-07	primary	3 & 5	29	26	27
40	Dec-07	primary	3-6	11	14	12
41	Dec-08	first	1-4	24	19	21
42	Nov-08	primary	6	14	35	24
43	Jun-08	primary	3-6	12	34	23
44	Feb-08	primary	1-5	10	24	17
45	Dec-07	primary	4-6	52	70	61
46	Dec-07	middle	5-6	30	19	24

47	Feb-08	primary	3-6	9	40	24
48	Dec-07	primary	1-6	7	20	13
49	Dec-08	middle	6-8	19	51	35
50	Mar-08	primary	3-6	10	21	15
51	Dec-07	middle	5-8	34	33	33
52	Dec-07	secondary	7,9&10	74	66	70
53	Dec-07	secondary	7	53	57	55
54	May-08	secondary	7	25	25	25
55	Dec-07	secondary	7-10	73	36	54

Table 24: Background information on the participating schools

To indicate how representative the range of school level deprivation is when comparing the participating schools with all other schools in the LA, a plot of RankSE factors shown (Figure 34).

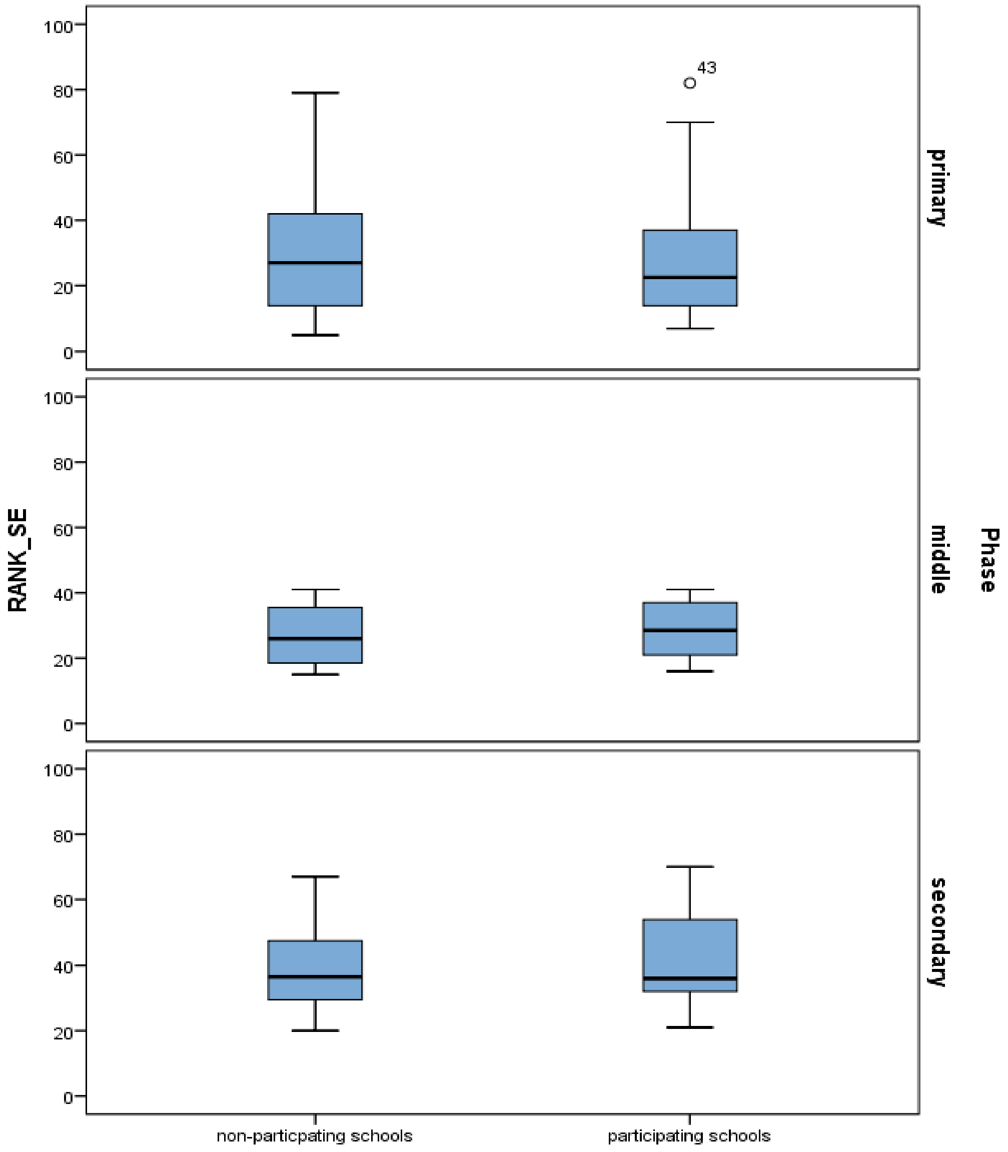


Figure 34: Comparison of deprivation context of participating and non-participating schools from the LA.

The distribution of RankSE percentile ranks for the participating schools is very similar to those of non-participating schools in the LA, although the participating secondary schools exhibit a distribution more skewed toward higher level of deprivation than is the case for the non-participating secondary schools.

5.2.2 Dependent and independent variables in the full dataset

The dependent variables for the MLM of the data were the 7 dimensions of the SEAL Survey which resulted from the confirmatory factor analysis (CFA) of the "About Me and My School" questionnaire (Hallam et al, 2003) which is described in the previous chapter. The student level independent variables for the MLM are drawn from other responses that students made to items on the survey including demographic information about the student (year group, and gender) and student self-rating responses to questionnaire items eliciting association with bullying and similar antisocial behaviour, both as a victim of such behaviours and also as the perpetrator. A summary descriptive analysis of the distribution of responses to these two bullying related items on the survey can be found below.

These associations with bullying behaviours as the victim and the bully were also aggregated to produce school level measures of antisocial behaviours by calculating the mean and standard deviation for both the victim and bully variables.

A number of school level independent variables were drawn from the Fischer Family Trust (FFT) database containing demographic and academic outcomes data for the years summer 2000 - summer 2006 (the period leading directly up to the administration of the SEAL Survey). Data was extracted from the FFT database for the participating schools and included measures of:

- the proportion of students taking for free school meals (FSM)
- a socio-economic status indicator based on the FFT SE model to determine "similar schools" combining FSM entitlement, a geodemographic indicator, the proportion of girls to boys and, where available, the mean and spread of the prior attainment of students on entry to the school. The SE model is widely used by schools as applied to estimates of student academic potential that are commonly used to inform student and school level target setting
- year on year school academic outcome measures leading up to the survey period
- the percentage of students attaining the expected outcomes for their age for the cohorts taking national examinations/tests in years 2002-2006 (for first schools the measure chosen was the percentage of students attaining in the national expectation

(level 2B) in writing, for primary/middle schools the outcome was the percentage of children averaging level 4 across the core curriculum subjects of English, maths and science, and for secondary schools the measure was the percentage of students attaining 5+A*-C in any subjects GCSE. These measures are all high stakes raw attainment indicators of school performance that were used in school "league tables" during that time and continued to be used in this way through the survey period)

- the school level contextual valued added measure based on the FFT SX model which, as discussed in Chapter 3, is a close match to the government/Ofsted CVA measure also used, during that time and on through the survey period, to rank schools in "league tables" and importantly as a key indicator of school effectiveness that was used by Ofsted to form a "pre-inspection hypothesis" judgments of the quality of academic outcomes in a school
- a measure of the trend in raw academic attainment and value added outcomes during the 5 years leading up to the survey period in the form of a raw attainment and contextual value added trend measure for the outcomes from the 2002-2006 period. These variables utilise the 5 level categorisation of school trajectories derived from my discussion of the longitudinal study of school outcomes conducted by Thomas et al (2007) contained in Chapter 3.

The school level measures drawn from the FFT database were all in the form of percentile ranks (1=high 100=low) in order to facilitate meaningful comparison between schools with different outcome measures. Dependent variables (all expressed as a percentage of the maximum possible outcome on the scale derived from responses to the SEAL Survey).

Code	Description	Range	Missing
SI	Self-image scale	0-100	3.7%
MF	Managing feelings scale	0-100	3.7%
MB	Managing behaviour scale	0-100	3.8%
Ind	Independence scale	0-100	3.7%
Resil	Resilience scale	0-100	3.4%
Frnd	Friendships scale	0-100	3.4%
Att	Attitudes to Teachers and School scale	0-100	4.6%

Table 25: Dependent variables in the SEAL Survey MLMs

The higher proportion of missing data for the dimension *Att* was most likely due to attrition as the items for this dimension being among the last in the survey.

A number of independent variables were included in the multilevel models based on insights gleaned from the literature review of school effectiveness models for both cognitive and non-cognitive outcomes. These are summarised in Table 26.

Code	**Description**	**Range**	**Missing (L1 only)**	**MLM level**
Year	Recode of the responses to the item "What year are you in" centred so that school year 6 = 0	-5 to 5	<0.1%	student
gender	Response to the item "Are you a boy or a girl"	0 = boy 1 = girl	<0.1%	student
Victim	Response to the item "I am sometimes bullied or picked on by other children"	0-4	2.0%	student
Bully	Response to the item "I sometimes bully or pick on other children"	0-4	1.9%	student
schvictim	School level mean of the "victim" variable	1.18-2.52		school
schbully	School level mean of the "bully" variable	0.38-1.38		school
sdschvictim	Standard deviation of the "victim" variable at the school level	1.12-1.58		school
sdschbully	Standard deviation of the "bully" variable at the school level	0,74-1.41		school
RankFSM	Percentile rank based on the proportion of students taking up free school meals	1-100		School
RankSE	Percentile rank based on the FFT SE model "similar schools" factor (see Chapter 3)	1-100		school
02RawRank	Percentile rank for the appropriate raw attainment threshold measure for the exam/test cohort for that year (2002-2006)	1-100		school
03RawRank				
04RawRank				
05RawRank				
06RawRank				
RawTrend	The longitudinal trend in raw outcome measures based on percentile ranks for the period 2002-2006	1-5		school
02SXRank	Percentile rank for the FFT SX measure for the school based on a contextual value added analysis of outcomes for the exam/test cohort for that year (2002-2006)	1-100		school
03SXRank				school
04SXRank				school
05SXRank				school
06SXRank				school
SXTrend	The longitudinal trend in contextual value added (FFT SX) measures based on percentile ranks for the period 2002-2006	1-5		school

Table 26: Independent variables in the SEAL Survey MLMs

5.2.3 Descriptive statistics for each Survey dimension

Overall means with 95% confidence intervals for each SEAL Survey dimension are presented below in chart form. They are shown against age (Year group) to indicate the association with the age of the student.

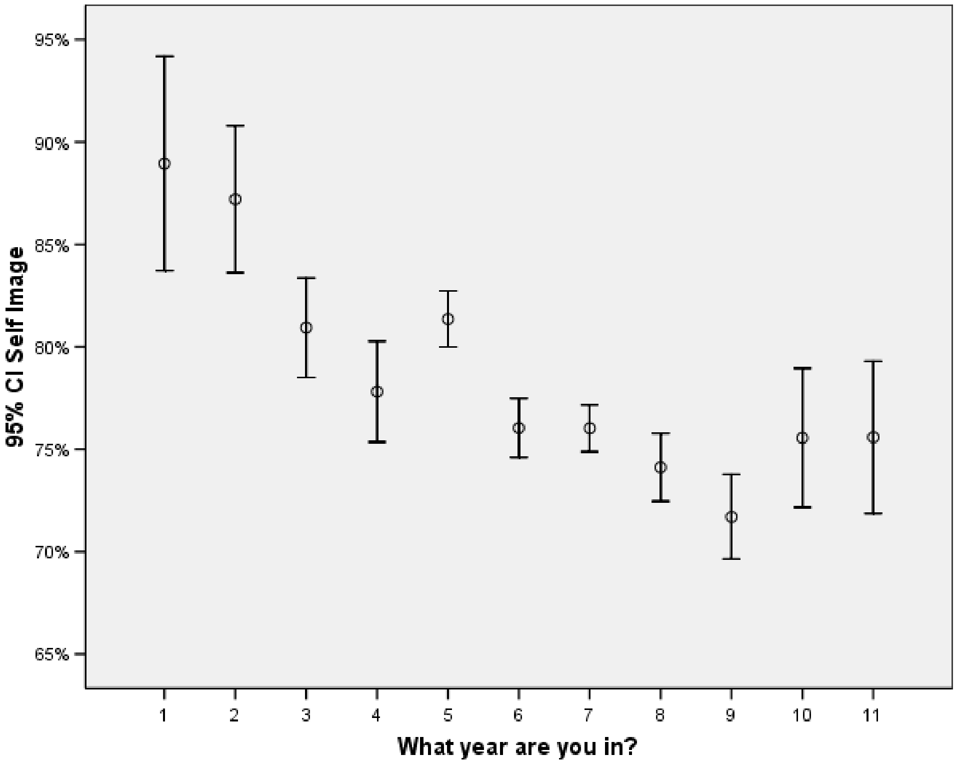


Figure 35: Means and 95% confidence intervals for the self-image dimension

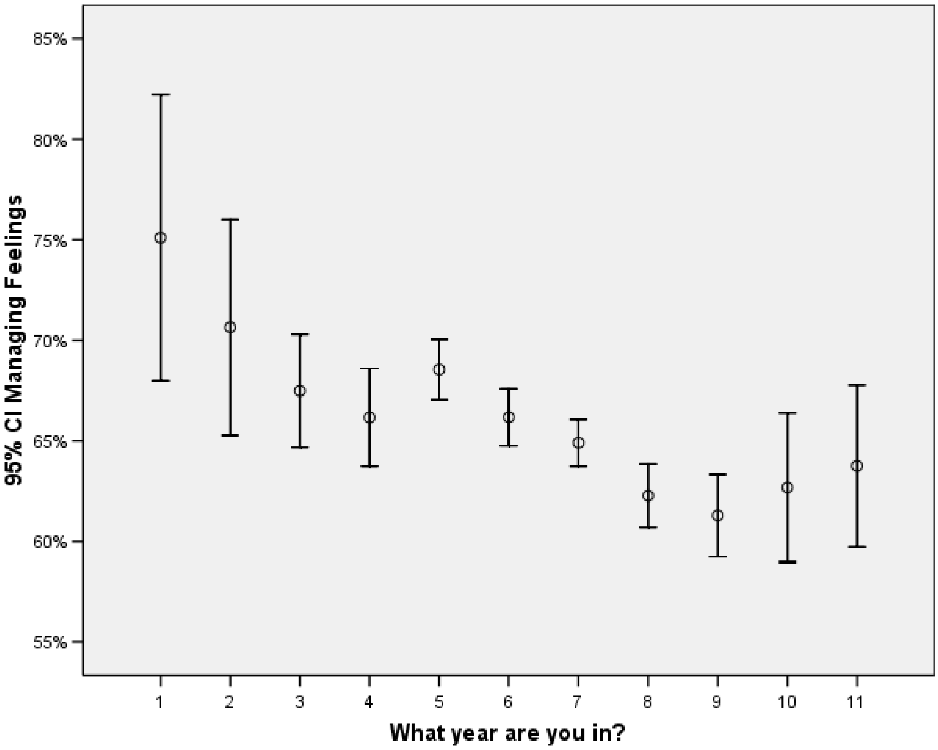


Figure 36: Means and 95% confidence intervals for the managing feelings dimension

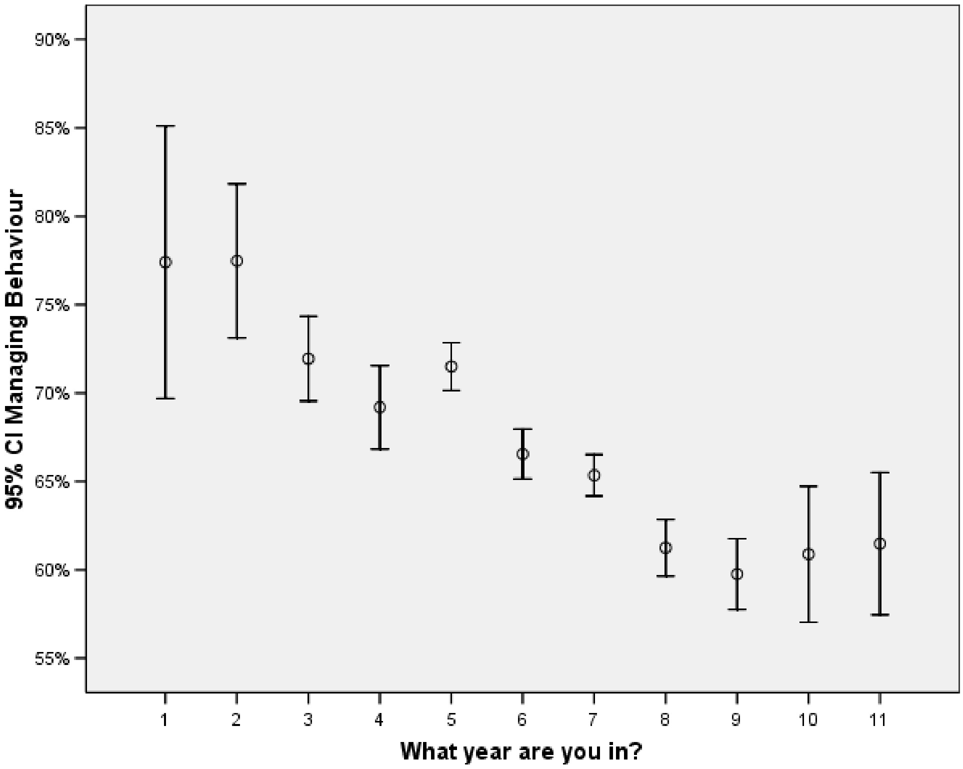


Figure 37: Means and 95% confidence intervals for the managing behaviour dimension

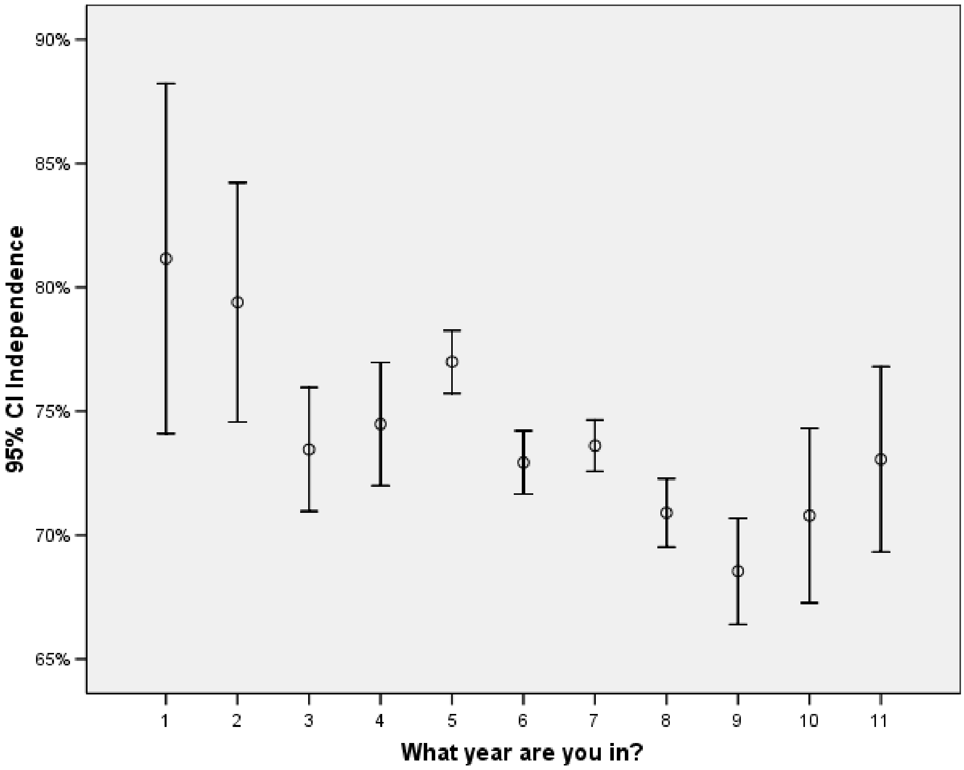


Figure 38: Means and 95% confidence intervals for the independence dimension

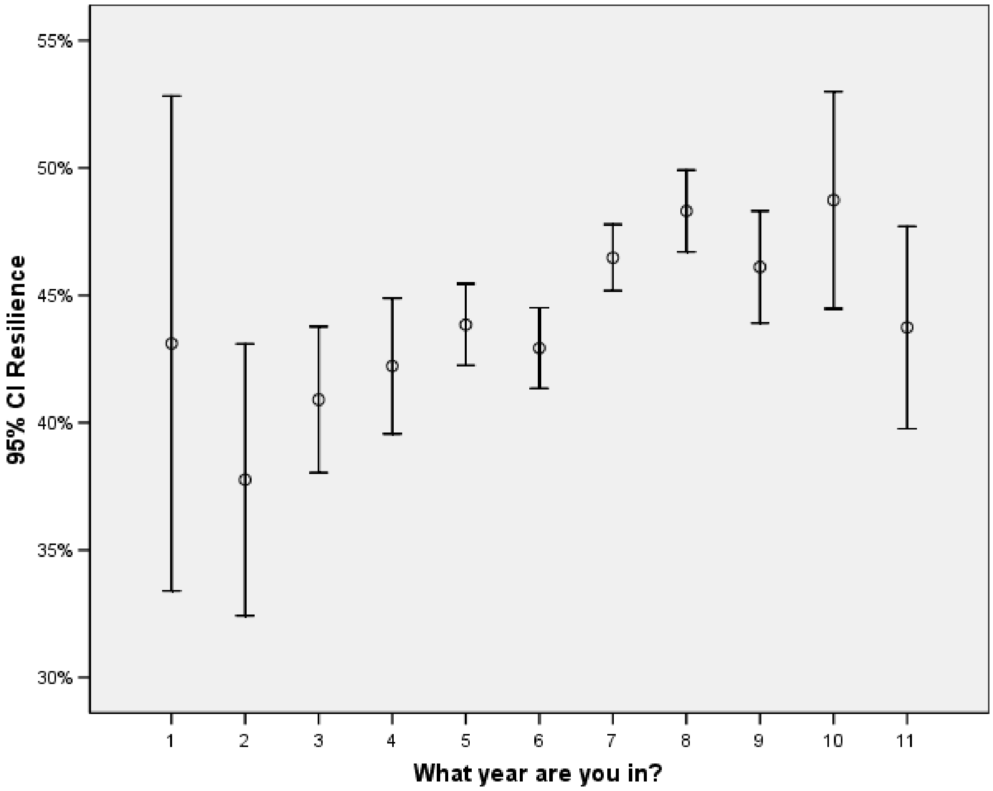


Figure 39: Means and 95% confidence intervals for the resilience dimension

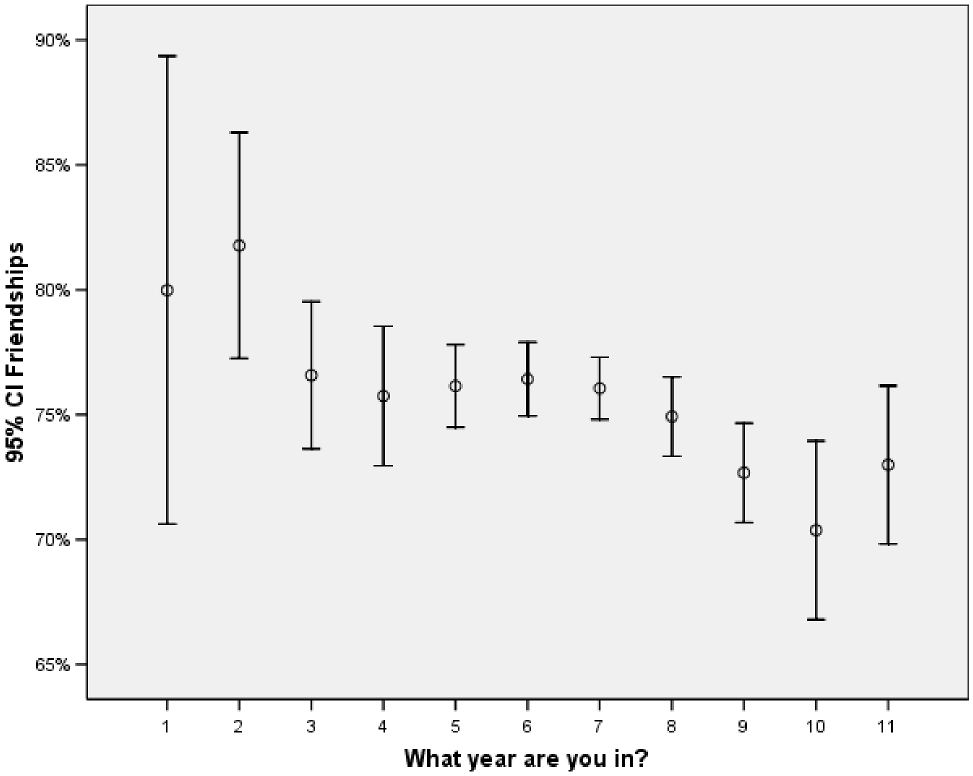


Figure 40: Means and 95% confidence intervals for the friendships dimension

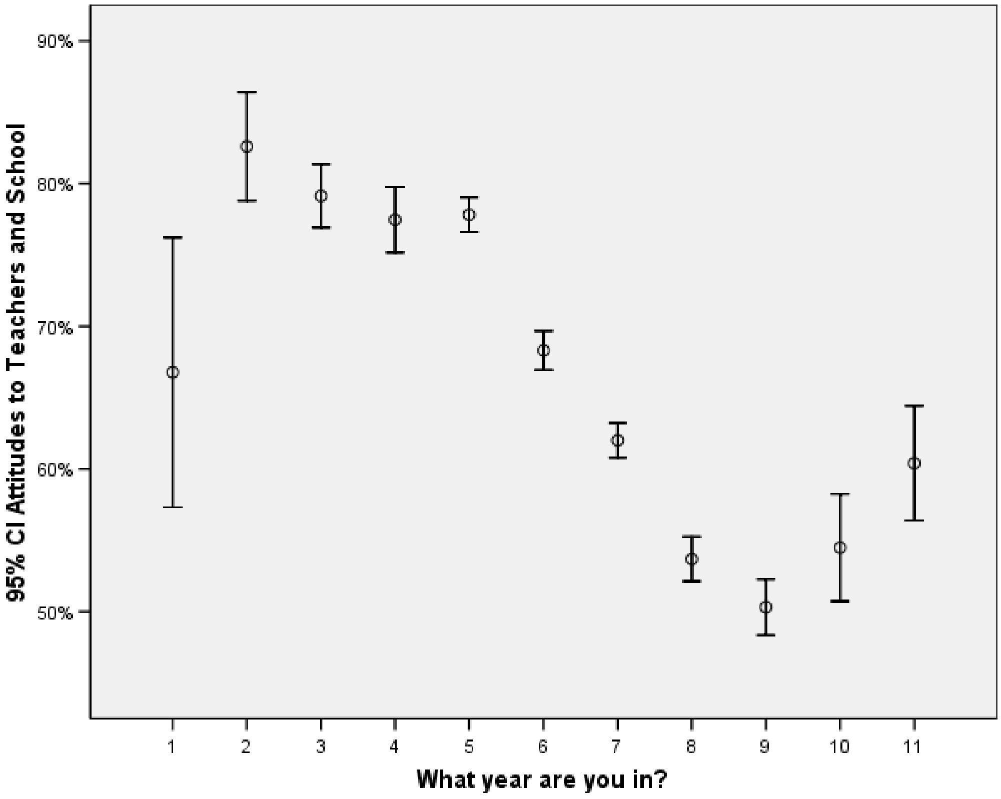


Figure 41: Means and 95% confidence intervals for the attitudes to teachers and school dimension

5.2.4 Descriptive statistics for the victim and bully independent variables

Cross tabulations of the *victim* and *bully* variables are provided to indicate the distribution of responses to the two items by gender and by age (school year).

victim	strongly agree	agree	neither	disagree	strongly disagree	Total
male	10.80%	18.70%	18.60%	22.30%	29.60%	4422
female	8.80%	19.30%	21.10%	24.60%	26.30%	4236
total	9.80%	19.00%	19.80%	23.40%	28.00%	8658

Table 27: crosstabs of the responses to the *victim* item (I am sometimes bullied or picked on by other children) by gender

bully	strongly agree	agree	neither	disagree	strongly disagree	Total
male	2.60%	7.30%	15.90%	25.20%	49.00%	4421
female	1.90%	3.60%	9.00%	24.70%	60.90%	4237
total	2.30%	5.50%	12.50%	24.90%	54.80%	8658

Table 28: crosstabs of the responses to the *bully* item (I sometimes bully or pick on other children) by gender

victim	strongly agree	agree	neither	disagree	strongly disagree	Total
Year 1	14.00%	27.20%	11.80%	29.40%	17.60%	136
Year 2	17.90%	26.20%	15.40%	17.90%	22.50%	240
Year 3	14.30%	25.10%	20.10%	15.90%	24.50%	804
Year 4	15.40%	20.90%	20.40%	15.20%	28.10%	755
Year 5	12.80%	22.10%	20.70%	18.70%	25.70%	1302
Year 6	9.00%	19.30%	21.20%	23.00%	27.40%	1536
Year 7	6.20%	17.10%	18.20%	27.20%	31.20%	1604
Year 8	7.70%	15.80%	19.30%	27.30%	29.90%	1257
Year 9	5.30%	12.10%	21.00%	32.90%	28.70%	849
Year 10	6.70%	15.40%	21.20%	23.10%	33.70%	104
Year 11	1.40%	11.40%	21.40%	32.90%	32.90%	70
total	9.80%	19.00%	19.80%	23.40%	28.00%	8658

Table 29: crosstabs of the responses to the *victim* item (I am sometimes bullied or picked on by other children) by age (year group)

bully	strongly agree	agree	neither	disagree	strongly disagree	Total
Year 1	7.40%	5.10%	5.10%	39.00%	43.40%	136
Year 2	8.30%	7.50%	5.40%	17.90%	60.80%	240
Year 3	4.50%	6.90%	10.80%	17.40%	60.40%	806
Year 4	3.60%	6.10%	12.70%	18.20%	59.40%	753
Year 5	2.70%	5.60%	12.80%	19.20%	59.70%	1303
Year 6	1.80%	5.50%	12.40%	25.80%	54.60%	1537
Year 7	0.70%	4.20%	10.70%	25.40%	59.00%	1603
Year 8	1.10%	5.90%	14.80%	30.30%	47.90%	1257
Year 9	0.60%	4.60%	17.10%	34.40%	43.30%	848
Year 10	5.70%	3.80%	11.40%	30.50%	48.60%	105
Year 11	4.30%	2.90%	12.90%	41.40%	38.60%	70
total	2.30%	5.50%	12.50%	24.90%	54.80%	8658

Table 30: crosstabs of the responses to the *bully* item (I sometimes bully or pick on other children) by age (year group)

Results of MLM analyses

SI – 2 level	null model																
	estimate	se		estimate	se		estimate	se		estimate	se		estimate	se		estimate	se
Fixed																	
intercept	78.558	0.487		77.723	0.269		79.683	0.346		85.102	0.457		74.456	4.247		76.268	3.452
Year				-1.401	0.123		-1.424	0.124		-1.609	0.129		-1.505	0.156		-1.408	0.129
Gender							-4.007	0.424		-4.511	0.422		-4.510	0.422		-4.484	0.422
Victim										-2.295	0.164		-2.307	0.165		-2.304	0.164
bully										-1.944	0.216		-1.985	0.218		-1.973	0.216
schvictim													(-0.833)	1.575		---	
schbully													(2.510)	2.557		---	
sdschvict													9.816	4.662		6.731	2.649
sdschbul													(-2.657)	3.822		---	
Random																	
L1 variance	386.487	5.944		385.070	5.918		380.941	5.855		365.893	5.654		365.789	5.652		366.795	5.654
L2 variance	8.751	2.395		(0.878)	0.628		(1.074)	0.679		(1.426)	0.762		(1.049)	0.662		---	
Model fit																	
deviance	74880.490			74793.870			74704.950			73589.130			73581.640			73587.160	
cases included	8506			8506			8506			8417			8417			8417	

Table 31: SEAL Survey dimension – Self Image (SI): 2-level models – students nested within schools (student level variables only)

SI - 3 level	**null model**															**optimum model**	
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	78.412	0.496		77.711	0.267		79.595	0.344		84.907	0.448		74.817	4.522		75.600	4.144
year				-1.369	0.128		-1.403	0.131		-1.602	0.131		-1.426	0.173		-1.391	0.152
gender							-3.883	0.427		-4.402	0.425		-4.406	0.425		-4.399	0.425
victim										-2.260	0.166		-2.280	0.166		-2.274	0.166
bully										-1.933	0.219		-1.962	0.220		-1.960	0.219
schvictim													(0.186)	1.715		---	
schbully													(1.062)	2.850		---	
sdschvict													8.262	4.960		7.161	3.176
sdschbul													(-1.757)	4.142		---	
Random																	
L1 variance	378.94	6.003		379.307	6.006		375.669	5.949		361.124	5.750		361.100	5.750		361.110	5.750
L2 variance	9.664	2.133		7.357	1.877		7.024	1.845		7.004	1.809		6.627	1.781		7.020	1.731
L3 variance	7.350	2.452		(0.142)	0.551		(0.290)	0.603		(0.333)	0.610		(0.390)	0.625		---	
Model fit																	
deviance	73410.18			73337.76			73255.41			72161.52			72155.840			72156.71	
cases included	8341			8341			8341			8253			8253			8253	

Table 32: SEAL Survey dimension – Self Image (SI): 3-level models – students nested within classes nested within schools (student level variables only)

The optimum models (2 level and 3 level models) were then carried forward for the stepwise addition of other school level factors for SES and academic outcome measures: RankFSM and RankSE – both produced non-significant point estimates (p>0.05); RawTrend and SXTrend – both produced non-significant point estimates (p>0.05); 02 to 06RawRank and 02 to 06SXRank – all produced non-significant point estimates (p>0.05)

MF - 2 level	**Null**													
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed														
intercept	66.490	0.515		65.784	0.494		64.666	0.535		71.948	0.565		83.161	5.882
year				-1.278	0.167		-1.257	0.167		-1.278	0.157		-1.411	0.174
gender							2.322	0.432		0.944	0.418		0.937	0.418
victim										-1.544	0.163		-1.516	0.163
bully										-5.225	0.215		-5.230	0.216
schvictim													(-3.173)	2.080
schbully													(0.647)	3.454
sdschvict													(-8.927)	6.264
sdschbul													(6.165)	5.083
Random														
L1 variance	397.662	6.116		395.190	6.079		393.861	6.058		358.257	5.539		358.113	5.537
L2 variance	10.098	2.691		8.621	2.373		8.532	2.352		6.510	1.894		5.554	1.690
Model fit														
deviance	75117.56			75059.730			75030.840			73442.420			73433.7	
cases included	8505			8505			8505			8416			8416	

Table 33: SEAL Survey dimension – Managing Feelings (MF): 2-level models – students nested within schools (student level variables only)

MF - 3 level	**null**													
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed														
intercept	66.442	0.511		65.870	0.472		64.698	0.514		71.884	0.548		82.405	5.951
year				-1.186	0.181		-1.161	0.181		-1.198	0.169		-1.340	0.196
gender							2.443	0.434		1.030	0.421		1.028	0.421
victim										-1.502	0.164		-1.479	0.165
bully										-5.212	0.217		-5.213	0.218
schvictim													(-2.562)	2.149
schbully													(-1.728)	3.548
sdschvict													(-9.529)	6.339
sdschbul													(7.439)	5.167
Random														
L1 variance	389.634	6.173		389.686	6.172		388.126	6.148		353.251	5.626		353.223	5.626
L2 variance	10.334	2.227		7.691	2.006		7.830	2.011		7.558	1.875		7.250	1.848
L3 variance	7.910	2.610		6.144	2.145		6.032	2.123		4.232	1.669		3.856	1.572
Model fit														
deviance	73639.740			73599.51			73567.95			72014.66			72007.210	
cases included	8340			8340			8340			8252			8252	

Table 34: SEAL Survey dimension – Managing Feelings (MF): 3-level models – students nested within classes nested within schools (student level variables only)

The optimum models (2 level and 3 level models) were then carried forward for the stepwise addition of other school level factors for SES and academic outcome measures: RankFSM and RankSE – both produced non-significant point estimates (p>0.05); RawTrend and SXTrend – both produced non-significant point estimates (p>0.05); 02 to 06RawRank and 02 to 06SXRank – all produced non-significant point estimates (p>0.05)

MB -2 level	**null**																
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**Se**		**estimate**	**se**
Fixed																	
intercept	68.246	0.540		67.014	0.433		64.236	0.468		69.421	0.532		74.201	5.665		69.679	0.491
year				-2.247	0.154		-2.196	0.152		-2.075	0.148		-2.230	0.164		-2.092	0.147
gender							5.761	0.407		4.062	0.392		4.053	0.392		4.092	0.392
victim										(0.201)	0.153		(0.219)	0.153		---	
bully										-6.062	0.201		-6.096	0.202		-6.008	0.196
schvictim													(-3.303)	1.997		---	
schbully													(5.270)	3.324		---	
sdschvict													(-3.351)	6.023		---	
sdschbul													(1.034)	4.886		---	
Random																	
L1 variance	366.322	5.636		358.664	5.517		350.442	5.391		314.881	4.869		314.644	4.866		315. 091	4.871
L2 variance	11.739	2.974		6.119	1.800		5.840	1.730		5.839	1.687		5.374	1.594		5.828	1.684
Model fit																	
deviance	74411.880			74208.810			74010.830			72339.770			72330.640			72405.480	
cases included	8503			8503			8503			8414			8414			8421	

Table 35: SEAL Survey dimension – Managing Behaviour (MB): 2-level models – students nested within schools (student level variables only)

MB - 3 level	**null**																
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	68.185	0.560		67.227	0.413		66.459	0.677		69.397	0.515		74.030	5.692		69.694	0.473
year				-2.077	0.166		-2.011	0.165		-1.926	0.160		-2.108	0.186		-1.945	0.159
gender							5.992	0.408		4.269	0.394		4.265	0.394		4.302	0.394
victim										(0.231)	0.154		(0.243)	0.154			
bully										-5.979	0.203		-6.011	0.204		-5.917	0.197
schvictim													(-2.784)	2.052			
schbully													(4.810)	3.385			
sdschvict													(-3.908)	6.056			
sdschbul													(1.385)	4.931			
Random																	
L1 variance	351.532	5.574		351.516	5.569		342.141	5.421		308.507	4.915		308.528	5.124		308.781	4.917
L2 variance	15.697	2.532		7.561	1.856		8.313	1.882		7.169	1.686		6.721	1.648		7.115	1.682
L3 variance	9.892	3.179		4.040	1.617		3.709	1.555		3.747	1.484		3.673	1.457		3.749	1.482
Model fit																	
deviance	72850.820			72723.240			72510.630			70889.550			70882.250			70956.030	
cases included	8338			8338			8338			8250			8250			8257	

Table 36: SEAL Survey dimension – Managing Behaviour (MB): 3-level models – students nested within classes nested within schools (student level variables only)

The optimum models (2 level and 3 level models) were then carried forward for the stepwise addition of other school level factors for SES and academic outcome measures. These are detailed in the next table.

MB	**2 level**			**3 level**	
	estimate	**se**		**Estimate**	**se**
Fixed					
intercept	68.197	0.869		68.403	0.853
year	-2.121	0.150		-1.981	0.163
gender	4.148	0.397		4.365	0.399
victim					
bully	-6.028	0.198		-5.935	0.200
sdschvict					
RankFSM	0.056	0.026		0.048	0.025
RankSE					
RawTrend					
Random					
Student variance	314.719	4.933		308.413	4.981
Class variance				7.033	1.697
School variance	5.376	1.615		3.527	1.455
Model fit					
Deviance	70415.630			68984.640	
cases included	8191			8029	

Table 37: SEAL Survey dimension – Managing Behaviour (MB): 2- & 3-level models with the addition of school level variables

Ind - 2 level	null																
	estimate	se		estimate	se		estimate	se		estimate	se		estimate	se		estimate	se
Fixed																	
intercept	74.810	0.334		74.466	0.272		74.283	0.330		77.623	0.403		74.753	3.718		77.015	0.236
year				-0.896	0.119		-0.894	0.119		-0.891	0.112		-0.836	0.139		-0.885	0.093
gender							(0.376)	0.384		(-0.527)	0.383		(-0.524)	0.383			
victim										(-0.254)	0.149		(-0.244)	0.150			
bully										-3.176	0.196		-3.154	0.198		-3.249	0.188
schvictim													(-1.118)	1.388			
schbully													(-2.205)	2.249			
sdschvict													(3.811)	4.100			
sdschbul													(1.320)	3.369			
Random																	
L1 variance	314.273	4.833		313.0636	4.812		313.034	4.812		301.898	4.665		301.908	4.664		303.085	4.670
L2 variance	3.055	1.092		(1.293)	0.671		(1.282)	0.668		(0.795)	0.530		(0.629)	0.480			
Model fit																	
deviance	73093.890			73041.830			73040.870			71965.010			71962.370			72041.13	
cases included	8506			8506			8506			8417			8417			8424	

Table 38: SEAL Survey dimension – Independence (Ind): 3-level models – students nested within schools (student level variables only)

Ind - 3 level	**null**																
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	74.792	0.335		74.510	0.278		74.275	0.335		77.571	0.411		75.388	3.989		76.972	0.260
year				-0.865	0.126		-0.861	0.126		-0.874	0.120		-0.836	0.151		-0.874	0.106
gender							(0.485)	0.387		(-0.427)	0.386		(-0.429)	0.386			
victim										(-0.259)	0.150		(-0.247)	0.151			
bully										-3.151	0.198		-3.133	0.200		-3.223	0.190
schvictim													(-1.170)	1.501			
schbully													(-2.305)	2.526			
sdschvict													(3.210)	4.364			
sdschbul													(1.562)	3.651			
Random																	
L1 variance	309.368	4.899		309.267	4.896		309.199	4.896		298.639	4.754		298.648	4.754		299.022	4.758
L2 variance	5.201	1.505		3.879	1.383		3.899	1.384		3.366	1.309		3.367	1.305		3.935	1.285
L3 variance	2.255	1.078		(0.937)	0.692		(0.925)	0.689		(0.595)	0.565		(0.443)	0.514			
Model fit																	
deviance	71648.470			71606.390			71604.820			70550.070			70547.890			70621.840	
cases included	8341			8341			8341			8253			8253			8260	

Table 39: SEAL Survey dimension – Independence (Ind): 3-level models – students nested within classes nested within schools (student level variables only)

The optimum models (2 level and 3 level models) were then carried forward for the stepwise addition of other school level factors for SES and academic outcome measures: RankFSM and RankSE – both produced non-significant point estimates (p>0.05); RawTrend and SXTrend – both produced non-significant point estimates (p>0.05); 02 to 06RawRank and 02 to 06SXRank – all produced non-significant point estimates (p>0.05)

Resil - 2 level	**null**																
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	44.008	0.574		44.553	0.508		48.539	0.569		56.914	0.609		71.532	6.382		73.936	5.990
year				0.924	0.179		0.843	0.179		0.411	0.170		(0.297)	0.188			
gender							-8.298	0.466		-8.422	0.452		-8.421	0.452		-8.298	0.477
victim										-4.909	0.176		-4.900	0.177		-4.987	0.171
bully										(-0.449)	0.232		(-0.444)	0.233			
schvictim													(0.553)	2.256			
schbully													(-0.018)	3.747			
sdschvict													-13.781	6.794		-13.234	(4.519)
sdschbul													(2.607)	5.513			
Random																	
L1 variance	478.671	7.351		478.100	7.342		460.721	7.075		418.381	6.469		418.339	6.469		418.639	6.471
L2 variance	12.705	3.351		8.570	2.496		9.457	2.653		7.534	2.199		6.576	2.001		6.976	2.083
Model fit																	
deviance	76940.420			76915.920			76604.750			74738.890			74733.480			74785.840	
cases included	8532			8532			8532			8415			8415			8420	

Table 40: SEAL Survey dimension – Resilience (Resil): 2-level models – students nested within schools (student level variables only

Resil - 3 level	**Null**																
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	44.052	0.555		44.582	0.483		48.524	0.544		56.869	0.593		70.930	3.783		73.504	5.812
year				0.941	0.190		0.855	0.190		0.436	0.180		(0.304)	0.206			
gender							-8.240	0.469		-8.398	0.455		-8.393	0.455		-8.261	0.450
victim										-4.874	0.177		-4.866	0.178		-4.950	0.172
bully										(-0.459)	0.235		(-0.452)	0.235			
schvictim													(0.595)	2.274			
schbully													(-0.446)	3.783			
sdschvict													-13.462	6.729		-12.948	4.388
sdschbul													(2.993)	5.494			
Random																	
L1 variance	471.978	7.463		471.809	7.460		454.966	7.194		413.068	6.578		413.034	6.577		413.300	6.580
L2 variance	7.648	2.291		7.312	2.254		6.809	2.156		6.427	1.995		6.389	1.989		6.421	1.993
L3 variance	10.069	3.109		6.6065	2.230		6.840	2.359		5.347	1.968		4.599	1.799		4.943	1.877
Model fit																	
deviance	75412.860			75390.830			75088.080			73267.300			73262.200			73314.100	
cases included	8365			8365			8365			8251			8251			8256	

Table 41: SEAL Survey dimension – Resilience (Resil): 3-level models – students nested within classes nested within schools (student level variables only)

The optimum models (2 level and 3 level models) were then carried forward for the stepwise addition of other school level factors for SES and academic outcome measures. These are detailed in the next table.

Resil	2 level			3 level			2 level		3 level			2 level		3 level	
	estimate	se		estimate	se		estimate	Se	estimate	se		estimate	se	estimate	se
Fixed															
intercept	73.338	5.583		73.263	5.406		75.817	5.718	75.725	5.539		83.289	6.506	83.281	6.197
year															
gender	-8.289	0.453		-8.238	0.457		-8.319	0.454	-8.268	0.457		-8.345	0.455	-8.295	0.459
victim	-4.991	0.173		-4.952	0.175		-4.998	0.174	-4.959	0.175		-5.015	0.174	-4.976	0.176
bully															
sdschvict	-14.307	4.234		-14.334	4.120		-16.265	4.444	-16.320	4.325		-17.209	4.520	-17.192	4.305
RankFSM	0.074	0.027		0.075	0.027										
RankSE							0.076	0.029	0.078	0.028					
RawTrend												-1.220	0.449	-1.257	0.421
Random															
Student variance	418.728	6.563		414.108	6.685		418.728	6.563	413.209	6.691		418.499	6.593	413.872	6.717
Class variance				5.870	1.975				5.864	1.975				5.965	1.985
School variance	5.531	1.799		3.751	1.614		5.531	1.799	3.975	1.670		5.465	1.796	3.386	1.532
Model fit															
Deviance	72737.510			71290.980			72737.510		70821.970			71968.520		70521.610	
cases included	8190			8028			8190		7977			8104		7942	

Table 42: SEAL Survey dimension – Resilience (Resil): 2- & 3-level models with the addition of school level variables

Frnd - 2 level	**null**																
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	76.151	0.326		76.038	0.323		76.172	0.391		84.732	0.471		80.660	4.530		84.481	0.412
year				-0.348	0.140		-0.349	0.140		-0.755	0.133		-0.792	0.163		-0.754	0.133
gender							(-0.275)	0.452		(-0.482)	0.436		(-0.490)	0.436			
victim										-4.751	0.170		-4.732	0.171		-4.758	0.170
bully										-1.011	0.224		-1.070	0.225		-0.975	0.221
schvictim													(-2.897)	1.671			
schbully													(2.281)	2.718			
sdschvict													(3.314)	4.954			
sdschbul													(2.700)	4.054			
Random																	
L1 variance	433.880	6.660		433.646	6.656		433.626	6.656		391.116	6.045		390.814	6.040		391.173	6.046
L2 variance	2.041	0.989		1.883	0.948		1.884	0.948		1.459	0.800		1.409	0.786		1.458	0.799
Model fit																	
deviance	76052.010			76045.900			76045.529			74131.890			74124.820			74133.1	
cases included	8532			8532			8532			8415			8415			8415	

Table 43: SEAL Survey dimension – Friendship (Frnd): 2-level models – students nested within schools (student level variables only)

Frnd – 3 level	**null**																
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	76.109	0.334		76.045	0.321		76.146	0.390		84.700	0.467		81.253	4.814		84.446	0.393
year				-0.363	0.149		-0.364	0.149		-0.786	0.138		-0.808	0.184		-0.794	0.131
gender							(-0.209)	0.456		(-0.447)	0.440		(-0.455)	0.440			
victim										-4.767	0.172		-4.753	0.172		-4.774	0.171
bully										-1.010	0.227		-1.061	0.228		-0.976	0.224
schvictim													(-2.642)	1.825			
schbully													(2.091)	3.014			
sdschvict													(2.313)	5.276			
sdschbul													(3.129)	4.392			
Random																	
L1 variance	429.480	6.792		429.366	6.790		429.371	6.790		386.439	6.155		386.368	6.154		386.447	6.156
L2 variance	8.704	2.182		8.650	2.170		8.622	2.168		8.882	2.044		8.467	2.015		9.267	1.987
L3 variance	(1.131)	0.986		(0.822)	0.833		(0.825)	0.884		(0.327)	0.669		(0.461)	0.715			
Model fit																	
deviance	74602.900			74597.090			74596.890			72722.370			72716.690			72723.640	
cases included	8365			8365			8365			8251			8251			8251	

Table 44: SEAL Survey dimension – Friendship (Frnd): 3-level models – students nested within classes nested within schools (student level variables only)

The optimum models (2 level and 3 level models) were then carried forward for the stepwise addition of other school level factors for SES and academic outcome measures. These are detailed in the next table.

Frnd	**2 level**		**3 level**			**2 level**		**3 level**	
	estimate	**se**	**estimate**	**se**		**estimate**	**se**	**estimate**	**se**
Fixed									
intercept	83.217	0.679	83.307	0.703		83.297	0.701	83.483	0.732
year	-0.824	0.134	-0.852	0.142		-0.828	0.131	-0.856	0.138
gender									
victim	-4.826	0.172	-4.844	0.174		-4.883	0.173	-4.898	0.175
bully	-1.008	0.224	-1.010	0.228		-1.024	0.226	-1.025	0.229
RankFSM	0.051	0.019	0.048	0.020					
RankSE						0.045	0.019	0.039	0.020
Random									
Student variance	391.050	6.128	386.546	6.244		390.967	6.144	386.575	6.262
Class variance			8.471	2.045				8.333	2.030
School variance	(1.225)	0.749	(0.374)	0.690		(1.003)	0.687	(0.224)	0.625
Model fit									
deviance	72101.620		70710.100			71665.510		70275.030	
cases included	8185		8023			8136		7974	

Table 45: SEAL Survey dimension – Friendship (Frnd): 2- & 3-level models with the addition of school level variables

Att - 2 level	**null**																
	estimate	**se**		**Estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	71.026	1.095		68.556	0.661		66.679	0.703		70.165	0.779		51.991	8.468		56.465	4.140
year				-3.586	0.178		-3.543	0.178		-3.440	0.177		-3.345	0.184		-3.388	0.174
gender							3.952	0.415		2.900	0.411		2.901	0.411		2.890	0.410
victim										(0.023)	0.160		(0.021)	0.160			
bully										-3.793	0.211		-3.820	0.212		-3.846	0.206
schvictim													(-0.680)	2.895			
schbully													(-3.233)	4.893			
sdschvict													(4.364)	8.855			
sdschbul													15.725	7.198		13.495	4.019
Random																	
L1 variance	377.061	5.829		363.761	5.622		359.765	5.561		345.295	5.340		345.255	5.339		345.450	5.341
L2 variance	60.395	12.442		18.769	4.371		19.772	4.560		21.684	4.900		17.126	4.017		17.699	4.103
Model fit																	
deviance	74047.190			73690.730			73600.470			73182.400			73170.870			73220.500	
cases included	8425			8425			8425			8416			8416			8421	

Table 46: SEAL Survey dimension – Attitudes to Teachers and School (Att): 2-level models – students nested within schools (student level variables only)

Att - 3 level	**null**																
	estimate	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**		**estimate**	**se**
Fixed																	
intercept	70.680	1.138		68.448	0.636		66.459	0.677		69.766	0.748		50.080	8.686		56.557	4.098
year				-3.356	0.235		-3.301	0.238		-3.203	0.238		-2.929	0.264		-3.097	0.236
gender							4.211	0.407		3.182	0.404		3.180	0.404		3.176	0.404
victim										(0.045)	0.159		(0.038)	0.159			
bully										-3.612	0.210		-3.633	0.210		-3.659	0.204
schvictim													(0.944)	3.127			
schbully													(-5.987)	4.949			
sdschvict													(4.287)	9.176			
sdschbul													16.873	7.336		13.124	4.098
Random																	
L1 variance	337.173	5.381		338.510	5.398		333.787	5.323		321.257	5.126		321.093	5.124		321.436	5.127
L2 variance	44.210	4.805		29.108	3.581		29.676	3.610		28.270	3.457		28.253	3.447		27.895	3.421
L3 variance	55.751	13.393		11.652	3.937		12.395	4.105		13.793	4.331		10.567	3.646		11.644	3.866
Model fit																	
deviance	72128.960			72002.740			71896.570			71504.810			71492.480			71544.200	
cases included	8261			8261			8261			8252			8252			8257	

Table 47: SEAL Survey dimension – Attitudes to Teachers and School (Att): 3 level models – students nested within classes nested within schools (student level variables only)

The optimum models (2 level and 3 level models) were then carried forward for the stepwise addition of other school level factors for SES and academic outcome measures: RankFSM and RankSE – both produced non-significant point estimates (p>0.05); RawTrend and SXTrend – both produced non-significant point estimates (p>0.05); 02 to 06RawRank and 02 to 06SXRank – all produced non-significant point estimates (p>0.05).

Chapter 6: Discussion of results

Following on from the presentation of the CFA and MLM data this chapter discusses the insights and inferences that might be drawn from these results. The discussion has been structured by the research questions posed in Chapter 4 for both universal and Family SEAL in order to establish the extent to which the research aims have been achieved.

For universal SEAL, the results show that a number of student and school level factors are associated with the non-cognitive outcomes from the SEAL Survey. School and Class level effects were observed to be modest in magnitude for most non-cognitive outcomes when compared to those found to be typical in similar analyses of cognitive outcomes as discussed in Chapter 3. These findings are broadly in line with those in previous studies of non-cognitive outcomes. For Family SEAL there is tentative evidence that a post intervention gain can be observed, especially for some participating children, and that the most substantial gains are likely to be observed in terms of increased social and emotional competence in the school environment.

6.1 Insights from the survey analysis for Universal SEAL

6.1.1 Factors that explain the variance in SEAL survey scores

USRQ1: What is the nature and contribution of student and school level factors that explain the variation in SEAL related non-cognitive outcomes?

The best fitting 2- and 3-level models for the dimension *self-image* indicate significant associations between students' *self-image* rating and their age (based on their year group) and gender. The gender effect is particularly strong with girls on average rating their *self-image* around 4.5% lower than boys which, is around 3 times the mean effect of progressing to the next year of schooling. Students reporting a prior association with bullying and anti-social behaviour (either as the 'victim' or/and the 'bully') tend to report lower *self-image* on average as their reported level of association with bullying behaviour increases. Those reporting the highest levels of association with bullying behaviour report self-image scores on average around 10-11% lower than those who indicate they have no association with bullying behaviours. The mean effect for victims is greater than that indicated by bullies, but only marginally so. It is not the case, indicated in this dataset, that bullies report a higher *self-image*. These data, along with the studies by Konu et al (2002a; 2002b) suggest that the self-image/self-esteem of bullies compared to that of victims may be more complex than any simplistic notion that bullies tend to possess a higher than average view of their self. The significant positive contribution that the increased spread at the school level of students' reporting being a *victim* of bullying behaviour is hard to interpret. This will be discussed in more detail in the section below.

For the *managing feelings* dimension of the survey the impact of progressing a year through the school system is similarly negative in terms of mean self-reported score with a reduction in managing feelings score of around 1.2 to 1.4% per year. Coupled with the result for self-image this suggests that students become more self-critical as they get older, possibly due to experience. The descriptive analysis of means with year in Figures 33 and 34 suggest a generally consistent fall in score with age, although a cautionary note needs to be sounded as these data are cross-sectional rather than longitudinal. As De Fraine et al (2005) point out, cross-sectional analyses of multiple cohorts like this are not the best source for a estimating the impact of age on non-cognitive outcomes. The impact of gender is different from that observed for *self-image*, with girls reporting, on average around 1% higher *managing feelings* scores than boys. Unlike *self-image*, the effect of gender on *managing feelings* scores appeared to be approximately twice as strong before the addition of the victim and bully factors to the models, suggesting that there is a gender imbalance in the level of reporting associations with bullying and anti-social behaviour.

Once again, students reporting a prior association with bullying and anti-social behaviour report lower *managing feelings* scores on average as their reported level of association as victims of bullying behaviour increases. Those reporting the highest levels on the victim scale report *managing feelings* scores on average around 7.5% lower than their peers at the other end of the scale. For bullies the effect is much more pronounced, with those reporting the highest levels of association with bullying reporting *managing feelings* scores over 25% lower than those who indicate they have no association with bullying behaviours.

The dimension *managing behaviour* also exhibits a significant effect of progressing through each year of the system, with an average fall of just over 2% in the *managing behaviour* score per year. Gender exhibits a particularly strong effect with girls reporting on average scores of around 4.2-4.3% higher than boys. Once again, the effect of gender was stronger before the addition of the factors *victim* and *bully* to the MLM. There is no significant effect on managing behaviour scores observed for victims but a strong negative association for bullies, with those reporting the highest levels of association with bullying behaviour having, on average managing behaviour scores 30% lower than those reporting no association with bullying. There are clearly issues of causality here but a link between general behaviour in the class and association with bullying or being antisocial to others is to be expected. Coupled with the associations between bullying and managing feelings described above this may indicate that of the SEAL programme, such as the Primary SEAL focus on bullying, and the focus on developing empathy and identifying and managing feelings have an important role to play. The alignment with national anti-bullying week may mean that a concentrated focus on bullying needs to be supplement with a sustained effort through the year. Utilising group based analyses through MLMs within the SER framework (as described in Chapter 3, section 3.1) may help identify schools which, over time reduce the level of bullying behaviours through the use of such resources for managing feelings, especially with bullies, which could in turn lead to the identification of useful improvement initiatives. The results of studies such as this reported in Kyriakides et al (2013) may provide useful insights for tackling bullying at the school level.

For the dimension *independence* once again mean scores fall as students progress through the years of schooling, though the fall under 1% per year is smaller for this dimension that the others considered so far. The effect of gender and association with bullying activity as a victim are not significant but students self-reporting association with bullying activity as bullies report lower independence scores with those at the highest end of the bully scale reporting means scores 16% lower than those reporting no association as bullies.

For *resilience* the simple descriptive analysis of means in Figure 37 suggests a different pattern with age than that observed for the other six dimensions, with a rising trend observed for each older cross-sectional cohort. However, in the MLMs this effect appeared more modest and eventually became non-significant once the full model was built including the factor for the spread of reported association with bullying behaviours as a victim (*sdschvict*). There was a very strong gender effect observed, with girls reporting on average *resilience* scores that were over 8% lower than boys. As may be expected the *resilience* scores of those reporting themselves as victims of bullying and antisocial behaviour are lower. Those reporting at the highest end of the victim scale have mean *resilience* scores 25% lower than those at the other end of the scale. Once again, the direction of causality cannot be indicated here, though one might postulate whether low resilience stems from being a victim of bullying and antisocial behaviour. For those reporting association with bullying and anti-social behaviours as bullies the impact on resilience was non-significant. There is a suggestion that there is a bipolar nature about the *resilience* scale, since the overall means are much lower than those reported for the other six dimensions. Thus the optimum score for *resilience* is possibly not up at the

100% end of the scale. Indeed a score of 100% might indicate a somewhat tough outer shell, or thick-skinned approach to pressing on with school work and persevering in relationships. It is noteworthy then that bullies do not report higher *resilience* scores than their peers which one might have expected if the *resilience* scale had this bipolar character. This may be due to the mix of relationship and work based *resilience* items in the scale.

The addition of certain school level factors appear to have more impact on resilience scores than scores from the other survey dimensions. There were significant though modest effects from the school level SES factors, RankFSM and RankSE, with a move of 10 percentile ranks to a ranking indicating greater deprivation(based on these national ranking scales) resulting in a mean *resilience* score that is around 0.75% higher. Thus greater deprivation at the school level seems to be linked to slightly higher *resilience* scores. Interestingly, by contrast the *RawTrend* factor suggests that schools with a poor or downward trend in headline measures of school level raw attainment during the previous 5 years tend to have students reporting lower levels of *resilience*. This effect resulted in a difference in mean *resilience* score of over 6% between those with the strongest performance indicated by raw attainment measures and those with the poorest performance trends. From the review of models cognitive outcomes in Chapter 3 and over the years of SER the negative association between levels of deprivation and raw attainment is a clear and lasting finding of effectiveness research. It is worth of note then that the effect of school level deprivation on *resilience* scores appears to be positive, while the impact of a school poor level trend in attainment on *resilience* scores appears to be negative.

For the *friendships* dimension, once again a negative effect of progressing through the years of schooling is observed, though the impact is modest with a decrease of under 1% in the mean *friendships* score for each school year. As with the dimension *independence* , the impact of gender was observed to be non-significant. The effects of association with bullying and antisocial behaviour, both as a victim and a bully, are significant and negative. The negative effect for bullies is relatively modest, with those reporting the highest association on the bully scale reporting *friendships* scores on average 5% lower than those with no bullying association. The negative effect for victims is much stronger making a difference of between 23 and 24% to mean reported *friendships* scores for those at each end of the victim scale. This negative, albeit modest, impact for bullies is unexpected if one might assume that they would tend to strong self-reporting of friendships, such as in a gang type grouping. The very considerable negative effect of being a victim on *friendships* scores, while understandable, is particularly worrying as it comes in effect as a ‘double whammy’ with a victim is in need of friends. Once again, there is no indication of causality here, as it is not clear whether victims report being bullied or picked on because they are social isolates, or whether they become social isolates because they are picked on. This outcome suggests peer mediation and social skills development initiatives through SEAL may have an important role to play.

As with the *resilience* dimension, the addition of school level deprivation measures (*RankFSM* and *RankSE*) to the models showed significant positive effects of the level of deprivation on *friendship* scores. These effects were even more modest than the effects for *resilience*, with a move of 10 percentile ranks to a ranking indicating greater deprivation resulting in a mean *friendships* score that is around 0.45 to 0.5% higher

Finally, in the dimension *Attitudes to Teachers and School* the strongest effect of age in terms of progress through the years of schooling was observed with a decrease in mean *attitudes* score of over 3% per year observed. The effect for gender was also relatively strong with girls, as might be expected, reporting higher scores (over 3% higher on average) for the *attitudes* dimension. Perhaps less expected was the non-significant effect of being bullied or picked on as indicated on the victim scale. This suggests that the experience of being bullied or picked on by peers does not lead to a negative impact on *attitudes to teachers and school*. This is despite the fact that one item in the *attitudes* scale is "I like coming to school". There also seems to be no evidence of resentment towards teachers for not dealing with the situation effectively. On the other hand, for those reporting an association with bullying and antisocial behaviours as bullies there is a strong negative effect on *attitudes to teachers and school* with those at the highest end of the bully scale reporting mean scores 18-19% lower than those at the other end of the scale. This is understandable as bullies are likely to be disaffected with school and also have poor relationships with teachers.

NOTE: when *bully* is treated as a nominal variable rather than ordinal it doesn't behave in an ordinal way for bully=1 possibly inaccurate reporting as this is rare and it may be that noise in the form of measurement error has crept in especially as the item is toward the end of the survey. The other 4 categories do follow the ordinal pattern where a trend is observed.

6.1.2 Partitioning the variance in SEAL survey scores between the student/school and student/class/school levels

USRQ2: Do potential school- and class-level effects exist in non-cognitive outcomes related to SEAL?

USRQ3: How do the school- and class-level effects observed for these non-cognitive outcomes compare with school- (and class-) level effects observed for cognitive outcomes?

Null models were produced by placing the variance at each level of the model into the random part of the multilevel model (MLM).

The proportion of the school level variance for each dimension was as follows:

SEAL survey dimension	Proportion of variance at the school level /%
SI	2.21
MF	2.48
MB	3.11
Ind	0.96
Resil	2.59
Frnd	0.47
Att	13.81

All values are significant (p≤0.05)

Table 48: The proportion of the variance at the school level for each SEAL Survey dimension in two level null models (students nested within schools).

These values for the school effect are modest for all the SEAL survey dimensions with the exception of the Attitudes to Teachers and School (Att) scale. Though modest in size, these levels of variance at the school level are however, in accordance with the findings reported from previous research. Gray (2004) concluded that the school effect for non-cognitive outcomes is not as large as it is for cognitive outcomes. In studies involving measures related to students' attitudes to school Gray concluded that between 5-9% of the variance was at the school level before the addition of any prior attitude measures. The dimension a*ttitudes to teachers and school* may therefore have an appreciable school effect compared with past research.

Gadeyne et al (2006) concluded that the difference between the school effect for non-cognitive outcomes is likely to be appreciably smaller than that for cognitive outcomes and this is indeed the case for proportion of the variance at the school level observed for the other six dimensions here.

The proportion of the school level variance for each dimension was as follows:

SEAL survey dimension	Proportion of variance at the class level	Proportion of variance at the school level
SI	2.49	1.86
MF	2.58	1.94
MB	4.27	2.62
Ind	1.65	0.71
Resil	1.59	2.06
Frnd	1.99	(0.26)
Att	11.59	12.75

All values are significant p(≤0.05) unless in brackets

Table 49: The proportion of the variance at the class and the school level for each SEAL Survey dimension in three level null models (students nested within classes, nested within schools)

The values for the variance at the class level in the null models are also relatively modest but generally higher than the proportion of the variance at the school level, with the exception of two dimensions, namely *resilience* and *attitudes to Teachers and School*, which both show a higher school level variance. For the *attitudes to Teachers and school* dimension the school effect holds up considerably (dropping by less than 8%) even after the addition of the class as an intermediate level in the model. For the other dimensions the addition of the intermediate level to the model results in a drop in the proportion of the variance at the school level to around 75-85% of the value estimated in the simpler, 2-level models.

The reduction in deviance (-2*log likelihood) for the addition of the extra level is significant for each of the seven dimensions, and so it is clear that the 3 level models provide a better fit to the data.

In their 2-level analysis of students in classes Gadeyne et al (2006) found class-level effects for non-cognitive outcomes were in the order of those for cognitive outcomes. The class level variance in the null models for these SEAL related dimensions suggest that after the addition of further explanatory variables, the class effect will end up being more modest than those normally expected for cognitive outcomes in English schools. This may be that some of the class level variance is absorbed by the school level in the more sophisticated, 3-level models constructed with this dataset

Opdenakker and Van Damme (2000) also conducted 3 level models in their analysis of wellbeing measures and found that in null models the proportion of the variance at the class level was 4.0% and at the school level was 6.5%. This compared with proportions of variance of around 20-30% for null models of math and language achievement.

The review of previous 3-level studies conducted by Landeghem et al (2000) for null models of a range of non-cognitive outcomes mainly associated with academic self-concept and motivation ranged from 3.6-8.2% for the proportion of the variance at the class level and 0.2 to 4% of the variance at the school level so the results of this study are in line with these studies, if on the lower side of those. It may be that the nature of the affective dimensions, focusing on self-image and feelings and also on social skills provides fewer opportunities for a school effect. That said, the nature of SEAL is such that it is exactly these kinds of affective outcomes that are at the heart of the programme and serious and sustained engagement with SEAL as a universal provision for all students (even if not universal in the sense that it is for all members of the school community) may result in wider variance between schools in the longer term.

Age and gender

The proportion of the school level variance for each dimension was as follows:

SEAL survey dimension	Proportion of variance at the school level /%		
	null model	year	year + gender
SI	2.21	(0.23)	(0.28)
MF	2.48	2.13	2.12
MB	3.11	1.68	1.64
Ind	0.96	(0.41)	(0.41)
Resil	2.59	1.76	2.01
Frnd	0.47	0.43	0.43
Att	13.81	4.91	5.21

All values are significant (p≤0.05) unless indicated in brackets

Table 50: The proportion of the variance at the school level for each SEAL Survey dimension in two level models (students nested within schools).

As can be seen in the table, the addition of basic student demographic factors of self-reported age (based on the coarse measure of the student year group) and self-reported gender has a marked impact on the proportion of the variance at the school level for some of the SEAL survey dimensions. The variance at the school level drops only a small proportion (<one quarter) for *managing feelings* and *friendships* , but falls by one quarter to half for *managing behaviour* and *resilience*. For *attitudes to teachers and school* the fall in school level variance is between one half and three quarters, and finally, for *independence* and *self-image* the variance falls to become no longer significant (p>0.05). This is to be expected based on the descriptive statistics showing a strong age related trend in these same dimensions. This is also in keeping with previous research on non-cognitive outcomes which have demonstrated significant dependency on the age of the student (Brookover et al, 1979; De Fraine et al, 2005). The descriptive statistics show this age dependency is not linear but that around the age of transition from primary to secondary the mean scores for some dimensions take a marked decline, especially the dimension *attitudes to teachers and school*.

The addition of the factor *gender* to the model already containing age has various effects. Where the point estimate for gender was non-significant as expected, there is very little change in the proportion of variance at the school level.

Intuitively, one might expect the addition of any explanatory variable that reduces the overall variance to also reduce the proportion of variance at the higher levels. This is not the case with gender, which caused the proportion of variance at the school level to rise slightly for both *resilience* and *attitudes to teachers and school* (the effect on *self-image* is not included as the proportion of variance at the school level is non-significant). This suggests possibly that the gender effect for these dimensions is not consistent between the higher level units, in this case, schools. Alternatively, and perhaps more likely, is that the influence of age on these outcomes is different for boys than it is for girls and so the addition of gender reintroduces variance between schools as schools in the data set represent the full range of compulsory schooling in the locality, namely first schools, primary schools, middle schools and secondary schools. The overall result of adjusting the models for gender is either a similar proportion of variance at the school level or an increase in the school level variance.

The proportion of the school level variance for each dimension was as follows:

SEAL survey dimension	level in MLM	Proportion of variance at that level /%		
		null model	Year	year + gender
SI	class	2.49	1.90	1.84
	school	1.86	(0.04)	(0.08)
MF	class	2.58	1.94	1.98
	school	1.94	1.52	1.50
MB	class	4.27	2.11	2.37
	school	2.62	1.11	1.05
Ind	class	1.65	1.24	1.25
	school	0.71	(0.30)	(0.30)
Resil	class	1.59	1.53	1.47
	school	2.06	1.36	1.46
Frnd	class	1.99	1.97	1.97
	school	(0.26)	(0.19)	(0.19)
Att	class	11.59	7.92	8.16
	school	12.75	3.07	3.30

All values are significant (p≤0.05) unless in brackets

Table 51: The proportion of the variance at the class and the school level for each SEAL Survey dimension in three level models (students nested within classes, nested within schools)

From the table above, based on the outputs of 3-level models, it is possible to consider the impact of the addition of basic student demographic factors of age and gender on the proportion of the variance at both the class and at the school level for each of the SEAL survey dimensions.

Simply on the addition of *year* as an explanatory variable for the dimensions *self-image* and *independence* causes the proportion of the variance at the school level to become non-significant, as it did for the simpler, 2-level models discussed above. This suggests that there is no observable school effect in the data for these dimensions of the SEAL Survey. The impact on the class effect is a fall in the variance of approximately a quarter but, after adjusting for age a modest class effect may still remain for these dimensions. For the *managing feelings* dimension a similar reduction of around one quarter was observed in the proportion of the variance at both the class level and the school level with the adjustment for students' age. *Resilience* and *friendship* saw only very small reductions in the class level variance, each less than 5%, whereas the school level variance for *resilience* fell to two thirds of the value in the null model.

As with the 2- level models, the dimensions of *managing behaviour* and *attitudes to teachers and school* saw the biggest reduction in the proportion of variance estimated at the higher level, in this case for both the class and school levels. For attitudes to teachers and school the class level variance fell to less than two thirds of the null model value and for managing behaviour it fell to less that 50%. In terms of the school level variance *managing behaviour* fell to just over a third of its null model value while the school level variance for *attitudes to teachers and school* fell to a mere quarter of the null model value, though still remained the largest of the school level variances at just over 3%.

The addition of the factor *gender* to models that have already been adjusted for students' age once again caused some higher level variances to increase by a small amount. This phenomenon was observed for the dimensions *managing behaviour* and *attitudes to teachers and school* at the class level and for *attitudes to teachers and school* and for *resilience* at the school level. Overall the effects of gender in addition to the adjustment for the students' age were relatively modest and the MLM analysis at this stage suggested that small but significant class effects might remain for each of the seven SEAL survey dimensions and school effects for four of the seven.

<u>The addition of potential proxy measures of prior affect – association with bullying and antisocial behaviour, either as the victim or the bully</u>

The format of these models is the same as those displayed in the equations shown for 2- and 3-level models for age and gender above as these additional factors are also student level factors independent variables in the MLM. These proportions of explained variance are much smaller than those of prior attainment for cognitive outcomes but none are trivial and these factors make a considerable contribution (around 10%) for four of the seven dimensions.

SEAL survey dimension	Total variance null model	Total variance after *victim* and *bully* added	Proportion of variance explained
SI	393.642	382.288	2.88%
MF	405.645	368.513	9.15%
MB	378.533	337.320	10.89%
Ind	317.563	306.178	3.59%
Resil	489.175	442.508	9.54%
Frnd	435.837	394.900	9.39%
Att	444.496	425.759	4.22%

Table 52: The effect on the total variance of adding *victim* and *bully* factors to the MLMs

Two-level models

The proportion of the school level variance for each dimension was as follows:

SEAL survey dimension	Proportion of variance at the school level /%			
	null model	Year	year + gender	year + gender + victim + bully
SI	2.21	(0.23)	(0.28)	(0.39)
MF	2.48	2.13	2.12	1.78
MB	3.11	1.68	1.64	1.82
Ind	0.96	(0.41)	(0.41)	(0.26)
Resil	2.59	1.76	2.01	1.77
Frnd	0.47	0.43	0.43	(0.37)
Att	13.81	4.91	5.21	5.91

Table 53: Summary of the proportion of variance at the school level in two-level MLMs

There were five SEAL Survey dimensions with significant proportions of variance remaining ($p<0.05$) at the school level after the previous stages of model building adjusting for students' age (expressed as their school year) and gender. The reduction in overall variance together with the slight drop in the proportion of variance at the school level for the *friendship* dimension means that the school level variance is no longer significant ($p>0.05$). The effect of adding the victim and bully factors to the models for *managing feelings* and *resilience* causes the proportion of variance at the school level to reduce by about 11-14% of their value from the previous model, while the school level variance for *managing behaviour* and *attitudes to teachers and school* rose by similar proportions.

SEAL survey dimension	level in MLM	Proportion of variance at that level /%			
		null model	year	year + gender	year + gender + victim + bully
SI	Class	2.49	1.90	1.84	1.90
	School	1.86	(0.04)	(0.08)	(0.09)
MF	Class	2.58	1.94	1.98	2.09
	School	1.94	1.52	1.50	1.16
MB	Class	4.27	2.11	2.37	2.27
	School	2.62	1.11	1.05	1.17
Ind	Class	1.65	1.24	1.25	1.11
	school	0.71	(0.30)	(0.30)	(0.20)
Resil	class	1.59	1.53	1.47	1.53
	school	2.06	1.36	1.46	1.26
Frnd	class	1.99	1.97	1.97	2.25
	school	(0.26)	(0.19)	(0.19)	(0.08)

Att	class	11.59	7.92	8.16	8.09
	school	12.75	3.07	3.30	3.80

All values are significant (p≤0.05) unless in brackets

Table 54: Summary of the proportion of variance at the school level in three-level MLMs

For the 3-level models all seven dimensions still have significant proportions of variance at the class level. There are slight reductions in the proportion of variance at the class level for *managing behaviour*, *independence* and *attitudes to teachers and school*, while *self-image*, *managing feelings* and *resilience* showed slight increases in the proportion of variance at the class level (<10%). *Friendships* showed a more marked increase of around 12-13% from the proportion of variance at the class level estimated in the previous model, adjusted only for age and gender. The changes to the proportion of variance at the school level were similarly small compared to the previous round of model building. The direction of movement in terms of increase or decrease was in opposition to the changes in the class level variance.

The results of the model building so far suggest that the utilisation of the variables available from the data set may result in small but significant class effects for six of the seven SEAL Survey dimensions and possible significant school effects, albeit similarly modest for three dimensions (*managing feelings*, *managing behaviour* and *resilience*. The exception, at both class and school level is *attitudes to teachers and school*, which has more substantial class and school effects. These effects are smaller than those observed for cognitive outcomes, falling at the lower end of the range normally reported for value added models (Luyten, 2003; Sammons, 2007).

This more substantial effect for the *attitudes to teachers and school* factor is both unsurprising and reassuring since one would expect a higher magnitude effect for a factor more proximally focused to the affect associated with the relationship between students and the teachers and school.

Adding school level variables – step 1 addition of school level aggregates of the *victim* and *bully* variables

The 2-level models containing the four student level explanatory variables *year*, *gender*, *victim* and *bully* were carried forward to the next round of model building. The school level mean and spread (standard deviation) of both the *victim* and *bully* variables had been calculated and these we entered into the 2-level models.

For five of the seven SEAL survey dimensions none of the four school level factors associated with bullying and antisocial behaviour produced significant point estimates for the school level variables. The exceptions to this were *self-image* and *resilience* for which the standard deviation of the victim variable across the school (*sdschvict*) was significant. This would suggest a potential peer effect of the spread of students reporting that they have been victims of bullying and antisocial behaviour at school. The point estimate for the *sdschvict* variable in the model for *self-image* was positive suggesting that the bigger the proportion of students reporting that they were victims of bullying or being picked on by other students, the more likely a student is to report a higher *self-image*. This is counter to the negative effect on *self-image* of the student self-reporting being a victim of bullying behaviours. One might postulate that being surrounded by others in the school who report being victims of bullying leaves an individual feeling more positive about her/himself, but this interpretation would require further testing through repeat administration of the survey and follow-up research.

In contrast to the peer effect on *self-image*, the peer effect of the spread of students reporting being victims of bullying behaviours on *resilience* is negative which is also in line with the negative effects on the *resilience* rating of self-reporting as a victim. This suggests that both being a victim and being around more people who consider themselves victims of bullying behaviours results in students rating themselves lower on the *resilience* scale.

In the final round of model building, a number of other school level factors were added to the best fitting model obtained thus far (indicated as the “optimum model” in the tables found in the Results chapter). Two of these factors related to the overall socio-economic status (SES) of students in the school (the percentile rank of the school based on Free School Meal eligibility of its students compared with all schools nationally – *RankFSM*, and the percentile rank for the FFT SE factor which FFT used to determine what they refer to as similar schools, based on the school Acorn score, the free school meal eligibility, the proportion of girls in the school the mean and spread of the prior attainment of students on entry to the school – *RankSE*). The other school factors related to the cognitive outcomes of students in the school using measures associated with school accountability found in school “league tables”. These factors were percentile ranks for the headline attainment statistic most relevant for each school in the dataset and also a contextual value added score (FFT SX) for the school, and these data were for the five year period leading up to the administration of the survey (2002-2006). Finally two trend measures in cognitive outcomes at the school level over this time period were included, one for the raw attainment statistic (*RawTrend*) and one for the contextual value added scores (*SXTrend*).

Of all of these factors only the SES factors *RankFSM* and *RankSE* and the *RawTrend* factor were shown to be significant in the "optimum models" carried forward from the penultimate stage of model building. The factors were only added individually to models as when added as a set their individual effects were non-significant. It is likely that the *RankFSM* and *RankSE* measures are correlated and it is also well known as a result of the whole body of school effectiveness research for cognitive outcomes (see Sammons 2007 for a summary) that raw attainment measures are associated with student SES so perhaps unsurprising that *RawTrend* might be significant.

The SES and trend in raw attainment factors were only found to be significant for the dimensions *managing behaviour* (*RankFSM* only), *resilience* (*rankFSM*, *rankSE* and *RawTrend*). In each case the SES factors were positively associated with the scores for the dimension. As percentile ranks in this case act as measures of increasing deprivation, this suggests that an increase in the mean deprivation experienced by the student body in the school results in higher ratings for *managing behaviour* and *resilience*. The effects are modest, with a shift in 10 places in percentile rank giving rise to a change of around 0.5% to the mean managing behaviour score and about 0.75% to the mean resilience score. It is important to note, however, that changes in percentile rank across the continuum do not necessarily represent a linear change in demand on the school. It is easier to move ten places in rankings when in the middle of the set of schools but more challenging to shift when in in the upper or lower quartile and especially in the head or tail group of the top/bottom 10% of schools. It may be that schools with particularly high levels of deprivation place a high priority on behaviour management and developing behavioural skills.

The *RawTrend* factor was negatively associated with *resilience* which suggests that students in schools with consistently low levels of raw threshold measures for cognitive outcomes, or with declining trends in these headline outcomes over time, are likely to rate themselves lower on the *resilience* scale. When *RawTrend* was combined in a model with *RankFSM*, the point estimate for the coefficient of *RawTrend* became non-significant ($p>0.15$), suggesting that *RankFSM* only should be added to the optimum model for *resilience*.

In the same way the 3-level models containing the four student level explanatory variables *year*, *gender*, *victim* and *bully* were also carried forward to the next round of model building. The school level mean and spread (standard deviation) of both the *victim* and *bully* variables had been calculated and these we entered into the 3-level models.

The results of this round of model building for both the mean and spread of the bull and victim variables, and for the SES and school level outcome factors confirmed the findings above for the 2-level models, namely, that only *self-image* and *resilience* demonstrate a peer effects for the variables *victim* and *bully*, and that *managing behaviour* and *resilience* demonstrate peer effects of SES. In each case the direction of these peer effects is the same as that observed in the analysis of 2-level models.

The absence of peer effects for non-cognitive outcomes is likely to be a function of the lower class and school level effects for non-cognitive outcomes compared with cognitive outcomes. It is also in line with some recent research on peer effects of both cognitive and non-cognitive outcomes of pre-school education on later learning in the elementary phase, conducted by Neidell and Waldfogel (2010), which found significant cognitive and non-cognitive peer effects, based on prior outcome measures, for cognitive outcomes, whereas no significant peer effects on non-cognitive outcomes were observed.

Bearing in mind that the region is relatively homogenous in terms of deprivation and other factors this present of significant school effects, however modest, is non-trivial and could suggest that early adoption of SEAL by some schools is having an effect that these effects may increase the variation in non-cognitive outcomes between schools. It certainly isn’t evidence of the potential for ranking schools on the basis of such outcomes. Such rankings are highly questionable for cognitive outcomes (Leckie and Goldstein, 2011) though some studies have shown they may have contributed to the overall quality of schools (Burgess et al, 2005).

6.2 Family SEAL Results

6.2.1 Analysis of parent and teacher Emotional Literacy Checklist data

		'control' students (N=13)		'concern' students (N=14)		Difference in means
		Mean	sd	Mean	sd	
Self-awareness	Parent rating	.627	.0881	.618	.1049	non sig
	Teacher rating	.813	.1531	.665	.1265	p<0.050
	Total	.720	.1547	.642	.1165	
difference in parent teacher means		p<0.010		non sig		
Self-regulation	Parent rating	.650	.1568	.629	.1888	non sig
	Teacher rating	.788	.2904	.598	.2411	p<0.100
	Total	.719	.2393	.613	.2131	
difference in parent teacher means		non sig		non sig		
Motivation	Parent rating	.719	.1451	.575	.1156	p<0.010
	Teacher rating	.837	.1584	.612	.1301	p<0.001
	Total	.778	.1604	.593	.1222	
difference in parent teacher means		p<0.100		non sig		
Empathy	Parent rating	.758	.0886	.807	.1651	non sig
	Teacher rating	.803	.2710	.643	.2305	non sig
	Total	.780	.1989	.725	.2138	
difference in parent teacher means		non sig		p<0.050		
Social skills	Parent rating	.931	.0597	.854	.1434	p<0.100
	Teacher rating	.885	.1528	.741	.1508	p<0.050
	Total	.908	.1160	.797	.1554	
difference in parent teacher means		non sig		p<0.100		
Overall percentage	Parent rating	.737	.0767	.696	.0998	non sig
	Teacher rating	.825	.1893	.652	.1180	p<0.010
	Total	.781	.1485	.674	.1096	
difference in parent teacher means		non sig		non sig		

Table 55: Mean scores from the parent and teacher Emotional Literacy checklists for both 'concern' and 'control' students. Results of tests of significant difference in mean scores are also show. (See Appendix for raw output data)

It is clear from the data in Table 55 above that there is a greater difference between the teacher scores for 'control' and 'concern' students than there is between the parent scores for the two groups of students. The teachers will be able to make comparisons between children and there could also be an element of self-fulfilling prophecy at work here as the teacher had already identified the students causing concern. The data also raises the question as to whether parents and carers have a more rounded view of children seeing them in a variety of contexts such as engaging with adults and siblings.

Results of one-way ANOVA reveal some interesting differences in scores. As the sample sizes were relatively small it was decided to consider differences in mean percentage score at the 90% significance level or higher ($p\leq 0.1$) as worthy of report. When comparing the mean teacher scores for control and concern students (Table 4.3.1) there were significant differences in the scores for self-awareness ($p<0.05$), self-regulation ($p<0.10$), motivation ($p<0.001$), social skills ($p<0.05$) and the overall mean score ($p<0.01$). The only other remaining mean score, for empathy, only just failed to be significant at the 90% significance level ($p=0.10$). In every case the mean scores for the control students were always higher than for the concern students.

In the mean of the parent percentage scores there were significant differences between 'control' and 'concern' students only for the aspects of motivation ($p<0.01$) and social skills ($p<0.1$). In both cases the mean score for the control students was higher than the mean score for the concern students.

When comparing the mean percentage scores for the parent and teacher responses for the 'concern' students there were significant differences for empathy ($p<0.05$) and for social skills ($p<0.1$). In each case the mean percentage score from the parents was higher than the mean percentage score from the teachers. As already highlighted previously, these two aspects form Faupel's (2003) subdivision of *social competence* based on Goleman's analysis of emotional intelligence (1996). For teachers in the school it would therefore suggest that any differences in emotional competence are manifest in social interactions demonstrated via the teachers' assessment of the students' empathy and social skills. It may be that the parents are able to observe a wider range of social interactions with siblings, parents, wider family members and friends rather than the relatively narrower set of student-student and student-teacher social interactions that take place at school.

For the 'control' students there was a significant difference in the parent and teacher mean percentage scores for self-awareness ($p<0.01$) and for motivation ($p<0.10$). In each case the mean percentage score from the teachers was higher than the mean percentage score from the parents. These two aspects are part of Faupel's (*ibid*) personal competence subdivision. Again, the narrower focus of the teacher, on the students' skills and capabilities demonstrated in the school context might not take into account uncertainties expressed by the students in the wider context of the family and home. There may also be an element of compliance here where 'control' students selected by the teachers who are 'model' pupils exhibit less compliant attitudes at home.

Despite drawing data from one pilot school the fact that significant differences in the sub-scale scores occur with such small sample sizes suggests that there is potential for the Emotional Literacy Checklists to be a useful tool in evaluating the impact of small group interventions such as Family SEAL. It may be that such a home-school collaborative intervention could bring the parent and teacher views more into line so that they see 'eye-to-eye' in terms of the child's social and emotional competence. Such an alignment of the home and school view of the student might be observed by the difference between the mean percentage scores being smaller in magnitude when pre-intervention surveys are compared with post intervention surveys. It may also result in a change of significance in the mean sub scale scores for the group of concern students when pre and post intervention surveys are analysed. If this is the case the patterns in the movement of scores could be investigated (such as parents/carers moving into line with the teacher view, or vice versa).

In some sense this presumes that the parent/carer and the teacher filling in the survey would also be the home and school representatives participating in the Family SEAL programme. This may not be true in all cases so it would be important to establish whether the survey respondents were also the representatives present at the Family SEAL sessions. One might tentatively suggest that where the survey respondents are not the home and school representatives present at the Family SEAL sessions and the parent and teacher results are still observed to come into line after Family SEAL, that this would provide stronger evidence of the impact of the programme.

6.2.2 Family SEAL Post-Pre Family SEAL comparison of mean ratings

	Non-concern children	
	Difference in parent ratings (N=22) Post – Pre Family SEAL	Difference in teacher ratings (N=30) Post – Pre Family SEAL
self-awareness	+3.1%	+6.3%**
self-regulation	+4.3%	+0.8%
motivation	+3.0%	+2.9%
empathy	+0.9%	-1.9%
social skills	+1.4%	+2.1%

Key: * p≤0.05, ** p≤0.01, *** p≤0.001

Table 56: Post Family SEAL gains in parent and teacher mean percentage ratings for non-concern children (FSRQs1&2)

	Concern children	
	Difference in parent ratings (N=15) Post – Pre Family SEAL	Difference in teacher ratings (N=22) Post – Pre Family SEAL
self-awareness	+6.5%	+7.1%**
self-regulation	+9.3%	+10.5%**
motivation	+7.7%*	+6.8%**
empathy	+4.7%	+9.4%***
social skills	0.0%	+5.1%**

Key: * p≤0.05, ** p≤0.01, *** p≤0.001

Table 57: Post Family SEAL gains in parent and teacher mean percentage ratings for concern children

Discussion of quantitative data

FSRQs1&2: The results of t-test analyses in Tables 42 and 43 demonstrate that post Family SEAL mean gains in parent and teacher ratings of social and emotional skills were reported for children across each of the five aspects of SEAL. The greatest mean gains were reported for the aspect of self-regulation (or managing feelings in the terminology of SEAL). The reported gains for this aspect were the highest in both parent and teacher ratings of 'concern' children, and in the parent ratings of 'non-concern' children. Teachers rated the greatest mean gains for 'non-concern' children in the aspect of self-awareness.

Despite the relatively small sample sizes in this study, statistically significant gains (at the 95% significance level or higher) were reported across all five aspects and for both 'concern' and 'non-concern' children. Application of more stringent criteria for assessing significance at the 95% level or higher, using the Bonferroni correction for experimentwise error for each multiple set of 5 t-tests, would lower the required p-value threshold to $p = \alpha/5 = 0.01$. After making this adjustment none of the gains reported by parents remain significant but there are still significant gains reported by teachers in the aspect of self-awareness for the non-concern children and in all five aspects for the concern children.

Thus, there is some tentative evidence to suggest that Family SEAL has at least a short term impact on the social and emotional skills of concern children as rated by teachers. These pilot results would merit further investigation via a more robust research design.

Shucksmith and colleagues (2007) report equivocal findings from a meta-analysis of the impact of targeted mental-health interventions in primary phase schooling (age 4-11). Gains from long-term (up to 2-3 years), multi-component programmes that included parental engagement were modest when compared with the investment required to sustain them. They did, however, identify evidence of significant gains in both social and emotional competence and improved academic achievement.

Reviews of interventions and programmes involving parental participation suggest that proximal outcomes such as increased social and emotional competence (Shucksmith et al, 2007) and more distal outcomes such as improved academic achievement (Desforges & Abouchaar, 2003) may require programmes like Family SEAL to be firmly embedded in broader, whole-school approaches to improve mental health in schools.

Qualitative results from parental evaluation questionnaires (FSRQ3)

FSRQ3a: Specific responses referring to gains solely for the children were limited. This may be due to the open-ended nature of the questionnaire instrument used in the evaluation and the fact that such evidence was gathered only at the end of the series of workshops. One parent expressed that her daughter had "*Learnt how to express her emotions calmly. Learnt to be nicer and address issues also knowing naughtiness isn't rewarded.*" Whilst another related that her child had learned "*How to be nice to sibling, stop hitting each other and understanding each other better.*" These responses correspond well with the reported gains in self-regulation of feelings discussed above.

FSRQ3b: The evaluation questionnaires revealed strong evidence that parents and carers valued the opportunities that Family SEAL provided to network with other parents with a high proportion of responses including "*Getting to know other parents*" as a key personal gain. Some also extended the networking theme referring to increased opportunities to get to know teachers at their child's school. A number of responses expressed comfort in the knowledge that as parents they were not alone in the problems they were facing including one parent who stated that "*it was nice to discover people have the same issues as me.*" Several parents of 'concern' children expressed how relieved they were to know that parents they considered to have 'perfect' children also wrestled with issues similar to their own. Some responses revealed that Family SEAL "*had an effect on the whole family*" which corresponds with the comments on improved relationships between siblings noted above.

RQ3c: Another key finding from the parent evaluation questionnaires was the value parents associated with spending quality one-to-one time with their child away from their siblings and other family pressures. One parent reported that Family SEAL "*Really opens your eyes to how much time you spend together*", whilst another, referring to how much she valued the one-to-one time with her son described Family SEAL as "*A real life learning experience.*" Some parents responded on behalf of their children including one who said her "*daughter enjoyed the fact that she had my one to one attention without any interruptions.*" Other parents reported how Family SEAL had influenced their parenting approach, such as one parent who stated that she had learned how "*to talk without shouting at them to make them listen.*"

One of the pilot schools felt they could make use of the session-by-session evaluation forms included in the Family SEAL materials (DfES, 2006). These yielded more extensive references to the impact of specific activities. This is well illustrated by the 'journey' that one parent described that she and her daughter made through a series of sessions, highlighting how specific activities had raised opportunities and concerns to be further explored back at home. The responses below are given under the titles for each Family SEAL session to which they refer, which align with the topics used in the Primary SEAL framework outlined in Table 1 above.

New beginnings

"Sarah loved playing the game and came out with answers to the questions that I already knew. No surprises yet except I am worried about how materialistic she is."

Going for Goals

"I think Sarah will love the star chart and aiming for her goal. Hopefully her enthusiasm will continue at home with all the distractions. She loves coming to SEAL and playing all the games. I think she is showing more confidence and our understanding of each other is growing."

Good to be me

"Sarah is not very good at talking about things so usually I watch for signs on how she is feeling. I think fridge magnets with faces and feelings on will be very helpful for Sarah to express herself. We will see."

Relationships

"It was lovely to spend time with Sarah without the distractions at home. I know Sarah looks forward to the sessions and we find it very beneficial to our home life as we have tried a few of the ideas and hopefully will continue to."

The limited qualitative evidence gathered from parents expressed the general benefits of Family SEAL most strongly. Specifically these related to the opportunity for social networking amongst parents and quality time with an individual child. It is difficult to say whether the gains accrued in the pilot project are due specifically to participation in Family SEAL or simply the result of parents engaging in social activities with one another and collectively with their children. There is limited evidence from the qualitative elements of this study to suggest that there were specific gains from engaging with the Family SEAL resources and activities. This may be due to the open-ended nature of the questionnaire instrument used in the evaluation and the fact that such evidence was gathered only at the end of the series of workshops. More specific evidence needs to be gathered possibly via structured questionnaires or interviews/focus groups with parents and triangulated by less subjective sources of evidence such as observations of parent workshops and activity sessions carried out by researchers.

Chapter 7: Conclusions and recommendations

This final chapter seeks to draw together all the elements of this thesis. It will summarise the knowledge gained from the review of the literature on school effectiveness studies of both cognitive and non-cognitive outcomes, and highlight the potential transferable insights gleaned from the more extensive body of work on effects in the cognitive domain. It will then summarise the main findings from both the universal SEAL and Family SEAL aspects of the current study before discussing the limitations and issues arising from aspects of the research design and the analysis methods employed. Finally, a number of recommendations for policy and practice will be made.

7.1 Summary of context

This study set out to determine whether the school effectiveness methodology, used in recent decades to estimate the size of effects on students' academic or cognitive outcomes, could also be usefully applied to measures of students' non-cognitive outcomes of education, with a particular focus on the development of social and emotional skills related to the Social and Emotional Aspects of Learning (SEAL) programme.

The context of the study was the familiarity of schools with measures of student attainment and progress in cognitive outcomes, and handling such data in order to inform decision making around student progress, curriculum development, and other aspects of school improvement through a process of data informed self- evaluation. A review of the development of ubiquitous student level attainment and progress data in the form of value added models made available via RAISEonline and Fischer Family Trust revealed that such data, though based on relatively sophisticated statistical models, are used widely in schools to monitor and evaluate progress. These measures of cognitive attainment and progress have been developed from insights gleaned though several decades of school effectiveness research which indicate that a range of student and school level factors exist that consistently help to explain significant proportions of the variance in student outcomes in cognitive tests and assessments. The development of value added measures of progress has been made possible by the introduction of national scale surveys of student demographic factors in the form of the pupil level annual school census, or PLASC, which has led to the development of the National Pupil

Database (NPD); one of the world's largest longitudinal data sets of students' academic progress and contextual factors.

School effectiveness studies have shown that the greatest proportion of the variance in academic outcomes is explained by students' prior academic attainment. This accounts for around 50% of the variation in outcomes. School effectiveness research has also shown that, even after adjusting for this wide range of student and school level factors that are beyond the control of the school, the size of the school effect on student outcomes remains significant with a value around 15%. CVA models for progress across various phases of education in England report a more modest range for the school effect of between 8-12%. Though seemingly modest in terms of scale, the upper end of this range of school effects would be equivalent to making a full grade more progress in *each* of the best 8 subject examined at age 16 through GCSE examinations and variance at the school level suggest that some schools will be adding more value than this to the outcomes for some of their students. While the development of VA and CVA models was driven predominantly by an agenda to evaluate differences between schools for accountability purposes such as tables of school performance (or league tables as they are more commonly known), these sophisticated measures also allow schools to evaluate the variation in student outcomes within their schools. Schools have the facility to monitor progress made by students through grouping by demographic factors such as gender, age in year, social economic status, ethnicity, level of additional educational need. What is not generally evaluated is the obvious "missing level" between that of the student and the school in the form of class groups, where effects have been shown to be as big or even greater than school effects.

Web-based databases such as RAISEonline and FFTlive, draw down data from the NPD to produce predetermined analyses of student progress that, over time, schools have become familiar with in order to inform their monitoring of student progress and subsequent development planning. These databases have afforded schools with sophisticated tools to look at variation within school as well as between schools, although FFT models may be more useful for within school analyses due to avoiding some of the issues associated with application of corrective adjustments (shrinkage) that are part of the multilevel modelling framework employed in RAISEonline VA and CVA models.

Despite such data being available to inform school improvement, longitudinal studies have shown that schools find it hard to sustain year on year improvement beyond a period of 2-3 years, either in relative terms indicated by value added progress made by students compared to those in other schools, or in absolute terms in the form of the raw attainment of students. Thus, maintaining data informed improvement remains a challenge for schools.

Similar measures for non-cognitive outcomes of schooling are not available. The introductory chapter to this thesis shows that, in the absence of available measures of progress, the introduction of a programmes like SEAL, and its focus on development of non-cognitive skills, may lead to schools utilising measures of students' progress and attainment in the cognitive domain as evidence for the impact of such programmes. This is particularly the case if, as is the case with SEAL, strong claims have been made for impact on the academic outcomes of schooling. This is problematic since the relationship between SEAL and academic outcomes is complex at best with limited evidence for any relationship between the two, and any effects of SEAL on academic outcomes are unlikely to be seen in a short time frame, and will be just one factor alongside many that will impact on academic outcomes. While non-cognitive outcomes do not receive the attention and focus poured upon high-stakes academic outcomes, they are nonetheless viewed as valuable outcomes of schooling. As the culture around data use suggests that we seek to "measure what value", then schools will naturally look to something that allows them to monitor and evaluate progress in these aspects of education, as well as gains made in the cognitive domain. The introduction of measures of non-cognitive outcomes focused on more proximal gains to be had from developing dispositions and skills promoted by a programme like SEAL has the potential to provide data to inform school improvement in this domain, without the additional tensions accompanying the accountability agenda associated with outcomes in the cognitive domain.

A note of caution needs to be sounded as the other half of the "measure what we value" adage is that we end up "valuing what we measure", and that can lead to unintended consequences. This might be through take up of measures for accountability purposes, both within schools, through target setting for students or holding teacher accountable for the outcomes of their students, and also between schools in terms of local and even national measures to compare schools. A recent government consultation proposed that student well-being and perceptions of schools could be quantified and incorporated within public measures of school performance in the form of "school report cards" (DCSF, 2009b).

7.2 Summary of literature review on non-cognitive measures

A review of the literature on studies of non-cognitive outcomes suggests that there may be much insight to be gained from utilising a school effectiveness research framework within the non-cognitive domain. That said, the extent of research into non-cognitive outcomes is far more limited than that available for cognitive outcomes, and so the sparse nature of the evidence suggests that it would be wise to be tentative in declaring the utility of measures of effectiveness in the non-cognitive domain.

One of a number of challenges presented by the study of non-cognitive outcomes lies in the wide range of outcomes that can be measured. The review (Section 3.2) in this thesis alone considered more than 25 different foci for non-cognitive outcomes including measures of wellbeing, various self-concepts such as self-esteem and academic self-concept, levels of anxiety, attitudes to study and school, social skills, behavioural outcomes, bullying behaviours, emotional awareness, and moral and citizenship values. While this issue is not unique to measures of non-cognitive outcomes it does make it challenging to compare the findings of studies of the range of various outcomes.

Another challenge for researchers is in formulating reliable and valid measurement models of non-cognitive outcomes. The nature and psychometric properties of the data collected can make this difficult and several of the studies reviewed in this thesis made compromises in the development of measurement models in order to achieve or approach generally acceptable model fit, both in terms of construct validity through various factor analysis methods (PCA, EFA and CFA) and in scale reliability, usually measured by Cronbach alpha. This was the case for scales developed utilising teacher ratings of students' non-cognitive outcomes, as well as for scales based on student self-report data. Similar issues were encountered in developing the measurement model associated with this study, and the implications of this psychometric challenge will be considered in more detail later in this chapter, where the limitations of the study are discussed.

These challenges notwithstanding, studies of non-cognitive outcomes based on the school effectiveness model reveal similar insights to those gleaned from more traditional cognitive focused studies. There are student and school level factors that show significant associations with the level of outcomes in the non-cognitive domain. Key independent variables include gender (in that boys tend to have less positive outcomes than girls), and a strong age trend is observed, such that older children self-report (and are reported to have) less positive outcomes over time. This was observed both in multiple cross-sectional studies utilising similar measures across different ages, as well as for cohorts moving through longitudinal research studies. Socioeconomic status and ethnicity were also found to be associated with some non-cognitive outcome measures. There were some non-cognitive outcomes, such as being in bullying behaviour, which did not show such associations.

Meta-analyses have reported a wide range of associations between non-cognitive and cognitive outcomes of schooling, while some more focused studies report low or no association so the evidence for associations between the two types of outcome appears to be equivocal, although subject specific and general academic self-concept measure do appear to have stronger associations with their cognate academic outcomes, though it is not clear what the causal relationships are in these cases.

As with studies of cognitive outcomes the application of variance partitioning available within a multilevel modelling framework indicates that significant school and class level effects are observed with non-cognitive outcomes. The school effects are, however, appreciably smaller than those observed for cognitive outcomes, when comparing similar models (i.e. null models, models adjusted only for contextual factors, value added type models adjusted for prior attainment, and contextual value added models adjusted for both prior attainment and context). Null models for cognitive outcomes reported proportions of the variance at the school level in the order of 20 to 30%, whereas null models for non-cognitive outcomes range from less than 1% through to 15% with many being at the lower end of that range. As with cognitive outcomes, adjusting for factors such as context and prior levels of attainment in the non-cognitive outcomes tends to reduce the proportion of the variance at the school level, although there were some exceptions to this where the school effect was found to increase from the level of the null model after adjusting for prior levels in VA type models. The often quoted school effect of around 8-15% for cognitive outcomes is determined after adjusting for prior attainment and contextual factors outside the control of the school, and is much larger than the school effect observed for non-cognitive outcomes in similar VA/CVA type models. For some non-cognitive outcomes adjusting for other factors in this way results no significant school effect remaining. In those studies where three-level MLMs were analysed the class level effect is usually similar to or greater than the school level effect in the same three-level model. Once again, this is similar to the pattern of variance portioning observed in three level studies of cognitive outcomes.

Thus, studies of non-cognitive outcomes suggest similar possibilities in terms of data for school improvement but with diminishing returns based on the lower levels of school and class/group level effects. This would suggest that the utility of non-cognitive outcome measures is likely to vary from scale to scale, and so the nature of the scale and magnitude of the school and class effect needs to be considered carefully. Again, this will be taken up in considering the outcomes of the present study below.

7.3 Summary of findings of the present study in the context of pre-existing body of knowledge

The measurement model derived via CFA for the SEAL Survey yielded 7 dimensions related to aspects of the SEAL programme: *self-image*, *managing feelings*, *managing behaviour*, *independence*, *resilience*, *friendships*, and *attitudes to teachers and school*. A basic descriptive analysis suggested a strong association with age. Analysis of two- and three-level multilevel models also indicated that a number of student and school level factors were significant in explaining the variance in the student outcomes. The association with age for all seven dimensions of was confirmed and the addition of age to each of the MLMs had the biggest effect on reducing the school and class level variance. Gender was also found to be significantly associated with five of the seven dimensions (*self-image*, *managing feelings*, *managing behaviour*, *resilience*, and *attitudes to teachers and school*).

Being involved with bullying behaviour had a more mixed ability to explain variance at the student and/or school level. Being a victim of bullying had a significant association with *self-image*, *managing feelings*, *resilience*, and *friendships*. Reporting being a bully was significantly associated with *self-image*, *managing feelings*, *managing behaviour*, *independence*, *friendships*, and *attitudes to teachers and school*. While the student level bullying factors were generally significant, there were, by contrast few significant peer-effects of bullying. Bullying behaviours as school level factors were only significant for *self-image* and *resilience* (standard deviation of reporting being a victim of bullying) and *for attitudes to teacher and school* (standard deviation of reporting being a perpetrator of bullying). Other school level context factors including the mean level of deprivation and absolute and relative trend in raw attainment and progress measures at the school level were only found to be significantly associated with *resilience* (school level socio-economic status and the trend in attainment levels reached in high stakes tests at the end of the period in school) and *friendships* (mean socioeconomic status only).

The size of school and class effects were greatest for the dimension a*ttitudes to teachers and school*, which had a school effect of 13.8% in a two level-null model and 12.8% in a three-level null model. The size of the class effect in the three-level null model was 11.6%. For the other six SEAL Survey dimensions the size of the school effect was much more modest ranging from just under 0.5% to just over 3% in two-level null models, and from non-significant to 2.6% in three level null models. The class effect for these dimensions ranged from 1.6% to 4.3% in three-level null models. In every case the class effect was greater than the school effect except for the dimension *resilience*.

After addition of the student level independent variables (*age*, *gender* and *victim* and *bully*) only four of the seven dimensions retained a significant school level effect: *managing feelings*, *managing behaviour* and *resilience* (all around 1.8%) and *attitudes to teachers and school* (just under 6%). In three-level models adjusted for the same student level factors these school effects were still significant but reduced to just above 1% for *managing feelings*, *managing behaviour* and *resilience*, and to 3.8% for *attitudes to teachers and schools*. By contrast class effects were significant for all seven dimensions ranging from just over 1% to 2.3% for all dimensions except attitudes to teachers and school which had an 8.1% class effect.

These findings were very similar to those observed in pre-existing school effectiveness type studies of non-cognitive outcomes where strong age and gender effects had been observed. The use of the *bully* and *victim* variables as explanatory variables in the MLMs was supported by improved model fit and significant point estimates for one or both of the variables across the whole set of seven dimensions. It's not possible to say whether the bullying behaviours associated with the responses to these items in the SEAL Survey are directly associated with the non-cognitive dimensions. It may be that they are capturing an element of other background factors at the student level not measured in this study. The relatively modest amount of variance explained (approximately 10%) after adjusting for the responses to the bullying items suggest they are not acting as proxies for prior levels of non-cognitive attainment and thus inclusion of prior levels of attainment for the corresponding SEAL Survey dimension, to form a value added or contextual value added type model, would reduce the size of the school and class effects still further potentially rendering more of these non-significant. The inclusion of other student level factors not collected in this study would also be beneficial, and findings from large scale longitudinal studies such as LOSO and EPPSE suggest that home and familial background factors could be useful additions to datasets for non-cognitive outcomes, as well as other classic student level factors drawn from school effectiveness studies such as ethnicity and socioeconomic status.

The magnitudes of the school and class level effects observed in this study are similar to those drawn from the pre-existing research base though at the lower end of the range of school and class level effects considered in the review of studies. The higher magnitude of school and class effect for attitudes to teachers and school is also in line with previous research

The observation that school level factors such as mean socioeconomic status and measures of school effectiveness and school improvement trends are generally not significantly associated with non-cognitive outcomes may be in part due to the low levels of school and class level variance left to explain, but the fact that these were also not significantly associated with the dimension *attitudes to teachers and school*, which had the highest school and class effects, lends weight to findings that non-cognitive outcomes are generally not highly associated with cognitive outcomes. There is scope therefore, to conduct studies to investigate whether other factors at school level can be found to be associated with the outcomes of SEAL with the cautionary note that the low levels of school and class level variance for most of the dimensions suggest that there may not be ‘rich pickings’ to be had in this area.

From the analysis of the impact of Family SEAL on students’ social and emotional competence the results suggest that students who are *not* a cause for concern in terms of their social and emotional development make only modest gains in competence in these outcomes through involvement with Family SEAL. The only significant post-intervention gain for non-concern children was in the teacher rating of *self-awareness*. For children identified as a cause for concern the results indicate that potentially greater gains are to be had especially in terms of teacher ratings of the aspects of SEAL. Significant post-intervention gains were observed in teacher ratings of all five aspects of SEAL with the greatest gains in *self-regulation* (managing behaviour) and *empathy*. This is an encouraging indication that engagement with parents and family in the development of non-cognitive skills might lead to the greatest gains being observed when the student is at school rather than at home.

7.4 Limitations of the present study

A number of limitations in the design and implementation of this study suggest that care needs to be taken in drawing inferences and also making recommendations for policy and practice.

The most challenging issue arose in working with data derived from the Universal SEAL Survey in order to develop a suitable measurement model, through establishing the construct validity and internal scale reliabilities of the model via factor analysis. The psychometric properties of the data collected from the SEAL Survey were such that extensive modifications were required after the initial CFA, in order to arrive at a measurement model that displayed appropriate model fit. This compromised the advantages of adopting the *a priori* approach of CFA to determine the underlying structure of latent variables. The *post-hoc* PCA/EFA that was carried out on the full data set suggests that the modifications to the model specification, made to the model to improve fit, may have impacted on some dimensions more than others. The dimensions *attitudes to teachers and school* and *resilience* were most closely related when comparing the CFA and PCA models. *Managing feelings* and *friendships* had all or the majority of core items in common, while *managing behaviour*, *independence* and *self-image* had few or no items in common between the CFA and PCA models and so are potentially more problematic in terms of their properties as measurement scales.

In balance to this the criteria used to determine good fit for the CFA measurement model were more stringent than some of those adopted in some previous studies. The requirement imposed was to reach a value approaching or above 0.95 for approximate fit indices and below 0.05 for the error approximation index (RMSEA). One of the challenges in working with literature reporting the development of measurement models, especially when this is described in reporting of a wider research study, is that it is often difficult for the reader to determine the number and nature of the steps taken during the process of model development. This is in no small part due to the space restrictions imposed by the word limits associated with the publication of journal articles. The numerous and extensive research reports available for the EPPSE study are extremely helpful therefore, in that the EPPSE researchers take the opportunity afforded by being able to write a far more expansive report to give a much greater level of detail about model development, though this material was usually contained in an appendix to the main report.

In the EPPSE study the final measurement model, derived via CFA, for the set of social-behavioural factors at the end of Year 9, was accepted as having goodness of fit “well within the conventional range of acceptability" (NFI = 0.91 and RMSEA = 0.043), although the researchers did indicate that a fit of 0.95 for NFI would be considered to be a sign of superior model fit (Sammons et al, 2011b: 104). It is worth noting that this was for factors derived from the SDQ teacher report data which indicates that the problems of data quality and psychometric suitability when working in the non-cognitive domain are not restricted to student self-report data.

In fact, self-report data caused some more substantial issues for development of some of the measurement models in the EPPSE study. This is illustrated by the development of the measurement model for the survey of Year 5 student dispositions. A total of ninety items were administered via two questionnaire instruments. They intended to utilise a combined EFA/CFA approach for model building which would match that proposed by Mulaik and Millsap (2000). An initial exploratory analysis (using principal components extraction and varimax rotation) yielded 22 factors with eigenvalues of 1 or more which accounted for 42% of the total variance. The results were described as "too unwielding [sic] to test as a theoretical structure in confirmatory factor analysis" (Sammons et al, 2008a: 53). So the items were divided into three main groups (student self-perceptions, student views of school and other factors) with the student self-perception group consisting of 6 factors with Cronbach alpha values ranging from a=0.50 to 0.83. Two CFA models were run using all six factors and one with 4 factors (removing the friendships factor with the lowest Cronbach alpha of a=0.5 and pupil values as it was felt the items did not relate to the school experience of the students). It is not clear from the report which items were allocated to each factor at this stage, but the model fit is described as unsatisfactory. The researchers then returned to the 6 factor model and removed a number of items on grounds of being highly skewed, having low factor loadings, or loading on multiple factors (Sammons et al 2008, 54). Also, a whole set of items was removed since it formed "a small, poor factor" (ibid). The resulting 5 factor model without modifications was described as having unacceptable fit so further items with cross-loading were removed which improved fit. Finally a further whole factor was removed since removal of an item cross loading on another factor dropped the Cronbach alpha to a=0.50). After this series of modifications, the final model with four factors, was described as being the best solution. The researchers report only two sets of model fit statistics from the model modification process, and it is not completely clear to which of the original models the initial fit statistics refer, as no details of the model structures are given other than the identity and loadings of the items for each factor in the final model. The initial CFA produced poor model fit with chi square statistic of 1092.148 (p<0.0001, df=183) and fit indices of GFI=0.923, CFI=0.874, NFI=0.853 and RMSEA

= 0.060. The final model had a chi square statistic of 455.423 (p<0.0001, df=146) and fit indices of GFI=0.951, CFI=0.919, NFI=0.898 and RMSEA = 0.049. Thus, a number of iterations of model modifications still made it challenging to determine a model with appropriate goodness of fit. Despite this authors still describe the questionnaires in the executive summary to the report as yielding “robust measures of pupils’ self-perceptions” (Sammons et al, 2008a: ii).

The development of the measurement model for the Year 2 survey of students’ dispositions was potentially even more challenging. In this case the researchers felt they were able to carry forward the results of the initial exploratory analysis which identified five factors (via PCA with varimax rotation). One of the factors was removed before progressing to CFA due to a low Cronbach alpha score of 0.4. The remaining four factors consisted of 19 items (one factor with 7 items and the other three consisting of four items). The researchers report several rounds of model modification, although detailed specifications of the intermediate models are unfortunately not identified. They do report a number of steps were involved in the refinement of the measurement model including making use of model modifications indicated by the software and items “that either loaded on other factors or cross loaded with other items were taken out through a series of re-runs of the model” (Sammons et al, 2008a: 48). They go on to indicate that the model statistics “still failed to reach acceptable criteria, so a number of models were tried pulling out different combinations of questions.” (*ibid*). Thus, the model development experience of the EPPSE researchers indicates how challenging it can be to work with self-report data from students on non-cognitive outcomes.

The issue of scale reliability for self-report data was indicated in the model development process described above. The resulting four factors in the final measurement model, reporting on 19 of the original 90 items had Cronbach alpha values ranging from 0.62 to 0.76 with two of the values being lower than the conventionally accepted threshold of 0.7, while the self-report survey of Year 2 dispositions yielded four similar factors with Cronbach alphas ranging from 0.52 to 0.69 (Sammons et al, 2008a). The problems of scale reliability working with data in the non-cognitive domain are also indicated in the reporting of other studies. For example, the original wellbeing instrument utilised in the LOSO study yielded eight factors and were described as demonstrating good psychometric properties with Cronbach alpha values ranging from 0.63 to 0.88 (Van Damme, 1997). The Rosenberg self-esteem scale (1979) modified by Salmivalli et al (1999) to score as one item through Guttman scaling, had a Cronbach alpha of 0.64.

In the Universal SEAL element of this study only makes use of student self-report data. Gadeyne et al (2006) have raised the importance of using multiple perspectives including teacher ratings and the views of parents on students' non-cognitive outcomes. They indicate that issues of measurement error in working with data on non-cognitive outcomes may be the reason why the size of the school effect for non-cognitive outcomes is appreciably smaller than that observed for cognitive outcomes. Gadeyne et al's suggestion is that the measurement error may derive from focusing only on students self-reports rather than utilising multiple measures of non-cognitive outcomes. They do recognise that including parent and teacher perspectives would bring additional implications in terms of the time and cost of data collection. They are also critical of the fact that previous studies had focused mainly on factors such as achievement motivation and academic self-concept with relatively few studies taking a broader focus on variables such as behavioural or social outcomes. The experience of the present study, and also that of the EPPSE study, suggest that the issue of measurement error is also prevalent when working with data on behavioural and social outcomes and the EPPSE study shows that data derived from teacher ratings are also not free of measurement error concerns.

A further limitation of the present study was associated with using a survey instrument designed for KS2 students (aged 7-11) across a much wider age range, namely the full primary to secondary continuum from ages 5 to 16. The intention was to allow for a common instrument to support discussion of the data across all year groups and especially to support he important transition from Primary to Secondary SEAL. These advantages may be outweighed by the increase in measurement error that comes from extending the data collection to participants beyond the original intended age range. The literacy levels of the youngest students in Years 1 and 2 made it necessary for Teaching Assistants to support the youngest students which may introduce another source of error. Differing interpretations of survey items across the older age range from those of KS2 students may also increase measurement errors at the older end of the age range. An alternative would be to follow the example of the EPPSE study by utilising different surveys across the age range with similar dimensions common to the measurement model for each age specific instrument, though that approach could explain some of the issues of model specification experienced by researchers on the EPPSE study.

While the descriptive data provided for the schools suggests that the participating schools are not dissimilar in terms of socioeconomic context to the other schools in the local authority, they were nonetheless a self-selecting group rather than a random sample and so are unlikely to be representative of all schools in the LA. This is especially the case for the primary phase schools which were under represented in the participating group which might therefore exacerbate any self-selection bias. Such bias will have an effect on the variance at the school and class level, and so will have impacted on the magnitude of the school and class level effects observed; a key focus of this study. This would possibly lead to under estimation of school and class effects due to self-selection of "already engaged" schools.

By contrast, the lack of a true prior attainment measure for each of the non-cognitive outcomes is likely to mean that the size of the school and class level effects are over estimated by this study, as adjusting for prior levels of attainment is likely to reduce overall variance, including the proportion at the class and school levels. This may not always be the case, however, as the EPPSE study found that adjusting for prior measures of non-cognitive outcomes can occasionally result in an increase in variance at the higher levels. It is hard to say whether this observation from the EPPSE study is an artefact of the wide distribution of the initial pre-school cohort across primary and secondary schools leading to a large number of schools with very few children in them, so this would need careful further study.

No measure of programme fidelity was made, neither in the universal SEAL nor the Family SEAL elements of this study. It has been argued earlier in this thesis that issues caused by the of programme fidelity are inherent in the non-statutory nature of programmes like SEAL, where schools, and teachers within schools, are given a great deal of autonomy to shape a curriculum that they perceive is right for them. Introduction of a more prescribed curriculum may have the unintended consequence of making it less likely that schools will engage with the programme. Nevertheless, capturing a measure of programme fidelity would be beneficial in any analysis where data from multiple schools are pooled together and could be a useful school process factor to help explain the variation between schools, though clearly the low school level effect suggests that this might have limited utility so consideration of cost vs benefits would be needed to determine whether this would be worthwhile from a practice and policy point of view.

For evaluating the impact of Family SEAL the data from all seven schools were pooled without adjusting for school characteristics nor the issues arising due to any lack of programme fidelity. In a more robust study there would be potential to adjust for a range of school characteristics and, with a suitable number of schools, to conduct a multilevel analysis to partition the variance in outcomes in order to determine the school level variance. Even with such a small number of schools engaged in the pilot it soon became clear that each school implemented

Family SEAL in a unique way. Some differences we noted in terms of delivery in the pilot schools included:

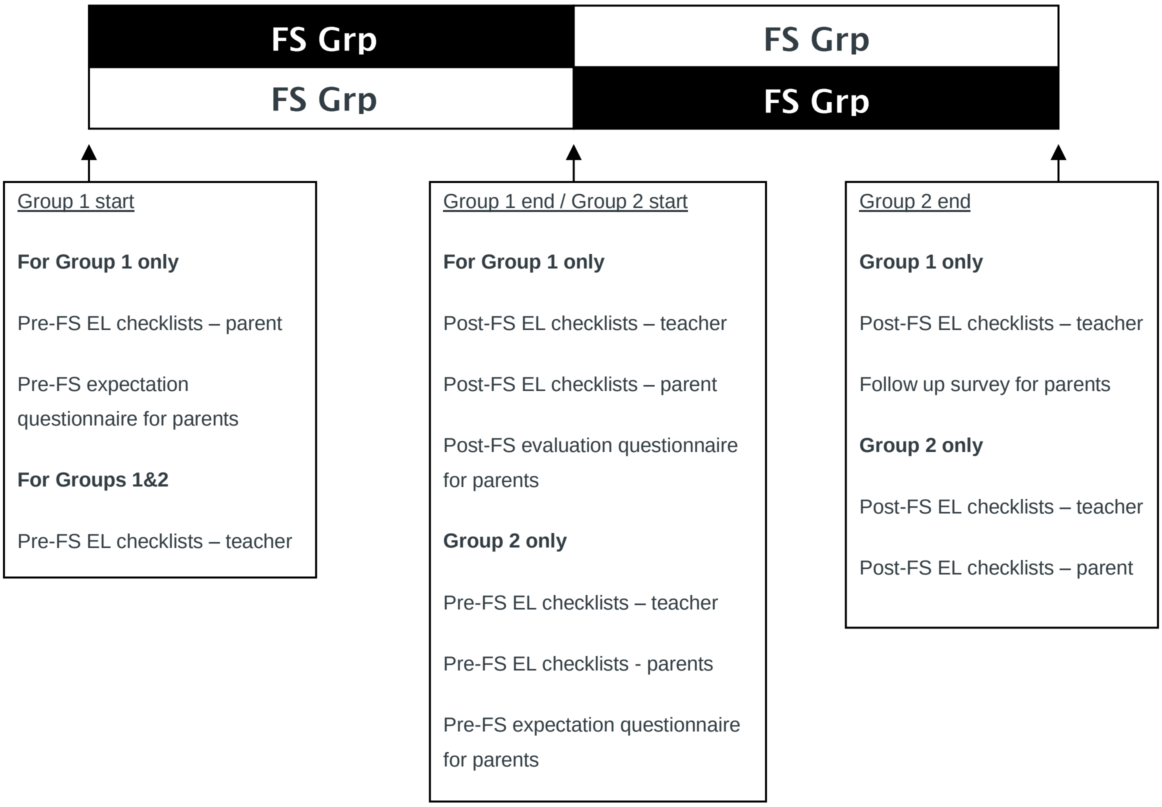
- varying use of certain Family SEAL resources, especially the digital presentations provided for each workshop session,
- varying degrees of focus on communicating and illustrating to parents the universal approach to SEAL adopted in the school.

The issue of programme fidelity may have been made more acute in this small scale study due to the involvement of external facilitators who brought to bear their experience of working on parenting programmes such as the Webster-Stratton (2000) Incredible Years programme. Webster-Stratton demands a high degree of fidelity in her own parenting programmes (Webster-Stratton, 2004). Humphrey et al (2008: 84) have also discussed the issue of fidelity in their evaluation of the small-group SEAL programme. They note that whilst some facilitators are relieved to have a full set of resources pre-prepared for them, others feel the need to adapt the resources to varying degrees in order to tailor the material for their group. The evidence Humphrey and colleagues gathered from the small group SEAL facilitators matches anecdotal comments we received that programme fidelity is more likely to become an issue as facilitators grow in confidence and experience as internal facilitators suggested they might be more prepared to modify and adjust the activities included in the Family SEAL resources after the initial experience and confidence gained from working once through the programme. Fidelity monitoring procedures could be implemented in further studies in the form of questionnaires to monitor elements of expected content and specific facilitator and participant behaviours. The resulting data would allow for an adjustment for programme fidelity to be made in the analysis. The issue of determining programme fidelity for Family SEAL might be more straightforward due to the smaller scale involved and the intensive nature of the intervention compared to the implementation of universal SEAL.

The lack of an experimental design in the Family SEAL study makes it impossible to draw causal conclusions as to whether gains in social and emotional literacy as reported by parents and the significant gains reported by teachers are due solely to involvement in Family SEAL rather than other interventions employed in the schools such as universal teaching of SEAL to all students in the school. While there is tentative evidence to suggest that students identified as a cause for concern in terms of their social and emotional development may have most to gain from Family SEAL, despite the brief time period between pre and post intervention measurements, it is not clear if the gains observed are derived from Family SEAL. This is also compounded by the limited qualitative data collected which suggest that parent report benefitting most from simply engaging with other parents and spending one to one time in school with their child, which are experiences not necessarily unique to Family SEAL. Thus it may be a case of form being more important than substance in any potential effect at work.

There is a likely source of bias in the findings derived from this small scale study in that all the participating families were self-selecting volunteers and so are unlikely to be representative of all families that might engage with Family SEAL. One might argue the effect of this self-selection bias is likely to be more pronounced in the very families that stand to gain most from engaging with the intervention, based on the analysis made in this study.

A further limitation of the evaluation of Family SEAL is the potential bias introduced by the non-blind ratings of students' social and emotional competence. While it would have been possible to organise the pilot project so that teachers were not aware which of their students were engaged in Family SEAL, such arrangements would almost certainly have had practical implications for each school in terms of suitable venues or the timing of Family SEAL, and also implications for parents such as the need to arrange child care for school age siblings if Family SEAL was run during school holidays or after school. Randomised control trials are both controversial and difficult to implement in educational settings, although Tymms and colleagues (2008) have recently argued to the contrary. This is especially so when an intervention is implemented for the first time. One alternative way forward where schools plan to run two or more Family SEAL groups during each year is to use a 'crossover design' such as that used by Humphrey et al (2008) in their evaluation of small group SEAL interventions. In this approach evidence is gathered at three time points during the year to allow for comparison between the two groups that will eventually all engage in Family SEAL.



Group 1 start

For Group 1 only

Pre-FS EL checklists – parent

Pre-FS expectation questionnaire for parents

For Groups 1&2

Pre-FS EL checklists – teacher

Group 1 end / Group 2 start

For Group 1 only

Post-FS EL checklists – teacher

Post-FS EL checklists – parent

Post-FS evaluation questionnaire for parents

Group 2 only

Pre-FS EL checklists – teacher

Pre-FS EL checklists - parents

Pre-FS expectation questionnaire for parents

Group 2 end

Group 1 only

Post-FS EL checklists – teacher

Follow up survey for parents

Group 2 only

Post-FS EL checklists – teacher

Post-FS EL checklists – parent

Figure 42: Crossover design for evaluating Family SEAL in schools running two groups in the same school year.

7.5 Recommendations for policy and practice

7.5.1 Evaluating universal SEAL through the SEAL Survey

This study applied a school effectiveness approach in order to examine the nature and extent of effects of specific non-cognitive outcomes of schooling related to the SEAL programme, both in terms of universal SEAL provision and the more intensive Family SEAL intervention for groups of children and their parents/carers. It set out to determine some of the student and school level characteristics associated with these non-cognitive outcomes and to estimate the size of school and class level effects in order to compare them with similar effects for cognitive outcomes. For Family SEAL it sought to establish whether the intervention might lead to gains in social and emotional competence for participating children and to compare any reported gains in the home or the school environment.

1) The size of the student level variance from the SEAL Survey self-report data together with significant associations for age and for gender with the seven non-cognitive

dimensions derived from the survey suggests that there is scope for the SEAL Survey data to provide useful insights into the level of attainment in each dimension and potentially to monitor developmental progress in each dimension. As the gains made through universal SEAL are likely to be established over long time periods adjustment for the general age trend would be needed in order to establish whether progress was being made over a time period spanning across, say a key stage, or several school years.

In line with school effectiveness data for cognitive outcomes, it would be useful to collect other student level data such as measures of socioeconomic status and ethnicity to examine whether these factors are also associated with each of the non-cognitive outcomes and so might be useful for looking at the progress of groups.

2) The magnitude of the class effects suggest that *within* school analysis might be fruitful determining the progress of groups of students and especially for class groups, though this is more appropriate in the primary phase, and in secondary schools where SEAL might be implemented through PSHE teaching in tutor groups. This might help to identify areas of best practice and effectiveness in SEAL within the school. As ever, such data should be triangulated with other sources to draw a balanced view of “what works” in universal SEAL provision.

NOTE: Data quality issues suggests that extra caution may need to be taken when making use of and drawing inferences the outcomes of the dimensions *self-image*, *independence* and, to a lesser extent, *managing behaviour*. Data quality issues also indicate that the use of these data for target setting and monitoring, particularly where it has a strong accountability focus is not advisable as the measures are not robust enough to support this.

3) The modest size of the school effects observed in this study, coupled with the data quality issues described above, indicate that the use of measures of non-cognitive outcomes to make judgments about the quality of provision and practice between schools should be proscribed. The limited size of the between school variance suggests that school level comparisons would be invidious. This has implications at both the local and national levels where such between school comparisons form a core part of public accountability policies.

Between school comparisons *might* be beneficial for identifying best practice or substantial progress through less formal self-evaluation analyses (such as at LA wide SEAL briefings for SEAL coordinators), where the focus is on identifying the most effective schools in terms of attainment or gains made. That said, the complex nature of schools always makes it difficult to be sure that differences in the way SEAL is being implemented between schools is the factor behind any limited between-school variation observed.

The provision of a value added analysis for schools within a LA region might be a useful service to help benchmark progress across the LA using the informal, self-evaluation approach describe above, though as the adjustment for prior attainment levels is likely to reduce school level variation even further this should be piloted first. If comparing the levels attained or progress made by multiple cohorts of students, schools/school leaders may need reminding about the relative nature of year on year value added comparisons, and that few schools maintain gains across multiple year groups for more than 2-3 years, even for cognitive outcomes.

NOTE: An exception to this might be the *attitudes to teachers and school* dimension which is the most robust of the seven SEAL Survey dimensions and also demonstrates a much stronger school effect. Any use of attainment or value added progress measures in this domain would need further research and validity and reliability checks to establish whether they are fit for purpose, especially where that is for more accountability focused comparisons.

4) The establishment of some form of implementation fidelity measure would be very useful for any future research into SEAL related outcomes, and may also help provide evidence of some of the school process factors that account for the attainment or progress gains made in the most effective schools. A potential cost-benefit consideration would be needed to weigh the time and costs associated with developing a programme fidelity measure with the insights that might be gleaned from it in the context of limited between-school variation.

7.5.2 Evaluating the impact of Family SEAL

Despite its limited scope the findings from small scale evaluation of Family SEAL merits more detailed research into its effectiveness.

1) An experimental or quasi-experimental study (such as one utilising the crossover design described above) should be undertaken to establish if the gains observed with school are as a result of engagement in Family SEAL.

After adjustment for experimentwise error, the only consistent significant post-programme gains in social and emotional skills were reported by teachers for those children who had been identified as causing prior concern in their social and emotional development. Whether this suggests that Family SEAL would be most effective as a targeted intervention for such ‘concern’ children rather than a universal one offered to all is impossible to conclude from the limited evidence here. Such a targeted approach to Family SEAL would, however, also need to be weighed in the light of limited qualitative evidence gathered from parents involved in this pilot study of the benefits they perceived of building wider social networks with other parents from a wider social group with whom they may not otherwise have associated.

2) Care should be taken if targeting Family SEAL only to children (and their parents/carers) causing concern in terms of the students’ social and emotional development. Any further research into the impact of Family SEAL could helpful incorporate a design element that allows such targeted approaches to Family SEAL to be included within the study so that gains of targeted and mixed/universal groups might be compared. Getting concern children to Family SEAL is the challenge but the involvement of parents of children not necessarily causing a concern seems to be a key part of the parental experience and the feeling of shared challenge and overcoming difficulties by all parents (the perception that raising children is a levelling experience) could be key to retaining the parents of non-concern children after initial fears that I’m not like them have subsided. Careful training and groups work skills on the part of facilitators is likely to be key in enabling the social mix to bear fruit.

Appendices

Appendix 1

About Me and My School Questionnaire

Name: ____________________________ Class: ___________

Age: _______ Please tick - Girl □ or Boy □

Here are some statements about you. Read each statement and then put a tick in one of the boxes. Make sure you do each statement.

		Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
1	I try to help people when they are unhappy.					
2	I often forget what I should be doing.					
3	I know what things I'm good at.					
4	I often lose my temper.					
5	I get annoyed when other people make mistakes.					
6	I can describe how I am feeling most of the time.					
7	I get upset if I don't do something well.					
8	I find it difficult to make new friends.					
9	I know when people are starting to get upset.					
10	If I find something difficult I still try to do it.					
11	I'm easily hurt by what others say about me.					
12	I calm down quickly after I have got angry or upset.					
13	Other children let me play with them.					
14	I laugh at other children when they get something wrong.					
15	I am usually calm.					
16	I have lots of friends at school.					
17	I find it easy to pay attention in class.					
18	I worry about the things I can't do well.					
19	I like my class.					
20	I work quietly in my class.					
21	I want to do well in my work.					
22	I sometimes leave the room without permission.					
23	I get on well with my teachers.					
24	I sulk or argue when I am told off.					
25	I can ask a question and wait for an answer.					
26	I can take turns.					

Appendix 1

27	I listen well in class.					
28	I am happy being me.					
		Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
29	I am good at some things.					
30	I can work without my teacher’s help.					
31	I get up and wander around the classroom.					
32	Playtime is fun.					
33	Our teachers are fair in the way they treat us.					
34	It is easy to work in my class.					
35	I can talk to my teacher about anything.					
36	I am sometimes picked on or bullied by other children.					
37	I can tell the teacher if anyone is unkind to me.					
38	I sometimes bully or pick on other children.					
39	I like coming to school.					
40	I like lunchtime.					

☺ Thank-you for your help ☺

SE&M = self-esteem and motivation, PE = perceptions of own emotions, AE = awareness of own emotions, AEO = awareness of emotions in others, ASW = anxiety about school work, SSR = social skills and relationships, ASRT = attitudes to school and relationships with teachers, AW = academic work

Appendix 2

Student SEAL self-rating survey – final model parameters

Regression Weights: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P
Q24 <--- Managing Feelings	-.658	.050	-13.040	***
Q10 <--- Independence	1.000			
Q39 <--- Att T&S	1.000			
Q11 <--- Resilience	-1.269	.137	-9.233	***
Q18 <--- Resilience	-1.080	.079	-13.718	***
Q7 <--- Resilience	-1.000			
Q20 <--- Managing Behaviour	.853	.033	25.719	***
Q15 <--- Managing Feelings	1.038	.058	18.044	***
Q12 <--- Managing Feelings	1.000			
Q30 <--- Independence	.741	.067	11.045	***
Q17 <--- Managing Behaviour	1.000			
Q16 <--- Friendships	1.167	.064	18.237	***
Q13 <--- Friendships	1.000			
Q35 <--- Att T&S	.820	.046	17.899	***
Q33 <--- Att T&S	.964	.043	22.586	***
Q23 <--- Att T&S	.903	.041	22.286	***
Q29 <--- Self-image	.907	.058	15.717	***
Q28 <--- Self-image	1.000			
Q27 <--- Managing Behaviour	.945	.030	31.177	***
Q22 <--- Managing Behaviour	-.524	.033	-15.998	***

Appendix 1

Standardized Regression Weights

	Estimate
Q24 <--- Managing Feelings	-.406
Q10 <--- Independence	.547
Q39 <--- Att T&S	.608
Q11 <--- Resilience	-.603
Q18 <--- Resilience	-.527
Q7 <--- Resilience	-.497
Q20 <--- Managing Behaviour	.642
Q15 <--- Managing Feelings	.689
Q12 <--- Managing Feelings	.593
Q30 <--- Independence	.355
Q17 <--- Managing Behaviour	.752
Q16 <--- Friendships	.764
Q13 <--- Friendships	.689
Q35 <--- Att T&S	.518
Q33 <--- Att T&S	.711
Q23 <--- Att T&S	.729
Q29 <--- Self-image	.553
Q28 <--- Self-image	.608
Q27 <--- Managing Behaviour	.808
Q22 <--- Managing Behaviour	-.400

Correlations: (Group number 1 - Default model)

			Estimate
Att T&S	<-->	Resilience	-.109
Att T&S	<-->	Friendships	.310
Independence	<-->	Resilience	.155
Managing Feelings	<-->	Independence	.715
Independence	<-->	Att T&S	.547
Independence	<-->	Friendships	.415
Resilience	<-->	Managing Behaviour	-.056
Managing Feelings	<-->	Managing Behaviour	.640
Friendships	<-->	Managing Behaviour	.249
Att T&S	<-->	Managing Behaviour	.705
Managing Feelings	<-->	Att T&S	.511
Managing Behaviour	<-->	Self-image	.436
Managing Feelings	<-->	Self-image	.511
Friendships	<-->	Self-image	.724
Att T&S	<-->	Self-image	.504
Independence	<-->	Self-image	.685
Resilience	<-->	Self-image	.345
Independence	<-->	Managing Behaviour	.798
Resilience	<-->	Friendships	.350
Managing Feelings	<-->	Resilience	.160
Managing Feelings	<-->	Friendships	.371
E18	<-->	e7	.201
E24	<-->	Resilience	-.241
E22	<-->	e24	.183

Appendix 1

Covariances: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P
Att T&S	<-->	Resilience	-.046	.015	-2.985	.003
Att T&S	<-->	Friendships	.152	.017	8.710	***
Independence	<-->	Resilience	.042	.015	2.744	.006
Managing Feelings	<-->	Independence	.244	.021	11.815	***
Independence	<-->	Att T&S	.199	.019	10.363	***
Independence	<-->	Friendships	.132	.016	8.248	***
Resilience	<-->	Managing Behaviour	-.026	.016	-1.600	.110
Managing Feelings	<-->	Managing Behaviour	.365	.025	14.655	***
Friendships	<-->	Managing Behaviour	.132	.018	7.459	***
Att T&S	<-->	Managing Behaviour	.430	.026	16.492	***
Managing Feelings	<-->	Att T&S	.270	.023	11.876	***
Managing Behaviour	<-->	Self-image	.209	.019	10.895	***
Managing Feelings	<-->	Self-image	.213	.021	10.365	***
Friendships	<-->	Self-image	.280	.020	13.797	***
Att T&S	<-->	Self-image	.224	.020	11.233	***
Independence	<-->	Self-image	.197	.017	11.762	***
Resilience	<-->	Self-image	.116	.018	6.561	***
Independence	<-->	Managing Behaviour	.314	.021	15.098	***
Resilience	<-->	Friendships	.129	.017	7.637	***
Managing Feelings	<-->	Resilience	.064	.017	3.771	***
Managing Feelings	<-->	Friendships	.171	.019	9.039	***
E18	<-->	e7	.195	.040	4.861	***
E24	<-->	Resilience	-.142	.023	-6.234	***
E22	<-->	e24	.186	.025	7.576	***

Variances: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P
Managing Feelings	.495	.043	11.484	***
Independence	.235	.031	7.487	***
Att T&S	.564	.043	13.081	***
Resilience	.320	.046	6.905	***
Friendships	.428	.033	13.089	***
Managing Behaviour	.659	.037	17.683	***
Self-image	.350	.033	10.762	***
E12	.911	.039	23.318	***
E15	.592	.033	18.026	***
E20	.682	.026	26.524	***
E22	.950	.032	29.663	***
E24	1.087	.038	28.332	***
E27	.312	.016	18.958	***
E10	.552	.031	17.659	***
E13	.472	.026	18.035	***
E16	.415	.032	13.011	***
E23	.406	.019	21.624	***
E39	.960	.037	26.092	***
E11	.900	.057	15.707	***
E18	.968	.050	19.435	***
E33	.512	.023	22.371	***
E35	1.034	.037	27.902	***
E7	.976	.049	20.081	***
E30	.894	.032	27.752	***
E28	.598	.029	20.328	***
E29	.652	.028	23.379	***
E17	.504	.022	22.553	***

Appendix 1

Residual Covariances (Group number 1 - Default model)

	Q29	Q28	Q17	Q30	Q7	Q35	Q33	Q18	Q11	Q39	Q23	Q16	Q13	Q10	Q27	Q24	Q22	Q20	Q15	Q12
Q29	.000																			
Q28	.000	.000																		
Q17	.031	.020	.000																	
Q30	.097	-.025	.031	.000																
Q7	.022	-.036	-.043	-.083	-.003															
Q35	.084	.070	-.031	-.053	-.079	.000														
Q33	-.047	.008	-.056	-.080	-.047	.029	.000													
Q18	-.024	-.024	-.083	-.155	-.003	-.072	-.012	-.003												
Q11	.033	.021	-.008	-.058	-.022	.007	.052	.013	-.005											
Q39	-.005	.002	.001	-.034	.005	-.017	.025	-.015	.062	.000										
Q23	-.031	.003	.026	-.055	-.056	-.016	.004	-.057	.027	-.030	.000									
Q16	.015	.001	.018	.023	-.012	.030	-.014	.025	-.003	-.013	-.009	.000								
Q13	-.033	.010	.061	.046	-.008	.053	.007	.007	-.008	-.038	.015	.000	.000							
Q10	.020	-.042	.019	.000	-.017	.055	.004	.020	.079	.050	.026	-.035	.011	.000						
Q27	.027	-.008	-.003	-.007	-.027	-.014	-.010	-.048	.067	.033	.014	-.012	.026	.008	.000					
Q24	.079	-.036	-.040	-.031	.078	.022	-.025	.012	-.052	-.052	-.077	.002	.013	-.023	-.030	.011				
Q22	.062	-.020	.015	.075	.092	.061	.002	.097	.034	-.009	-.066	.015	.028	-.024	-.013	.045	.006			
Q20	-.034	-.066	-.002	-.062	.040	-.007	-.028	.006	.078	-.011	.030	-.070	-.044	-.029	.009	-.018	-.006	.000		
Q15	-.015	.012	.037	-.036	.023	-.041	-.047	-.001	.042	.035	.015	-.025	-.003	-.001	-.034	.003	-.070	.025	.000	
Q12	-.020	.032	.002	-.050	-.016	.021	-.050	-.010	-.068	-.019	.021	.016	.040	.029	-.058	.030	.019	-.009	.020	.000

Standardized Residual Covariances (Group number 1 - Default model)

	Q29	Q28	Q17	Q30	Q7	Q35	Q33	Q18	Q11	Q39	Q23	Q16	Q13	Q10	Q27	Q24	Q22	Q20	Q15	Q12
Q29	.000																			
Q28	.000	.000																		
Q17	1.259	.812	.000																	
Q30	4.299	-1.100	1.230	.000																
Q7	.860	-1.395	-1.526	-3.150	-.067															
Q35	3.152	2.612	-1.015	-1.929	-2.557	.000														
Q33	-2.037	.332	-2.097	-3.345	-1.761	.971	.000													
Q18	-.940	-.918	-2.881	-5.765	-.093	-2.275	-.426	-.076												
Q11	1.249	.770	-.265	-2.091	-.693	.210	1.873	.386	-.099											
Q39	-.187	.061	.030	-1.173	.141	-.471	.795	-.463	1.826	.000										
Q23	-1.459	.160	1.072	-2.537	-2.298	-.580	.180	-2.321	1.066	-1.030	.000									
Q16	.643	.033	.719	.993	-.446	1.102	-.613	.952	-.095	-.459	-.429	.000								
Q13	-1.507	.448	2.581	2.092	-.303	2.034	.335	.269	-.297	-1.418	.742	.000	.000							
Q10	.977	-2.067	.833	.000	-.736	2.263	.193	.862	3.274	1.951	1.321	-1.690	.567	.000						
Q27	1.248	-.385	-.095	-.310	-1.096	-.526	-.402	-1.921	2.586	1.172	.633	-.560	1.228	.403	.000					
Q24	3.112	-1.409	-1.383	-1.181	2.580	.702	-.919	.402	-1.658	-1.589	-3.131	.089	.525	-.963	-1.197	.269				
Q22	2.626	-.842	.559	3.030	3.318	2.096	.064	3.427	1.159	-.305	-2.860	.627	1.203	-1.104	-.517	1.571	.151			
Q20	-1.390	-2.691	-.075	-2.458	1.405	-.242	-1.064	.217	2.667	-.339	1.238	-2.804	-1.848	-1.293	.332	-.613	-.208	.000		
Q15	-.633	.484	1.343	-1.442	.832	-1.412	-1.858	-.030	1.433	1.147	.656	-1.003	-.110	-.048	-1.396	.097	-2.668	.912	.000	
Q12	-.754	1.179	.072	-1.807	-.509	.650	-1.783	-.304	-2.094	-.543	.805	.568	1.532	1.163	-2.133	.954	.637	-.307	.629	.000

Calculations for Raykov's Scale Reliability

Self-image

Q28 I am happy being me.

Q29 I am good at some things.

ρ_Y = (1+ 0.907)^2/ (1+0.907)^2 + (0.598 + 0.652) = 3.636649/4.886649 = 0.744

Managing Feelings

Q12 I calm down quickly after I have got angry or upset.

Q15 I am usually calm.

Q24 I sulk or argue when I am told off.

ρ_Y = (1+ 1.038 + 0.658)^2/ (1+ 1.038 + 0.658)^2 + (0.911 + 0.592 + 1.087 + 2x0.183 + 2x0.241) = 7.268416/10.706416= 0.679

Managing Behaviour

Q17 I find it easy to pay attention in class.

Q20 I work quietly in my class.

Q22 I sometimes leave the room without permission.

Q27 I listen well in class.

ρ_Y = (1+ 0.853 + 0.524 + 0.945)^2/ (1+ 0.853 + 0.524 + 0.945)^2 + (0.504 + 0.682 + 0.950 + 0.312 + 2x0.183) = 11.035684/13.849684= 0.797

Independence

Q10 If I find something difficult I still try to do it.

Q30 I can work without my teacher's help.

ρ_Y = (1+ 0.741)^2/(1+ 0.741)^2 + (0.552 + 0.894) = 3.031081/4.477081 = 0.677

Appendix 1

Resilience

Q7 I get upset if I don't do something well.

Q11 I'm easily hurt by what others say about me.

Q18 I worry about the things I can't do well.

ρ_Y = (1+ 1.080 + 1.269)^2/ (1+ 1.080 + 1.269)^2 + (0.976 + 0.900 + 0.968 + 2x0.201 + 2x0.241) = 11.215801/14.943801= 0.751

Friendships

Q13 Other children let me play with them.

Q16 I have lots of friends at school.

ρ_Y = (1+ 1.167)^2/(1+ 1.167)^2 + (0.472 + 0.415) = 4.695889/5.582889 = 0.841

Attitudes to teachers and school

Q23 I get on well with my teachers.

Q33 Our teachers are fair in the way they treat us.

Q35 I can talk to my teacher about anything.

Q39 I like coming to school.

ρ_Y = (0.903 + 0.964 + 0.820 + 1)^2/ (0.903 + 0.964 + 0.820 + 1)^2 + (0.406 + 0.512 + 1.034 + 0960) = 13.593969/16.505969 = 0.824

These results suggest that the reliability of the scales associated with the seven identified dimensions in the final CFA model are approaching or above the generally accepted cut off value of 0.7. These values are in line with the Cronbach's alpha values for the sub-scales reported in the Teacher and Parent Emotional Literacy Checkilsists (Faupel 2003) utilised widely in evaluating the impact of social and emotional learning programmes including national evaluations of SEAL (Humphrey et al, 2010), and in the study of Family SEAL in this thesis.

Appendix 3

Family SEAL evaluation One way ANOVA analysis of NFER Parent and Teacher Survey results

‘Concern’ students identified by the class teacher as causing social and emotional concerns

Oneway ANOVA for KS2 "concern" students

Descriptives

						95% Confidence Interval for Mean			
		N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
Self-awareness	Parent rating	14	.618	.1049	.0280	.557	.678	.5	.9
	Teacher rating	14	.665	.1265	.0338	.592	.738	.4	.9
	Total	28	.642	.1165	.0220	.596	.687	.4	.9
Self-regulation	Parent rating	14	.629	.1888	.0505	.520	.738	.3	.9
	Teacher rating	14	.598	.2411	.0644	.459	.737	.3	.9
	Total	28	.613	.2131	.0403	.531	.696	.3	.9
Motivation	Parent rating	14	.5750	.11561	.03090	.5082	.6418	.35	.75
	Teacher rating	14	.6116	.13013	.03478	.5365	.6867	.44	.88
	Total	28	.5933	.12221	.02310	.5459	.6407	.35	.88
Empathy	Parent rating	14	.807	.1651	.0441	.712	.902	.5	1.0
	Teacher rating	14	.643	.2305	.0616	.510	.776	.3	.9
	Total	28	.725	.2138	.0404	.642	.808	.3	1.0
Social skills	Parent rating	14	.8536	.14340	.03832	.7708	.9364	.60	1.00
	Teacher rating	14	.7411	.15083	.04031	.6540	.8282	.38	.94
	Total	28	.7973	.15536	.02936	.7371	.8576	.38	1.00
Overall percentage	Parent rating	14	.696	.0998	.0267	.639	.754	.5	.8
	Teacher rating	14	.652	.1180	.0315	.584	.720	.5	.8
	Total	28	.674	.1096	.0207	.632	.717	.5	.8

Appendix 1

The mean scores are decimal equivalents of percentage scores, so for example, the mean parent rating for self-awareness of 0.618 corresponds to a mean score of 61.8%

Appendix 1

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Self -awareness	Between Groups	.016	1	.016	1.161	.291
	Within Groups	.351	26	.013		
	Total	.367	27			
Self -regulation	Between Groups	.006	1	.006	.138	.714
	Within Groups	1.219	26	.047		
	Total	1.226	27			
Motiv ation	Between Groups	.009	1	.009	.619	.438
	Within Groups	.394	26	.015		
	Total	.403	27			
Empathy	Between Groups	.189	1	.189	4.700	.040
	Within Groups	1.045	26	.040		
	Total	1.234	27			
Social skills	Between Groups	.089	1	.089	4.091	.054
	Within Groups	.563	26	.022		
	Total	.652	27			
Ov erall percentage	Between Groups	.014	1	.014	1.169	.290
	Within Groups	.310	26	.012		
	Total	.324	27			

Appendix 1

Oneway ANOVA for KS2 "control" students

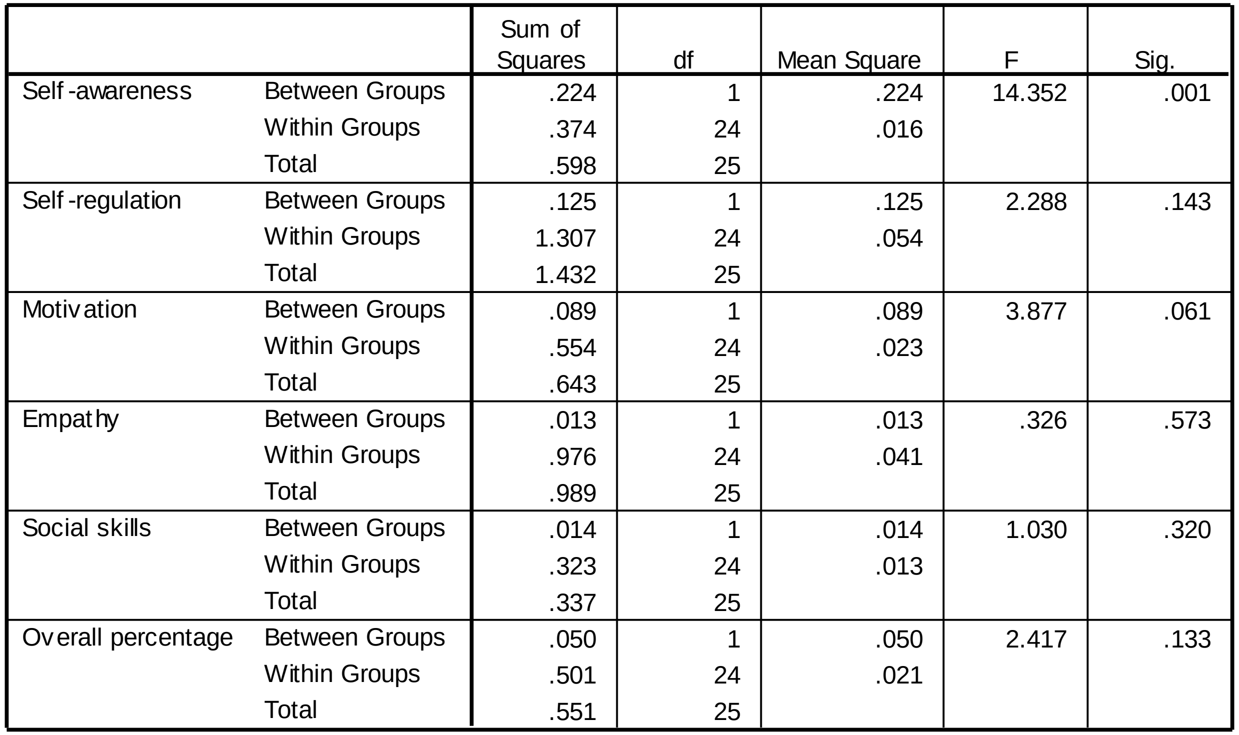
Descriptives

						95% Confidence Interval for Mean			
		N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
Self-awareness	Parent rating	13	.627	.0881	.0244	.574	.680	.5	.8
	Teacher rating	13	.813	.1531	.0425	.720	.905	.4	1.0
	Total	26	.720	.1547	.0303	.657	.782	.4	1.0
Self-regulation	Parent rating	13	.650	.1568	.0435	.555	.745	.3	.8
	Teacher rating	13	.788	.2904	.0805	.613	.964	.3	1.0
	Total	26	.719	.2393	.0469	.623	.816	.3	1.0
Motivation	Parent rating	13	.7192	.14511	.04025	.6315	.8069	.40	.95
	Teacher rating	13	.8365	.15840	.04393	.7408	.9323	.44	1.00
	Total	26	.7779	.16040	.03146	.7131	.8427	.40	1.00
Empathy	Parent rating	13	.758	.0886	.0246	.704	.811	.6	.9
	Teacher rating	13	.803	.2710	.0752	.639	.967	.3	1.0
	Total	26	.780	.1989	.0390	.700	.861	.3	1.0
Social skills	Parent rating	13	.9308	.05965	.01654	.8947	.9668	.80	1.00
	Teacher rating	13	.8846	.15277	.04237	.7923	.9769	.50	1.00
	Total	26	.9077	.11603	.02276	.8608	.9546	.50	1.00
Overall percentage	Parent rating	13	.737	.0767	.0213	.691	.783	.5	.8
	Teacher rating	13	.825	.1893	.0525	.711	.939	.4	1.0
	Total	26	.781	.1485	.0291	.721	.841	.4	1.0

The mean scores are decimal equivalents of percentage scores, so for example, the mean parent rating for self-awareness of 0.627 corresponds to a mean score of 62.7%.

Appendix 1

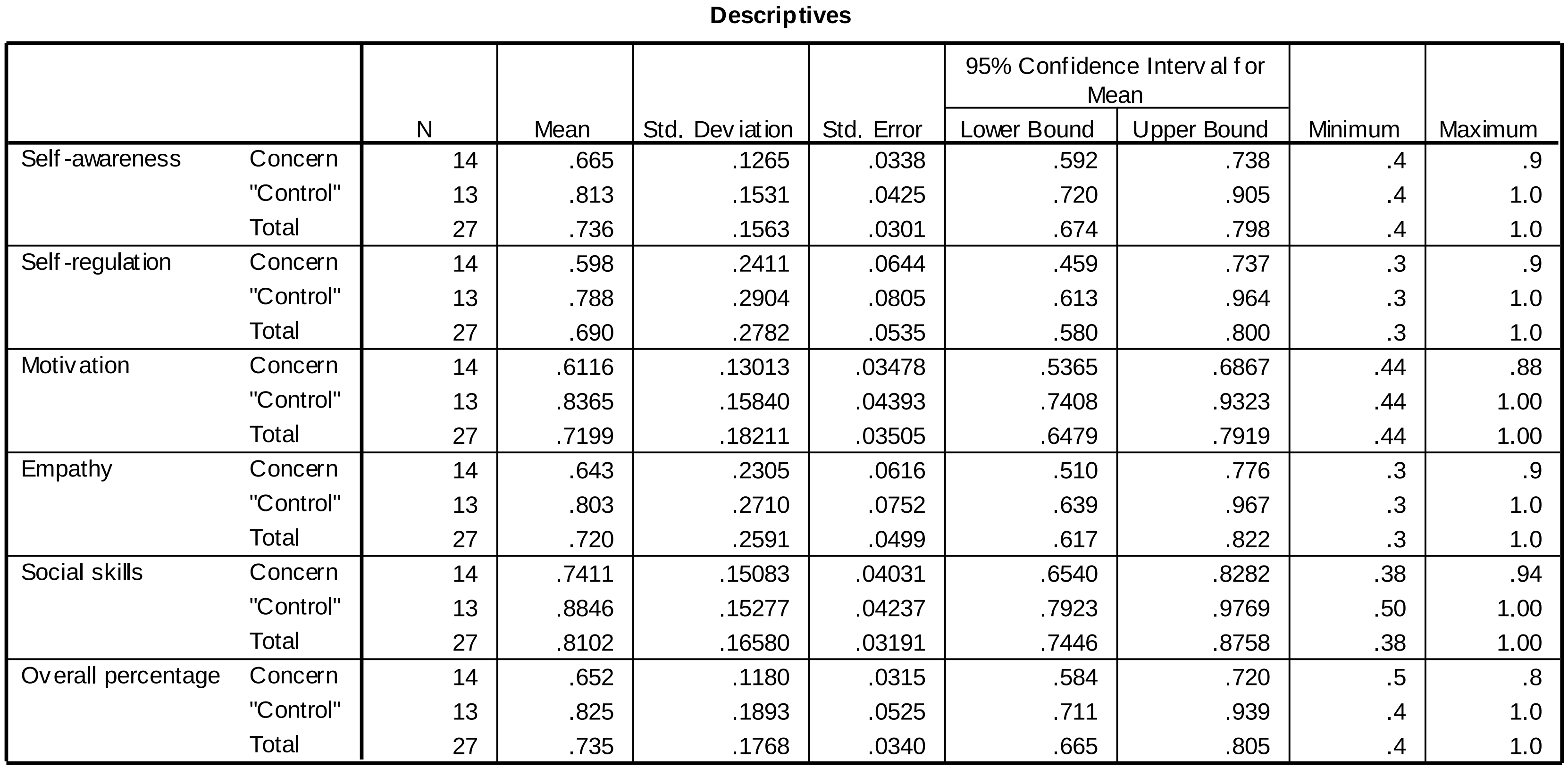
ANOVA



		Sum of Squares	df	Mean Square	F	Sig.
Self-awareness	Between Groups	.224	1	.224	14.352	.001
	Within Groups	.374	24	.016		
	Total	.598	25			
Self-regulation	Between Groups	.125	1	.125	2.288	.143
	Within Groups	1.307	24	.054		
	Total	1.432	25			
Motivation	Between Groups	.089	1	.089	3.877	.061
	Within Groups	.554	24	.023		
	Total	.643	25			
Empathy	Between Groups	.013	1	.013	.326	.573
	Within Groups	.976	24	.041		
	Total	.989	25			
Social skills	Between Groups	.014	1	.014	1.030	.320
	Within Groups	.323	24	.013		
	Total	.337	25			
Overall percentage	Between Groups	.050	1	.050	2.417	.133
	Within Groups	.501	24	.021		
	Total	.551	25			

Appendix 1

Oneway ANOVA for KS2 teacher responses



Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Self-awareness	Concern	14	.665	.1265	.0338	.592	.738	.4	.9
	"Control"	13	.813	.1531	.0425	.720	.905	.4	1.0
	Total	27	.736	.1563	.0301	.674	.798	.4	1.0
Self-regulation	Concern	14	.598	.2411	.0644	.459	.737	.3	.9
	"Control"	13	.788	.2904	.0805	.613	.964	.3	1.0
	Total	27	.690	.2782	.0535	.580	.800	.3	1.0
Motivation	Concern	14	.6116	.13013	.03478	.5365	.6867	.44	.88
	"Control"	13	.8365	.15840	.04393	.7408	.9323	.44	1.00
	Total	27	.7199	.18211	.03505	.6479	.7919	.44	1.00
Empathy	Concern	14	.643	.2305	.0616	.510	.776	.3	.9
	"Control"	13	.803	.2710	.0752	.639	.967	.3	1.0
	Total	27	.720	.2591	.0499	.617	.822	.3	1.0
Social skills	Concern	14	.7411	.15083	.04031	.6540	.8282	.38	.94
	"Control"	13	.8846	.15277	.04237	.7923	.9769	.50	1.00
	Total	27	.8102	.16580	.03191	.7446	.8758	.38	1.00
Overall percentage	Concern	14	.652	.1180	.0315	.584	.720	.5	.8
	"Control"	13	.825	.1893	.0525	.711	.939	.4	1.0
	Total	27	.735	.1768	.0340	.665	.805	.4	1.0

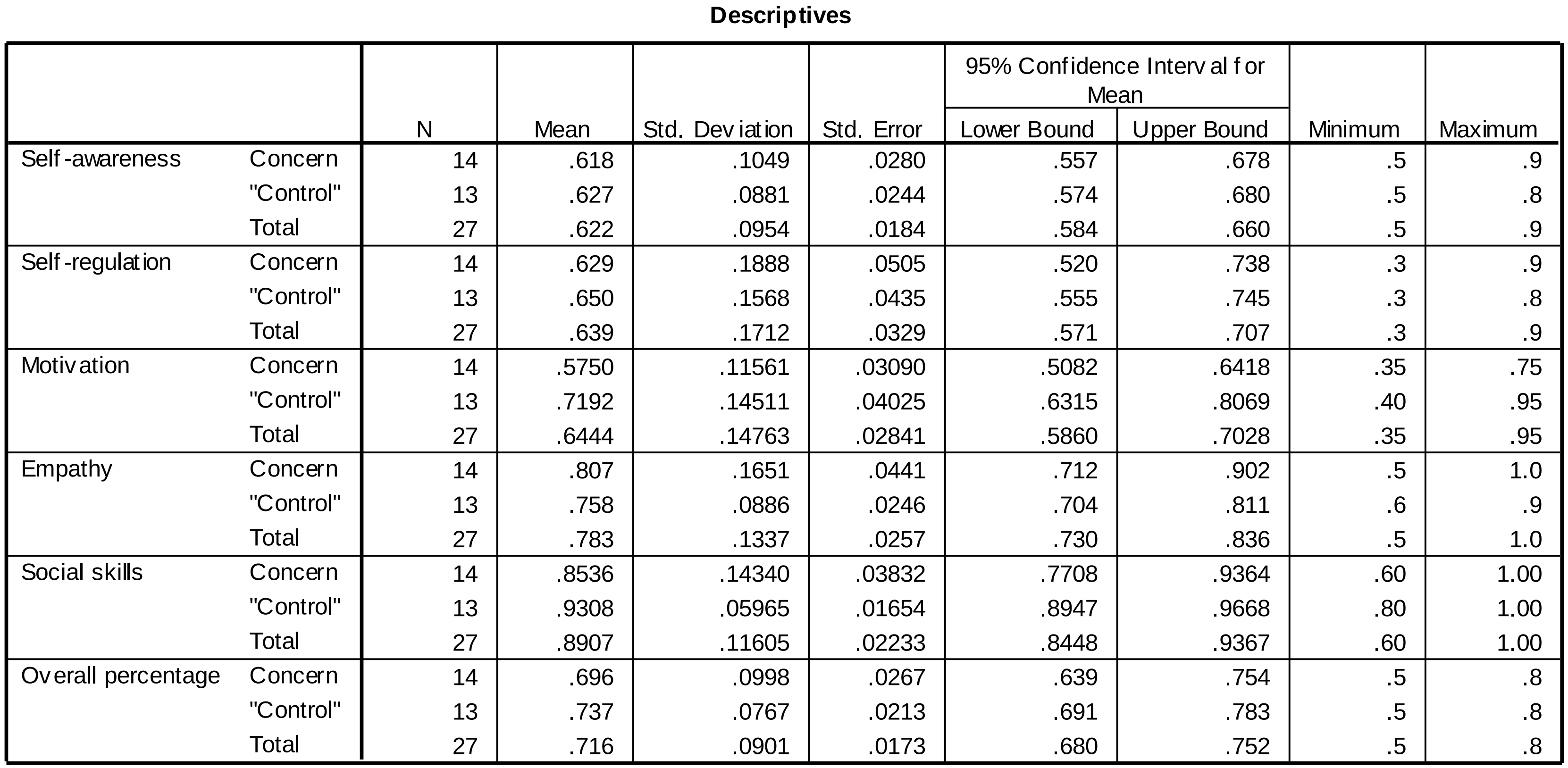
Appendix 1

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Self -awareness	Between Groups	.146	1	.146	7.478	.011
	Within Groups	.489	25	.020		
	Total	.635	26			
Self -regulation	Between Groups	.244	1	.244	3.451	.075
	Within Groups	1.768	25	.071		
	Total	2.012	26			
Motiv ation	Between Groups	.341	1	.341	16.358	.000
	Within Groups	.521	25	.021		
	Total	.862	26			
Empat hy	Between Groups	.173	1	.173	2.744	.110
	Within Groups	1.572	25	.063		
	Total	1.745	26			
Social skills	Between Groups	.139	1	.139	6.030	.021
	Within Groups	.576	25	.023		
	Total	.715	26			
Ov erall percentage	Between Groups	.202	1	.202	8.277	.008
	Within Groups	.611	25	.024		
	Total	.813	26			

Appendix 1

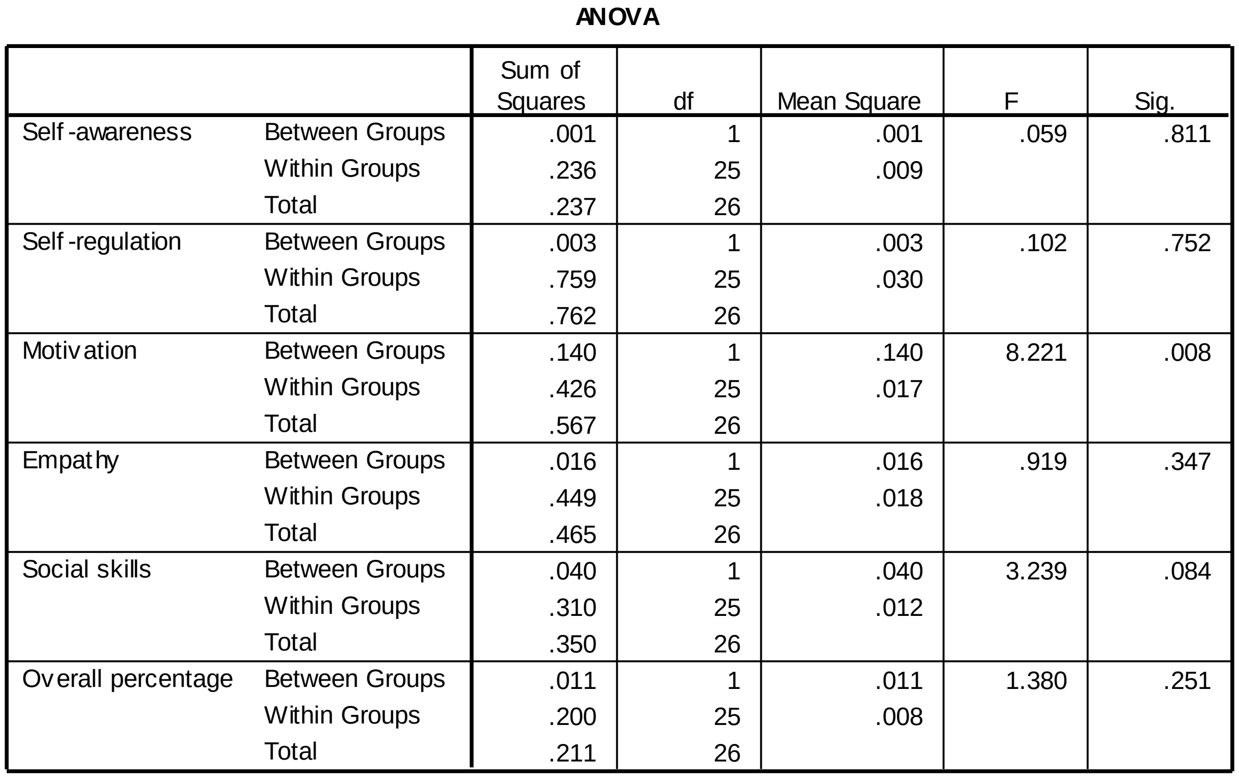
Oneway ANOVA for KS2 parent responses



Descriptives

		N	Mean	Std. Dev iation	Std. Error	95% Conf idence Interv al f or Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Self -awareness	Concern	14	.618	.1049	.0280	.557	.678	.5	.9
	"Control"	13	.627	.0881	.0244	.574	.680	.5	.8
	Total	27	.622	.0954	.0184	.584	.660	.5	.9
Self -regulation	Concern	14	.629	.1888	.0505	.520	.738	.3	.9
	"Control"	13	.650	.1568	.0435	.555	.745	.3	.8
	Total	27	.639	.1712	.0329	.571	.707	.3	.9
Motiv ation	Concern	14	.5750	.11561	.03090	.5082	.6418	.35	.75
	"Control"	13	.7192	.14511	.04025	.6315	.8069	.40	.95
	Total	27	.6444	.14763	.02841	.5860	.7028	.35	.95
Empathy	Concern	14	.807	.1651	.0441	.712	.902	.5	1.0
	"Control"	13	.758	.0886	.0246	.704	.811	.6	.9
	Total	27	.783	.1337	.0257	.730	.836	.5	1.0
Social skills	Concern	14	.8536	.14340	.03832	.7708	.9364	.60	1.00
	"Control"	13	.9308	.05965	.01654	.8947	.9668	.80	1.00
	Total	27	.8907	.11605	.02233	.8448	.9367	.60	1.00
Ov erall percentage	Concern	14	.696	.0998	.0267	.639	.754	.5	.8
	"Control"	13	.737	.0767	.0213	.691	.783	.5	.8
	Total	27	.716	.0901	.0173	.680	.752	.5	.8

Appendix 1



ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Self-awareness	Between Groups	.001	1	.001	.059	.811
	Within Groups	.236	25	.009		
	Total	.237	26			
Self-regulation	Between Groups	.003	1	.003	.102	.752
	Within Groups	.759	25	.030		
	Total	.762	26			
Motivation	Between Groups	.140	1	.140	8.221	.008
	Within Groups	.426	25	.017		
	Total	.567	26			
Empathy	Between Groups	.016	1	.016	.919	.347
	Within Groups	.449	25	.018		
	Total	.465	26			
Social skills	Between Groups	.040	1	.040	3.239	.084
	Within Groups	.310	25	.012		
	Total	.350	26			
Overall percentage	Between Groups	.011	1	.011	1.380	.251
	Within Groups	.200	25	.008		
	Total	.211	26			

List of References

Arbuckle, J. L. (2006) *Amos 7.0 User's Guide* (Chicago, SPSS Inc.).

Baumeister, R., Smart, L., and Boden, J. (1996) Relation of threatened egotism to violence and aggression: The dark side of high self-esteem, *Psychological Review*, 103, 5-33.

BBC (2004) “Results changed to boost learning”, *BBC News Online*, accessed at http://news.bbc.co.uk/1/hi/education/3761248.stm on 5th February 2008.

Bell B. A., Morgan G. B., Kromrey J. D., Ferron J. M. (2010) The impact of small cluster size on multilevel models: a Monte Carlo examination of two-level models with binary and continuous predictors, in Xe H., Elliott M. R., (editors) *JSM Proceedings, Survey Research Methods Section*, Vancouver, BC: American Statistical Association, 4057–4067.

Borghans, L., Duckworth, A. L., Heckman, J. J. and ter Weel, B. (2008) The economics and psychology of personality traits, *Journal of Human Resources*, 43 (4), 972-1059.

Borghans, L., Golsteyn, B. H. H., Heckman, J. J. and Humphries, J. E. (2010) IQ, achievement, and personality. Unpublished manuscript, University of Maastricht and University of Chicago (revised from the 2009 version).

Bradshaw, J., Hoelscher, P. & Richardson, D. (2007) An Index of Child Well-being in the European Union, *Social Indicators Research*, 80, 133-177.

Brookover, W., Beady, C., Flood, P., Schweitzer, J. and Wisenbaker, J. (1979) *Schools, social systems and student achievement - schools can make a difference*. New York: Praeger.

Brown, T. A. (2006) *Confirmatory Factor Analysis for Applied Research*, New York, The Guildford Press.

Browne, M. W. (1984) Asymptotically distribution-free methods for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.

Burgess, S., McConnell, B., Propper, C. and Wilson, D. (2004) Girls Rock, Boys Roll: An Analysis of the Age 14–16 Gender Gap in English Schools, *Scottish Journal of Political Economy*, 51(2), 209-229.

Burrell, G. and Morgan, G. (1979) Sociological Paradigms and Organisational Analysis. London: Heinemann Educational Books, cited in Cohen, L., Manion, L. and Morrison, K. (2000) *Research Methods in Education*, London: RoutledgeFalmer.

Cohen, L., Manion, L. and Morrison, K. (2000) *Research Methods in Education*, London: RoutledgeFalmer.

Costa, P. T. and McCrae, R.R. (1988) From Catalog to Classification: Murray's Needs and the Five-Factor Model, *Journal of Personality and Social Psychology*, 55(2), 258–65.

Craig, C. (2007) The potential dangers of a systematic, explicit approach to teaching social and emotional skills (SEAL), (Glasgow, Centre for Confidence and Well-being).

Creemers, B.P.M. and Kyriakides, L. (2006) Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model, *School Effectiveness and School Improvement*, 17(3), 347-366.

Creemers, B.P.M. and Kyriakides, L. (2008) *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*, London: Routledge.

Critchlow, J. and Coe, R. (2003) Serious flaws arising from the use of the median in calculating value added measures for School Performance Tables in England, *Paper presented to the 29th International Association for Educational Assessment (IAEA) Annual Conference*.

Crowley, S. L. and Fan, X. (1977) Structural equation modeling: Basic concepts and applications in personality research, *Journal of Personality Assessment*, 68, 508–531.

CSFC (2010) *School Accountability First Report of Session 2009–10: Volume I Report, together with formal minutes*, House of Commons Children, Schools and Families Committee, London: The Stationery Office Limited.

DCSF (No date) *Social and Emotional Aspects of Learning (SEAL)*. Available online at: http://www.dfes.gov.uk/ibis/department_policy/seal01.cfm (accessed April 2008).

DCSF (2004) *Lincoln Birchwood Junior School*. Available online at: http://www.dcsf.gov.uk/cgi-bin/performancetables/dfepx1_04.pl?School=9252245&Mode=Z&Type= (accessed April 2008).

DCSF (2007a) *ISP and SEAL at Birchwood Junior School, Lincoln*. Available online at: http://www.standards.dfes.gov.uk/primary/casestudies/isp/seal/birchwood/ (accessed May 2007.

DCSF (2007b) *Case Study Raising Attainment Plans*. Available online at: http://www.standards.dfes.gov.uk/primary/casestudies/isp/seal/birchwood/birchwood_app_008907.pdf (accessed May 2007).

DCSF (2008) *Department for Children, Schools and Families: Achievement and Attainment Tables*. Available online at: http://www.dcsf.gov.uk/cgi-bin/performancetables/group_07.pl?Mode=Z&Type=LA&No=925&Base=p&Phase=p&F=101&L=150&Year=07 (accessed April 2008).

DCSF (2008) *Targeted Mental Health in Schools Project - Using the evidence to inform your approach: a practical guide for headteachers and commissioners*. Nottingham, Department for Children, Schools and Families.

DCSF (2009) *Leading on Intervention: Strengthening the quality of everyday inclusive teaching: The waves of intervention model*. [Online: accessed at http://nationalstrategies.standards.dcsf.gov.uk/node/41795 on 27th August 2009].

DCSF (2009b) *School Report Card: Prospectus*, Department for Children Schools and Families, Nottingham: DCSF Publications.

de Bilde, J., Van Damme, J., Lamote, C. and De Fraine, B. (2013) Can alternative education increase children's early school engagement? A longitudinal study from kindergarten to third grade, *School Effectiveness and School Improvement*, 24(2), 212-233.

De Fraine, B., Van Landeghem, G., Van Damme, J. and Onghena, P. (2005) An analysis of Well-Being in Secondary Schools with Multilevel Growth Curve models and Multilevel Multivariate Models, *Quality and Quantity*, 39, 297-316.

Desforges, C. and Abouchaar, A. (2003) *The Impact of Parental Involvement, Parental Support and Family Education on Pupil Achievements and Adjustment: A Literature Review.* Nottingham, Department for Education and Skills.

DfE (2013) Secondary School Accountability Consultation, Department for Education [Online] accessed at https://www.gov.uk/government/consultations/secondary-school-accountability-consultation, August 2013.

DfES (2004) *Pupil Achievement Tracker Quick Tour*. Available online at: http://www.standards.dfes.gov.uk/performance/word/QuickStart.doc?version=1 (accessed 16th December 2006).

DfES/Ofsted (2004) *A New Relationship with Schools*, Department for Education and Skills / Office for Standards in Education, [online] accessed at http://www.ofsted.gov.uk/assets/3666.pdf, accessed on 14 April 2007.

DfES (2005) *Primary National Strategy Excellence and Enjoyment: social and emotional aspects of learning guidance*, London, Department for Education and Skills.

DfES (2006) *Primary National Strategy Excellence and Enjoyment: social and emotional aspects of learning Family SEAL*. Department for Education and Skills.

DfES (2006) *Achievement and Attainment Tables: LA Conference 2006*. Available online at: http://www.standards.dfes.gov.uk/performance/powerpoint/presentationLAv03.ppt?version=1 (accessed 3rd December 2006).

DfES (2006a) *GUIDANCE FOR LOCAL AUTHORITIES ON TARGET SETTING Part 1: LA Targets for Key Stages 2, 3, 4, Looked after Children, Minority Ethnic Groups, Attendance, Early Years' Outcomes*, [online] available at http://www.standards.dfes.gov.uk/ts/docs/guide1.doc , accessed 20 March 2007.

DfES (2006b) National Curriculum Assessments at Key Stage 2, and Key Stage 1 to Key Stage 2 Value Added Measures for England 2004/2005 (Final), DfES Statistical First Release, [online] available at http://www.dfes.gov.uk/rsgateway/DB/SFR/s000660/Addition2.xls , accessed 20 March 2007.

DfES (2006c) *Secondary Schools (GCSE and equivalent) Achievement and Attainment Tables 2006: More information*, [online] available at http://www.dfes.gov.uk/performancetables/schools_06/s8.shtml , accessed 13 April 2007.

DfES (2007a) Secondary National Strategy Social and Emotional Aspects of Learning, Department for Education and Skills).

DfES (2007b) Social and Emotional Aspects of Learning for secondary schools (SEAL): Introductory booklet, Department for Education and Skills).

DfES (2007a) *Intensifying Support Programme (ISP) and Social and Emotional Aspects of Learning (SEAL) Birchwood Junior School – April 2006.* [Online: accessed at http://nationalstrategies.standards.dcsf.gov.uk/node/88432?uc=force_uj on 26th August 2009].

DfES (2007b) *Secondary National Strategy Social and Emotional Aspects of Learning*. London, Department for Education and Skills.

Downey, C. and Kelly, A. (2007) Are value-added scores getting the measure of school performance in the UK?, *International Congress for School Effectiveness and Improvement (ICSEI)* (Potorož, Slovenia).

Downey C, Kelly A and Brown A (2008) Evaluating a programme to develop social and emotional skills in primary school students, presented at *The International Congress for School Effectiveness and Improvement (ICSEI)*, Auckland, NZ, 7th Jan 2008.

Edmunds, L. and Stewart-Brown, S. (2003) Assessing Emotional and Social Competence in Primary School and Early Years Settings: A Review of Approaches, Issues and Instruments, *Sure Start Evidence and Research Series*, Annesley, Nottinghamshire, DfES Sure Start.

Elchardus, M., Kavadias, D. and Siongers, J. (1998) Hebben scholen een invloed op de warden van jongeren? Een empirisch onderzoek naar de doeltreffendheid van waarde vorming in her secundair onderwijs (Do schools influence the values of the young? An empirical examination of the effectiveness of the teaching of values in lowewr seciondary education). Brussel, Belgium: VUB, Van Landeghem, G., Van Damme, J., Opdenakker, M-C., De Fraine, B. and Onghena, P. (2002) The Effect of Schools and Classes on Noncognitive Outcomes, *School Effectiveness and School Improvement*, 13(4), 429-451.

Faupel, A. (2003) *Emotional Literacy: Assessment and Intervention - Ages 7 to 11*, London, nferNelson.

FFT (2004a) *FFT Data Analysis Project – Value Added Development – 0304*, presentation given at FFT 2004 Regional Meetings, [online] downloaded from FFT secure server, accessed 15 March 2007.

FFT (2004b) *FFT Data Analysis Project – Annual Review – 0304*, presentation given at FFT 2004 Regional Meetings, [online] downloaded from FFT secure server, accessed 15 March 2007.

FFT (2005a) *Using ACORN data in Value-Added Analyses*, guidance paper, The Fischer Family Trust, [online] downloaded from FFT secure server, accessed 15 March 2007.

FFT (2005b) FFT Technical Information Brief: KS2-KS4 contextualised value added measures, guidance paper, The Fischer Family Trust, [online] downloaded from FFT secure server, accessed 15 March 2007.

FFT (2005c) Regional Meetings 2005 – VA methodology, presentation given at FFT 2005 Regional Meetings, [online] downloaded from FFT secure server, accessed 15 March 2007.

FFT (2005d) *Fischer Family Trust: Supplement to the PANDA (KS2)*, the Fischer Family Trust, [online] available at http://www.fischertrust.org/assets/PANDA/Example_FFT_EvalBook_KS2.PDF , accessed on 13 April 2007.

FFT (2005e) *Fischer Family Trust: Supplement to the PANDA (KS34)*, the Fischer Family Trust, [online] available at http://www.fischertrust.org/assets/PANDA/Example_FFT_EvalBook_KS34.PDF , accessed on 13 April 2007.

FFT (2006a) Secondary Training Materials Handout 4 – Anon Sch Est-Act, The Fischer Family Trust, [online] downloaded from FFT secure server, accessed 15 March 2007. Similar report available at http://www.fischertrust.org/assets/perfdata/ExampleAnalyses/England/KS3EstAct_AY2005-06_SCH.PDF

FFT (2006b) *Summary: Trends in Estimates (England) Key Stage 2, Key Stage 3, Key Stage 4*, guidance document published by the Fischer Family Trust, [online] downloaded from FFT secure server, accessed 18 March 2007.

FFT (2006c) *DfES Guidance and FFT Estimates*, guidance document published by the Fischer Family Trust, obtained via personal communication on 22 March 2007.

FFT (2007) [online] available at http://www.fischertrust.org/performance.htm, accessed on 16 March 2007.

FFT (2007a) *Summary Secondary Training & Support Materials*, the Fischer Family Trust, [online] available at http://www.fischertrust.org/assets/perfdata/Training/Secondary/Overview_Secondary_14032006_NoNotes.ppt.pdf, accessed on 11 April 2007.

FFT (2007b) *Analyses to Support Self-Evaluation (Example – KS2)*, the Fischer Family Trust, [online] available at http://www.fischertrust.org/assets/perfdata/ExampleAnalyses/England/Self_Evaluation/Example_Junior_FFT_EvalBook_KS2.PDF , accessed on 13 April 2007.

FFT (2007c) *Analyses to Support Self-Evaluation (Example – KS34)*, the Fischer Family Trust, [online] available at http://www.fischertrust.org/assets/perfdata/ExampleAnalyses/England/Self_Evaluation/Example_Secondary_FFT_EvalBook_KS34.pdf , accessed on 13 April 2007.

Fitz-Gibbon, C. T. (1991) Multilevel Modelling in an Indicator System, in: S. W. Raudenbush & J. D. Willms (Eds) *School, Classrooms and Pupils: International Studies of Schooling from a Multilevel Perspective* (San Diego, Academic Press).

Fitz-Gibbon, C. T. (1997) The Value Added National Project Final Report: Feasibility Studies for a National System of Value-Added Indicators, (London, School Curriculum and Assessment Authority).

Forsetlunda, L. , Chalmers, I. & Bjørndala, A. (2007) When Was Random Allocation First Used To Generate Comparison Groups In Experiments To Assess The Effects Of Social Interventions? *Economics of Innovation and New Technology*, 16(5), 371-384.

Gadeyne, E., Ghesquire, P. and Onghena, P. (2006) Psychosocial educational effectiveness criteria and their relation to teaching in primary education, *School Effectiveness and School Improvement*, 17(1), 63-85.

Gardner, H. (1983) *Frames of mind: the theory of multiple intelligences* (New York, Basic Books).

Gilbert, C. (2006) 2020 Vision - Report of the Teaching and Learning in 2020 Review Group, (Nottingham, DfES Publications).

GNN (2003) *(Government News Network) Value added results show more rounded picture of primary schools' progress - Miliband*. Available online at: http://www.gnn.gov.uk/content/detail.asp?NavigatedFromSearch=True&NewsAreaID=2&ReleaseID=101891 (accessed 06/09/2007 2007).

Goldacre, B. (2010) A 'shoot-out' between methods won't help us teach more children to read: Schools need large, robust randomised trials to help them decide which teaching methods to use, *The Guardian*, Saturday 31st July 2010.

Goldstein, H. (1997) *Value added data for schools: a commentary on a paper from SCAA*. Available online at: http://www.cmm.bris.ac.uk/team/HG_Personal/scaavadd.html (accessed 27/08/2007.

Goldstein, H. (2001) Using Pupil Performance Data for Judging Schools and Teachers: scope and limitations, *British Educational Research Journal*, 27(4), 433-442.

Goldstein, H. (2003) *Multilevel Statistical Models* (London, Arnold).

Goldstein, H. (2007) Evidence and education policy - some reflections and allegations, *Royal Society of Statistics (RSS)* (York.

Goldstein, H. and Cuttance, P. (1988) A note on national assessment and school comparisons, *Journal of Education Policy*, 3, 197–202.

Goldstein, H. and Thomas, S. (1995) School effectiveness and "value-added" analysis, *Forum*, 37(2), 36–38.

Goleman, D. (1996) *Emotional Intelligence: Why it can matter more than IQ* (London, Bloomsbury Publishing).

Goodman, R. (1997) The Strengths & Difficulties Questionnaire: A Research Note, *Journal of Child Psychology & Psychiatry*, 38, 581-586.

Gorard, S. (2006) Value-added is of little value, *Journal of Education Policy*, 21(2), 235-243.

Gray, J., Jesson, D. & Jones, B. (1986) The search for a fairer way of comparing schools' examination results, *Research Papers in Education*, 1(2), 91–122.

Gray, J., Jesson, D., Goldstein, H., Hedger, K. and Rasbash, J. (1995) A multi-level analysis of school improvement: Changes in schools' performance over time, *School Effectiveness and School Improvement*, 6, 97 – 114.

Gray, J., Goldstein, H. and Jesson, D. (1996) Changes and improvements in schools' effectiveness: Trends over time, *Research Papers in Education*, 11(1), 35 – 51.

Gray, J., Goldstein, H. and Thomas, S. (2001) Predicting the future: The role of past performance in determining trends in institutional effectiveness at A-level. , *British Educational Research Journal*, 27(4), 1 – 15.

Gray, J. (2004) School effectiveness and the 'other outcomes' of secondary schooling: a reassessment of three decades of British research, *Improving Schools*, 7(2), 185-198.

Grisay, A. (1996) Evolution des acquis cognitifs et socio-affectifs des élèves au cours des années de collège (Evolution of cognitive and affective development of students in lower secondary education). Liege, Belgique: Universite de Liege, cited in Van Landeghem, G., Van Damme, J., Opdenakker, M-C., De Fraine, B. and Onghena, P. (2002) The Effect of Schools and Classes on Noncognitive Outcomes, *School Effectiveness and School Improvement*, 13(4), 429-451.

Hague, D (2005) "Mike Fischer, Serial Entrepreneur" [online] *Oxford Science Enterprise Centre* website, [online] accessed at:

http://www.science-enterprise.ox.ac.uk/html/MIke_Fischer.asp , accessed on 3rd March 2007.

Hallam, S., Rhamie, J. and Shaw, J. (2006) Evaluation of the Primary Behaviour and Attendance Pilot, (London, DfES).

Halsey, K., Judkins, M., Atkinson, M. and Rudd, P. (2005), *New Relationship with Schools: Evaluation of Trial Local Authorities and Schools*, Nottingham, DfES Publications, [online] accessed at http://www.dfes.gov.uk/research/data/uploadfiles/RR689.pdf , accessed 16 March 2007.

Hammersley, M. (1997) Educational Research and Teaching: a response to David Hargreaves' TTA lecture, *British Educational Research Journal*, 23(2), 141-161.

Hammersley, M and Atkinson, P. (2007) *Ethnography: Principles in practice* (3rd Edn), Abingdon: Routledge.

Hargreaves, D. (1996) *Teaching as a Research-Based Profession: possibilities and prospects*. Teacher Training Agency Annual Lecture (London, Teacher Training Agency). Available online at http://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/TTA%20Hargreaves%20lecture.pdf accessed November 2011.

Hayduk, L., Cummings, G.G., Boadu, K., Pazderka-Robinson, H., and Boulianne, S. (2007) Testing! Testing! One, Two Three – Testing the theory in structural equation models!, *Personality and Individual Differences*, 42(2), 841-50.

Hansford, B.C. and Hattie, J.A. (1982) The Relationship Between Self and Achievement/Performance Measures, *Review of Educational Research*, 52(1), 123-142.

Heckman, J., Malofeeva, L., Pinto, R. and Savelyev, P. A. (2010) Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. University of Chicago, Department of Economics. [online] Available at http://schubertcenter.case.edu/synapseweb46/documents/en-US/Heckman%20article%202010%20-%20adult%20outcomes.pdf accessed February 2013

Helmke, A. (1989) Affective Student Characteristics and Cognitive Development: Problems, Pitfalls, Perspectives, *International Journal of Educational Research*, 13(8), 915-932.

Hoyle, R. B. and Robinson, J. C. (2002) League tables and school effectiveness: a mathematical model, *Proc. R. Soc. Lond. B*, 270, 113-119.

Hu, L. and Bentler, P. M. (1995) Evaluating model fit, in R. H. Hoyle (ed.), *Structural Equation Modeling: Concepts, Issues, and Applications*, Sage: Thousand Oaks, CA, pp. 76–99.

Hu, L., and Bentler, P. M. (1999) Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling*, 6(1), 1–55.

Huebner, E.S., Gilman, R. and Laughlin, J.E. (1999) A multimethod investigation of the multidimensionality of children's well-being reports: discriminant validity of life satisfaction and self-esteem, *Social Indicators Research*, 46(1), 1-22.

Humphrey N (2008) Key issues in the evaluation of school-based SEL programmes. Paper presented at ESRC seminar series: *The school as a location for the promotion and support of* *mental health*, Southampton, June 2008. [Online: accessed at http://www.abdn.ac.uk/rowangroup/documents/Neil%20Humphrey.ppt on 27th October 2008].

Humphrey N, Kalambouka A, Bolton J, Lendrum A, Wigelsworth M, Lennie C and Farrell P (2008) *Primary Social and Emotional Aspects of Learning (SEAL): Evaluation of Small Group Work,* Department for Children, Schools and Families (DCSF) /University of Manchester.

Jöreskog, K. G. & Sörbom, D. (1984) *LISREL-VI user's guide (3rd ed.)*, Mooresville, IN, USA, Scientific Software.

Kallestand, J. H., & Olweus, D. (2003) Predicting teachers' and schools' implementation of the Olweus Bullying Prevention Program: A multilevel study, *Prevention and Treatment*, 6(1), article 21.

Kaptchuk, T.J. (2001) The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology*, 54(6), 541-549.

Konu, A.I., Lintonen, T.P. and Rimpelä, M.K. (2002a) Factors associated with schoolchildren's general subjective well-being, *Health Education Research*, 17(2), 155-165.

Konu, A.I., Lintonen, T.P. and Autio, V.J. (2002b) Evaluation of Well-Being in Schools – A Multilevel Analysis of General Subjective Well-Being, *School Effectiveness and School Improvement*, 13(2), 187-200.

Knuver, J.W.M. (1993) De relatie tussen klas- en schoolkenmerken en het affectief functioneren van leerlingen [The relationship between class and school characteristics and the affective functioning of students]. Groningen, The Netherlands: Rijksuniveritiet Groningen, RION, cited in Van Landeghem, G., Van Damme, J., Opdenakker, M-C., De Fraine, B. and Onghena, P. (2002) The Effect of Schools and Classes on Noncognitive Outcomes, *School Effectiveness and School Improvement*, 13(4), 429-451.

Knuver, A.W.M. and Brandsma, H.P. (1993) Cognitive and Affective Outcomes in School Effectiveness Research, *School Effectiveness and School Improvement*, 4(3), 189-204.

Kreft, I.G. (1996) *Are multilevel techniques necessary? An overview including simulation studies*, cited in: C.T. Fitz-Gibbon (1997) *The Value Added National Project Final Report: Feasibility Studies for a National System of Value-Added Indicators*, London, School Curriculum and Assessment Authority.

Kreft, I. & De Leeuw, J. (1998) *Introducing Multilevel Modelling*, London, Sage Publications.

Kyriakides, L., Kaloyirou, C., & Lindsay, G. (2006) An analysis of the revised Olweus Bully/Victim questionnaire for students using the Rasch model, *British Journal of Educational Psychology*, 76(4), 781–801.

Kyriakides, L., Bosker, R., Muijs, D., Papadatos, Y., & Van Petegem, P. (2011). *Designing evidence-based strategies and actions to face bullying by considering socio-ethnic diversities in school populations and evaluating their effects (European Commission's Daphne III Programme JLS/DAP/2007-1/226)*, Nicosia, Cyprus: University of Cyprus, available online at http://www.ucy.ac.cy/data/jls/publications/FINAL%20REPORT.pdf accessed September 2013.

Kyriakides, L., Creemers, B.P.M., Papastylianou, D. and Papadatou-Pastou, M. (2013) Improving the School Learning Environment to Reduce Bullying: An Experimental Study, *Scandinavian Journal of Educational Research*, (in press).

Lenkeit, J. (2013) Effectiveness measures for cross-sectional studies: a comparison of value-added models and contextualised attainment models, *School Effectiveness and School Improvement*, 24(1), 1-25.

Luyten, H. (2003) The size of school effects compared to teacher effects: an overview of the research literature, *School Effectiveness and School Improvement*, 14(1), 31-35.

Maas, C.J.M. and Hox, J.J. (2005) Sufficient Sample Sizes for Multilevel Modeling, *Methodology*, 1(3), 86–92.

Mangan, J., Pugh, G. and Gray, J. (2005) Changes in Examination Performance in English Secondary Schools over the Course of a Decade: Searching for Patterns and Trends Over Time, *School Effectiveness and School Improvement*, 16(1), 29 - 50.

Marsh, H. W. and Parker, J W. (1984) Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, 47, 213-231.

Marsh, H. W. (1991) The failure of high ability high schools to deliver academic benefits: The importance of academic self-concept and educational aspirations, *American Educational Research Journal*, 28(2), 445-480.

McDonald, R.P. and Ho, M.-H.R. (2002) Principles and Practice in Reporting Statistical Equation Analyses, *Psychological Methods*, 7(1), 64-82.

McPherson, A. (1992) Measuring added value in schools, *NCE briefing 1*, London: National Commission on Education.

Michel, D. (2007) SEAL throughout the school, *SEAL Headteachers' Conference*, London, DCSF.

Miles, J. and Shevlin, M. (2007) A time and a place for incremental fit indices, *Personality and Individual Differences*, 42 (5), 869-74.

Miliband, D. (2003) The Annual Leadership Lecture October 2003, Nottingham, National College of School Leadership.

Miliband, D. (2004) Personalised Learning: Building a new relationship with schools, *North of England Education Conference*.

MoE Singapore (2009) *Desired Outcomes of Education* [online] available at http://www.moe.gov.sg/education/desired-outcomes/ accessed on 25th June 2013.

Moody, I. (2001) A case-study of the predictive validity and reliability of Key Stage 2 test results, and teacher assessments, as baseline data for target-setting and value-added at Key Stage 3, *The Curriculum Journal*, 12(1), 81-101.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D. and Ecob, R. (1988). *School matters. The junior years.* Somerset, UK: Open Books.

Mortimore, P. (1998) *The Road to School Improvement*, Lisse, The Netherlands, Swets & Zeitlinger.

Mourshed, M., Chijioke, C. and Barber, M. (2010) *How the world's most improved school systems keep getting better*, London, McKinsey & Company.

Muijs, D. (2004) *Doing Quantitative Research in Education*, London: SAGE.

Mulaik, S. A. and Millsap, R. E. (2000) Doing the four-step right, *Structural Equation Modelling*, 7(1), 36-74.

Myers, K. & Goldstein, H. (2004) Opinion: Adding Value?, *Education Journal*, p.15.

National Statistics (2006) Beginners Guide to UK Geography: Super Output Areas (SOAs) [online] accessed at http://www.statistics.gov.uk/geography/soa.asp#3layers accessed 16 March 2007.

Neidell, M. and Waldfogel, J. (2010) Cognitive and Noncognitive Peer Effects in Early Education, *The Review of Economics and Statistics*, 92(3), 562–576.

NHS Health Advisory Service (1995) *Together We Stand: The Commissioning, Role and Management of Child and Adolescent Mental Health Services*, London, HMSO.

NICE (2007) *Mental Wellbeing of Children Public Health Intervention Guidance.* [Online: accessed at http://www.nice.org.uk/nicemedia/pdf/MentalWellbeingFieldworkReport.pdf accessed on 18th August 2009].

NICE (2008) *Promoting children's social and emotional wellbeing in primary education -NICE public health guidance 12*. London, National Institute for Health and Clinical Excellence.

Noortgate, W. V. d., Opdenakker, M.-C. & Onghena, P. (2005) The Effects of Ignoring a Level in Multilevel Analysis, *School Effectiveness and School Improvement*, 16(3), 281 - 303.

Nuttall, D. (1991) An instrument to be honed: Tables do not reflect schools' true performance, *Times Educational Supplement*, September 13, p. 22.

Nuttall, D., Goldstein, H., Prosser, R. & Rasbash, J. (1989) Differential school effectiveness, *International Journal of Educational Research*, 13, 769–776.

Ofsted (2003a) *Lincoln Birchwood Junior School: Inspection report*. Available online at: http://www.ofsted.gov.uk/provider/files/784407/urn/120508.pdf (accessed April 2008).

Ofsted (2003b) *PANDA Report for an anonymous Secondary School A*. Available online at: http://www.ofsted.gov.uk/assets/3569.pdf (accessed 16th December 2006).

Ofsted (2005) Information Sheet 2: Explanation of KS4 scoring system, accessed online at http://www.ofsted.gov.uk/assets/Internet_Content/Shared_Content/Files/PANDA_guidance/interpretingdata_infosheet2_explainks4scores.doc on 5th February 2008

Ofsted (2005a) Old PANDA Training Information, Office for Standards in Education, [online] accessed at http://www.ofsted.gov.uk/portal/site/Internet/menuitem.455968b0530071c4828a0d8308c08a0c/?vgnextoid=dcd68c9acea71110VgnVCM1000003507640aRCRD , accessed on 14 April 2007.

Ofsted (2005b) *Developments in online school data – important changes for 2006*, Office for Standards in Education, [online] accessed at http://www.ofsted.gov.uk/assets/Internet_Content/Shared_Content/Files/Introductory_leaflet.pdf , accessed on 14 April 2007.

Ofsted (2005) *Healthy minds: Promoting emotional health and well-being in schools*. London, Office for Standards in Education.

Ofsted (2007) Developing social, emotional and behavioural skills in secondary schools: A five term longituduinal evaluation of the Secondary National Strategy pilot, (London, Office for Standards in Education).

Ofsted (2007) *What is RAISEonline?* Office for Standards in Education, [online] accessed at http://www.ofsted.gov.uk/portal/site/Internet/menuitem.455968b0530071c4828a0d8308c08a0c/?vgnextoid=5728ed8d9712d010VgnVCM1000003507640aRCRD , accessed on 14 April 2007.

Ofsted (2008) *Lincoln Birchwood Junior School: Inspection report*. Available online at: http://www.ofsted.gov.uk/provider/files/898001/urn/120508.pdf (accessed April 2008).

Ofsted (2009) Indicators of a school's contribution to well-being, London: Office for Standards in Education, Children's Services and Skills.

Ofsted (2011) Lincoln Birchwood Junior School: Inspection report. Available online at: http://www.ofsted.gov.uk/provider/files/1975925/urn/120508.pdf (accessed September 2013).

Olweus, D. (1996) *The revised Olweus Bully/Victim Questionnaire for Students*, Bergen, Norway: University of Bergen, cited in Kyriakides, L., Bosker, R., Muijs, D., Papadatos, Y., & Van

Petegem, P. (2011). *Designing evidence-based strategies and actions to face bullying by considering socio-ethnic diversities in school populations and evaluating their effects (European Commission's Daphne III Programme JLS/DAP/2007-1/226)*, Nicosia, Cyprus: University of Cyprus, available online at http://www.ucy.ac.cy/data/jls/publications/FINAL%20REPORT.pdf accessed September 2013.

Opdenakker, M.-C. and Van Damme, J. (2000) The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research, *School Effectiveness and School Improvement*, 11(1), 103 – 130.

Opdenakker, M.-C. and Van Damme, J. (2000) Effects of Schools, Teaching Staff and Classes on Achievement and Well-Being in Secondary Education: Similarities and Differences Between School Outcomes, *School Effectiveness and School Improvement*, 11(2), 165-196.

Prais, S. J. (2001) Grammar Schools' Achievements and the DfEE's Measures of Value-added: an attempt at clarification, *Oxford Review of Education*, 27(1), 69-73.

QCA (2001) *Supporting school improvement: Emotional and behavioural development.* London, Qualifications and Curriculum Authority

QCA (2007) *A framework of personal, learning and thinking skills*. Available online at: http://www.qca.org.uk/qca_5866.aspx (accessed 19th February 2008).

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2005a) *MLwiN Version 2.02.* Centre for Multilevel Modelling, University of Bristol.

Rasbash, J., Steele, F., Browne, W.J. and Prosser, B. (2005b) *A User's Guide to MLwiN, version 2.0*. Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W. and Bryk, A. S. (2002) *Hierarchical linear models. Applications and Data Analysis Methods*, London: Sage.

Ray, A. (2006) *School Value Added Measures in England: A paper for the OECD Project on the Development of Value-Added Models in Education Systems*, London, DfES.

Raykov, T. (2001) Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints, *British Journal of Mathematical and Statistical Psychology*, 54, 315-323.

Raykov, T. (2004) Behavioral scale reliability and measurement invariance evaluation using latent variable modelling, *Behavior Therapy*, 35, 299-331.

Rosenberg, M. (1965) *Society and the Adolescent Self-Image*, Princeton, NJ: Princeton University Press.

Rosenberg, M. (1979) *Conceiving the Self*, New York: Basic Books.

Rutter, M., Maughan, B., Mortimore, P. and Ouston, J. (1979) *Fifteen thousand hours: Secondary schools and their effects on children*. London: Open Books.

Salmivalli, C., Kaukiainen, A., Kaistaniemi, L. and Lagerspetz, K.M.J. (1999) Self-Evaluated Self-Esteem, Peer-Evaluated Self-Esteem, and Defensive Egotism as Predictors of Adolescents' Participation in Bullying Situations, *Personality and Social Psychology Bulletin*, 25(10), 1268-1278.

Salovey, P. & Mayer, J. D. (1990) Emotional Intelligence, *Imagination, cognition and personality*, 9, 185-211.

Salway Ash (2012) *New Mission Statement*, [online] http://salwayashschool.org/main/?p=1083, accessed September 2013).

Sammons, P. & Smees, R. (1998) Measuring Pupil Progress at Key Stage 1: using baseline assessment to investigate value added, *School Leadership & Management*, 18(3), 389-407.

Sammons, P. (2007) *School effectiveness and equity: making connections*, Reading: CfBT Education Trust.

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B. and Grabbe, Y. (2007a) *Effective Pre-school and Primary Education 3-11 Project (EPPE 3-11): Influences on Pupils' Attainment and Progress in Key Stage 2: Cognitive Outcomes in Year 5 Full Report*, London: Institute of Education, University of London.

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., Barreau, S. and Grabbe, Y. (2007b) *Effective Pre-school and Primary Education 3-11 Project (EPPE 3-11): Influences on Pupils' Development and Progress in Key Stage 2: Social/behavioural Outcomes in Year 5*. Research Report No. DCSF-RR007, Nottingham: DfES Publications.

Sammons, P., Sylva, K., Siraj-Blatchford, I., Taggart, B., Smees., R. and Melhuish, E.C. (2008a) *Influences on pupils' self-perceptions in primary school: Enjoyment of school, Anxiety and Isolation, and Self-image in Year 5*, London: DCSF / Institute of Education, University of London.

Sammons, P., Sylva, K., Siraj-Blatchford, I., Taggart, B., Smees, R and Melhuish, E.C. (2008b) *Exploring pupils' views of primary school in Year 5*, London: DCSF / Institute of Education, University of London.

Sammons, P., Sylva, K., Melhuish, E.C., Siraj-Blatchford, I., Taggart, B., Toth, K., Draghici, D. and Smees, R. (2011a) *Effective Pre-School, Primary and Secondary Education Project (EPPSE 3-14): Influences on Students' Attainment and Progress in Key Stage 3: Academic Outcomes in English, Maths and Science in Year 9*, London: Institute of Education, University of London / DFE.

Sammons, P., Sylva, K., Melhuish, E.C., Siraj-Blatchford, I., Taggart, B., Draghici, D., Toth, K. and Smees, R. (2011b) *Effective Pre-School, Primary and Secondary Education Project (EPPSE 3-14):*

Influences on Students' Development in Key Stage 3: Social-behavioural Outcomes in Year 9, London: Institute of Education, University of London / DFE.

Sammons, P., Sylva, K., Melhuish, E.C., Siraj-Blatchford, I., Taggart, B., Smees, R., Draghici, D. and Toth, K. (2011c) *Effective Pre-school, Primary and Secondary Education 3-14 Project (EPPSE 3-14): Influences on Students' Dispositions in Key Stage 3: Exploring Enjoyment of School, Popularity, Anxiety, Citizenship Values and Academic Self-Concept in Year 9*, London: Institute of Education, University of London / DFE.

Satorra, A. & Bentler, P. M. (1988) Scaling corrections for chi-square statistics in covariance structure analysis, *American Statistical Association 1988 proceedings of the business and economics section*, Alexandria, VA, American Statistical Association, 308-313.

Satorra, A. & Bentler, P. M. (1994) Corrections to test statistics and standard errors on covariance structure analysis, in: A. v. Eye & C. C. Clogg (Eds) *Latent Variables Analysis*, Thousand Oaks, California, Sage, 399-419.

Saunders, L. (1999) A Brief History of Educational 'Value Added': How Did We Get To Where We Are?, *School Effectiveness and School Improvement*, 10(2), 233-256.

SCAA (1994) *Value-added Performance Indicators for Schools*, London, School Curriculum and Assessment Authority.

SCAA (1997) *Making Effective Use of Key Stage 3 Assessments*, London, School Curriculum and Assessment Authority.

Schagen, I. and Hutchison, D. (2003) Adding Value in Educational Research – the marriage of data and analytical power, *British Educational Research Journal*, 29(5).

Schagen, I. (2006) The use of standardized residuals to derive value-added measures of school performance, *Educational Studies*, 32(2), 119-32.

Sharp, S. (2006) Assessing Value-Added in the First Year of Schooling: Some results and methodological considerations, *School Effectiveness and School Improvement*, 17(3), 329 – 346.

Sharp, S. and Croxford, L. (2003) Literacy in the First Year of Schooling: A Multilevel Analysis, *School Effectiveness and School Improvement*, 14(2), 213 - 231.

Shucksmith, J., Summerbell, C., Jones, S. and Whittaker, V. (2007) *Mental wellbeing of children in primary education (targeted/indicated activities)*, London: National Institute of Clinical Excellence.

Simon, S.D. (2001) Is the randomized clinical trial the gold standard of research? *Journal of Andrology*, 22(6), 938-943.

Slade, M. and Priebe, S. (2001) Are randomised controlled trials the only gold that glitters? *The British Journal of Psychiatry*, 179, 286-287.

Smith, D. and Tomlinson, S. (1989) *The school effect: A study of multi-racial comprehensives* (London, Policy Studies Institute).

Smith, P., O'Donnell, L., Easton, C. and Rudd, P. (2007) Secondary Social, Emotional and Behavioural Skills (SEBS) Pilot Evaluation, *DCSF Research Report RR003* (Nottingham, Department for Children, Schools and Families).

Snijders, T. A. B. and Bosker, R. J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, London: Sage.

Spradbery, J. and Cashin, G. (2006) “Supporting Schools…Using Data”, conference for Local Authority School Improvement Staff, London, 17th October 2006.

Steer, A. (2005) *Learning Behaviour - The Report of the Practitioners Group on School Behaviour and Discipline*. Available online at: http://www.dfes.gov.uk/behaviourandattendance/uploads/Learning%20Behaviour%20(published).pdf (accessed 18th February 2008).

Sylva, K., Sammons, P., Melhuish, E.C., Siraj-Blatchford, I. and Taggart, B. (1999) *The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 1 - An Introduction to the EPPE Project*, London: DfEE / Institute of Education, University of London.

Taylor, J. and Nguyen, A. H. (2006) *An Analysis of the Value Added by Secondary Schools in England: Is the Value Added Indicator of Any Value?* Available online at: http://www.lums.lancs.ac.uk/files/economics/6421/download/ (accessed 26th September 2006).

TDA (2007a) *Professional Standards for Teachers – Advanced Skills Teachers*, London, Training and Development Agency for Schools.

TDA (2007b) *Professional Standards for Teachers – Excellent Teachers*, London, Training and Development Agency for Schools.

TDA (2007c) *Professional Standards for Teachers – Qualified Teacher Status*, London, Training and Development Agency for Schools.

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

Thomas, S. & Goldstein, H. (1995) Questionable value, *Education*, p. 17.

Thomas, S. & Mortimore, P. (1996) Comparison of value added models for secondary school effectiveness, *Research Papers in Education*, 11(1), 5-33.

Thomas, S., Sammons, P. and Mortimore, P. (1997) Stability and consistency in secondary schools’ effects on students’ GCSE outcomes over three years, *School Effectiveness and School Improvement*, 8, 169 – 197.

Thomas, S., Smees, R., MacBeath, J., Robertson, P. and Boyd, B. (2000) Valuing pupils’ views in Scottish schools, *Educational Research and Evaluation*, 6(4), 281–316.

Thomas, S. (2001) Dimensions of Secondary School Effectiveness: Comparative Analyses Across Regions, *School Effectiveness and School Improvement*, 12(3), 285 - 322.

Thomas, S., Peng, W. J. and Gray, J. (2007) Modelling patterns of improvement over time: value added trends in English secondary school performance across ten cohorts, *Oxford Review of Education*, 33(3), 261–295

Thomson, D. (2007) *Using RAISEonline in self-evaluation. Understanding what the inspectors will see. (Secondary)*. Available online at: www.egfl.org.uk/export/sites/egfl/categories/admin/data/_docs/raise_online/raiseSec.pdf (accessed 30/08/2007).

Thomson, D. and Knight, T. (2006) *Valued added models and methods: Differences between KS2-KS4 CVA and FFT SX*, presentation given at Fischer Family Trust 2006 Conference, [online] downloaded from FFT secure server (accessed 15 March 2007).

Tymms, P. and Dean, C. (2004) Value Added in the Primary School League Tables A Report for the National Association of Head Teachers, Durham, CEM Centre, University of Durham.

Tymms P.B., Merrell, C. and Coe, R.J. (2008) Educational Policies and randomized Controlled Trials. *The Psychology of Education Review,* 32, 3-7 & 26-29.

UNICEF (2007) An overview of child well-being in rich countries: A comprehensive assessment of the lives and well-being of children and adolescents in the economically advanced nations, *Innocenti Report Card 7*, Florence, Italy, The United Nations Children’s Fund.

Van Damme, J., De Troy, A., Meyer, J., Minnaert, A., Lorent, G., Opdenakker, M.-C., and Verduyckt, P. (1997) Succesvol doorstromen in de aanvangsjaren van het secundair onderwijs –bijlagen [Successful passing through the first years in secondary education – appendices]. K.U. Leuven, LIVO, cited in Opdenakker, M-C. and Van Damme, J. (2000) Effects of Schools, Teaching Staff and Classes on Achievement and Well-Being in Secondary Education: Similarities and Differences Between School Outcomes, *School Effectiveness and School Improvement*, 11(2), 165-196.

Van Damme, J., Van Landeghem, G., De Fraine, B. Opdenakker, M.-C., and Onghena, P., (2000) Maakt de school het verschil? Effectiviteit van scholen, leraren en klassen in de eerste graad van het middelbaaronderwijs (Does the school make a difference? cited in Van Landeghem, G., Van Damme, J., Opdenakker, M-C., De Fraine, B. and Onghena, P. (2002) The Effect of Schools and Classes on Noncognitive Outcomes, *School Effectiveness and School Improvement*, 13(4), 429-451.

Van Damme, J., De Fraine, B. Van Landeghem, G., Opdenakker, M.-C. and Onghena, P. (2002) A New Study on Educational Effectiveness in Flanders: An Introduction, *School Effectiveness and School Improvement*, 13(4), 383-397.

Van de gaer, E., De Fraine, B., Pustjens, H., Van Damme, J., De Munter, A. and Onghena, P. (2009) School effects on the development of motivation toward learning tasks and the development of academic self-concept in secondary education: a multivariate latent growth curve approach, *School Effectiveness and School Improvement*, 20(2), 235-253.

Van Landeghem, G., Van Damme, J., Opdenakker, M.-C., De Fraine, B. and Onghena, P. (2002) The Effect of Schools and Classes on Noncognitive Outcomes, *School Effectiveness and School Improvement*, 13(4), 429-451.

Waters, M. (2007) SEAL and the curriculum aims, *SEAL Headteachers' Conference*, London, DCSF.

Weare, K. and Gray, G. (2003) What Works in Developing Children’s Emotional and Social Competence and Wellbeing?, Nottingham: DfES Publications (Department for Education and Skills).

Weare K (2004) *Developing the emotionally literate school*, London, Paul Chapman Publishing.

Weare, K. (2007) SEAL - why it matters and how it works, *Secondary SEAL Conference* (The Barbican Centre, London, Optimus Education).

Webster-Stratton C (2000) *Goals for the Incredible Years Programs.* [Online: accessed at http://www.incredibleyears.com/library/paper.asp?nMode=1&nLibraryID=464 on 1st September 2009].

Webster-Stratton C (2004) *Quality Training, Supervision, Ongoing Monitoring, and Agency Support: Key Ingredients to Implementing The Incredible Years Programs with Fidelity.* [Online: accessed at http://www.incredibleyears.com/library/items/quality-key-ingredients-fidelity-04.pdf on 1st September 2009].