The London School of Economics and Political Science

Feelings, Friends and Behaviour: Noncognitive attributes of pupils at English secondary schools

Amy Challen

A thesis submitted to the Department of Social Policy of the London School of Economics for the degree of Doctor of Philosophy

London, August 2013

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 97,055 words.

Abstract

The noncognitive features of pupils' experience of school are important: they can affect academic attainment; they have an independent influence on outcomes in later life; and to the extent that they are related to pupils' wellbeing they have intrinsic importance. I present four empirical papers on the emotional health, friendships, and behaviour of pupils in English secondary schools. The first two empirical papers present the results of a large pragmatic controlled trial of an intervention intended to promote pupils' resilience and mental health. I estimate the intervention impact on symptoms of poor mental health, behaviour, absence from school, academic attainment, and popularity. I find small and short-lived impacts on depressive symptoms, absence, and popularity, and a small but more lasting impact on academic attainment. I find no impact on anxiety scores or behaviour. The third paper examines behaviour incidents at school. Poor behaviour is a major challenge to the effectiveness of schooling, and the data I have represents a substantial improvement over previous attempts to measure pupil behaviour. I find that demographic characteristics are strong predictors of the number of incidents per pupil, but they do not explain much of the overall variance in incidents. Incident rates per lesson vary strongly by context within the school, suggesting that schools could influence behaviour by modifying the environment. However, a pupil's rank in terms of behaviour is remarkably persistent over different contexts and through time, suggesting that the tendency to misbehave is a stable noncognitive trait. The fourth paper looks in detail at the impact of time of day and day of the week on behaviour. I find that the strong and persistent day-of-the-week and time-of-day patterns I observe are not due to selective reporting or misreporting, and are not due to endogenous timetabling. Since schedule adjustment could be almost costless, it could be highly cost effective even if the impact on behaviour were much smaller than estimated in my observational data.

Acknowledgements

I gratefully acknowledge the contribution of the hundreds of school and Local Authority staff and the thousands of pupils involved in the UK Resilience Programme and evaluation. Particular and abiding thanks to Lucy Bailey.

Field trials should not be done alone, and I have spent six years being consistently grateful to many colleagues at the Centre for Economic Performance who made the practical elements possible, particularly Linda Cleavely, Nigel Rogers, and John MacIntosh. Thanks to Richard Layard for making it happen in the first place, and for his continued support.

Thanks also to Jane Gillham and Karen Reivich of the Penn Resiliency Project, Philip Noden, Sandra McNally and Steve Gibbons for general advice. Stephen Pam provided extremely valuable advice and discussion on the two chapters on behaviour (Chapter 4 and Chapter 5). Chapter 5 also owes a lot to Alex Bryson.

General thanks to Ashwini Natraj, Mirko Draca, Gill Wyness, Richard Murphy, Kati Szemeredi, Giulia Ferrari, James Vernoit, Tom Cunningham, Marty McGuigan, Ralf Martin, Ferdinand Rauch, Evie Watt and Nele Warrinnier, as well as many others at LSE, for their many and various contributions to life and work. Specific thanks to Alex Lembcke for years of free Stata advice, to Joe Joannes for making stuff work properly, and to Richard Freeman for the poetry. Profound thanks to Steve Todd.

Note on previous work and publications

The data used in this thesis was originally collected as part of the UK Resilience Programme (UKRP) evaluation. The first three chapters in this thesis taken together are an evaluation of this project, while the last two chapters explore related areas using data from UKRP schools. I am therefore submitting this thesis as a series of related publishable papers rather than as a monograph, as there are many linkages between the chapters but no single overriding research question.

The UKRP evaluation was funded by the Department for Children, Schools and Families, and for them I co-wrote three short reports which summarised the intervention impact at three points in time. These were:

Challen, A. R., Machin, S. J., Noden, P., & West, A. (2009). UK Resilience Programme Evaluation: Interim Report: Department for Children, Schools and Families.

Challen, A. R., Machin, S. J., Noden, P., & West, A. (2010). UK Resilience Programme Evaluation: Second Interim Report: Department for Education.

Challen, A. R., Machin, S. J., Noden, P., & West, A. (2011). UK Resilience Programme Evaluation: Final Report: Department for Education.

I am listed as a co-author on the quantitative sections of these reports. I designed the evaluation, helped organise the intervention and training, managed the evaluation, collected and processed the data, carried out the analyses and wrote the reports.[1] Stephen Machin provided advice on the evaluation set up and contributed some analysis and some text to the first interim report (2009). He also read and commented on the second report (2010). The work I present in this thesis draws on the UKRP reports, and I refer to these on many points. However, the analytical work presented in this thesis is entirely new, and much more substantial than the evaluation work presented in the reports.

I also co-wrote a policy article, which presented no empirical results but which summarised the research on UKRP and PRP (the curriculum on which UKRP is based)

[1] There was also a qualitative section in each report which was separate from the quantitative section. I had no input into the qualitative work, which was by Philip Noden and Anne West.

and which discussed the challenges of implementation. I wrote the bulk of the article, while Lucy Bailey added to this. The article was published as follows:

Challen, A. R., & Bailey, L. (2012). The UK Penn Resilience Programme: A summary of research and implementation. *Psychology of Education Review, 36*(2), 32-39.

In addition, I have written a short article for a psychology journal which is currently under review. This presents an evaluation of the first year of UKRP, but uses a different quantitative approach and is written for a different audience to this thesis. I carried out all the analyses and wrote the paper. Jane Gillham reviewed and amended the text several times and made suggestions for further analyses, as well as taking on the final preparation of the manuscript. She had also provided advice on outcome measures and trial design at the start of the UKRP project. Stephen Machin did not contribute to this paper. This paper has been accepted for publication and will appear online in late 2013, appearing in print at a later date:

Challen, A. R., Machin, S. J., & Gillham, J. E. (forthcoming). The UK Resilience Programme: a school-based universal non-randomised pragmatic controlled trial. *Journal of Consulting and Clinical Psychology*.

Thus although all the chapters in this thesis are single authored, I will make frequent reference to co-authored works which have covered different aspects of the UKRP intervention and evaluation, such as the results of the satisfaction surveys published in the first interim report. Any overlap with published work will mainly occur in the descriptions of UKRP in Chapter 1. Chapter 2 presents an evaluation of UKRP on the same outcome variables as were used in the UKRP reports, but uses a different approach and the empirical work is entirely new. The work in Chapters 3, 4 and 5 has not been published in any form.

Since I wrote Chapters 2-5 as stand-alone articles there may also be some overlap in terms of description of the background to the project and the population samples involved. I have tried to minimise this by including much of the background in Chapter 1, but there may still be some repetition.

Table of Contents

List of Figures

Appendix to Chapter 2

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix to Chapter 2

List of Abbreviations

CDI	Children’s Depression Inventory
DCSF	Department for Children, Schools and Families
DfE	Department for Education
DfES	Department for Education and Skills
EBD	Emotional and behaviour disorder
FSM	Free School Meals
GCSE	General Certificate of Secondary Education
IDACI	Income Deprivation Affecting Children Index
IRR	Incidence rate ratio
LA	Local Authority
NPD	National Pupil Database
Ofsted	Office for Standards in Education
RCMAS	Revised Children’s Manifest Anxiety Scale
RCT	Randomised controlled trial
SD	Standard deviation
SE	Standard error
SEAL	Social and Emotional Aspects of Learning
SEN	Special Educational Needs

Introduction

The primary purpose of state-funded schools in England is to promote pupils' academic attainment.[2] Given the importance of cognitive skills to attainment and future opportunities it is not surprising that so much research and practice focuses on these (Card, 1999). But schools also fulfil other roles. They provide a large part of the social context in which children develop: pupils of secondary age in the UK spend at least one third of their waking hours at school during term time, so even if schools were simply a form of childcare they would still have a strong influence on pupils' everyday wellbeing (Hagell, Peck, Zarrett, Giménez-Nadal & Symonds, 2012; Hattie, 2009). This is the first reason why the environment schools provide could be important: because of the intrinsic value of children's wellbeing. In addition, schools socialise pupils, teaching them behavioural and social norms which (if appropriate and well taught) will prepare them for work and the adult world (Bowles, Gintis & Osborne, 2001). Most children's close friendships are formed through school (Burgess & Umaña-Aponte, 2011; Moody, 2001), and relationships with teachers provide important instances of relationships with adults outside of the family (Parsons, 1959). The judgement reached in Brown v. Board of Education recognised the wider responsibility schools had in child development in the USA: "It [education] is the very foundation of good citizenship. Today it is a principal instrument in awakening the child to cultural values, in preparing him for later professional training, and in helping him to adjust normally to his environment." (Brown v. Board of Education, 1954). Schools in the UK may serve the same functions.

Thus whether they intend to or not, schools can be highly influential in fostering a range of noncognitive attributes in their pupils, and these may persist into adulthood.[3] Many of these behaviours, attitudes and social abilities are highly valuable in later life: noncognitive skills play an important role in earnings equations, over and above the contribution of cognitive ability and educational attainment (Segal, forthcoming;

[2] For example, the first aspect of schools which inspectors in England should report on is 'the achievement of students' (Ofsted, 2012a).
[3] Cognitive traits may be defined as general intelligence and the ability to solve abstract problems, leaving 'noncognitive' traits to cover a broad range of skills, attitudes, personality traits and preferences, and this is how I will be using the term. However, noncognitive traits are likely to involve cognitive processes, and their development may be influenced by cognitive skills, and to this extent the term is inexact (Borghans, Duckworth, Heckman & ter Weel, 2008).

Lindqvist & Vestman, 2011; Bowles et al., 2001), and characteristics such as self-control, emotional regulation, perseverance, and agreeableness contribute to success in many areas of life (Almlund, Duckworth, Heckman & Kautz, 2011). Thus the noncognitive characteristics that schools help to develop are important to a child's future wellbeing. Moreover, noncognitive characteristics enter into the production function for academic attainment. For any given individual, concentration, focus, patience, perseverance, hard work and self-control will make a difference to what a pupil of a given level of cognitive ability can achieve (Duckworth & Seligman, 2005). Since secondary school classes usually number 20-30 pupils (DCSF, 2007a), these capacities can easily have an impact on the attainment of other pupils. This is most salient with disruptive behaviour, but many characteristics could affect the outcomes of peers (e.g. academic ambition, Burgess & Umaña-Aponte, 2011). Thus the noncognitive features of pupils' experience of school are important: they have an impact on pupils' happiness; they can affect academic attainment; and they have an independent influence on outcomes in later life. But can schools really influence these characteristics?

There certainly appears to be a widespread belief that schools can affect their pupils' capacities for self-management and other non-cognitive abilities, and that these are an important part of a school's role (DfE, 2010a). However, much discussion in this area suggests that schools could and should do better at this. For example the Confederation of British Industry, a lobbying organisation for UK businesses, regularly bemoans school leavers' lack of readiness for employment and the lack of provision in schools (BBC, 2011; CBI, 2012). Ensuring that schools are pleasant and enjoyable places to be is also deemed to matter by many children and teachers (DfE, 2010a), and the 2006 Education and Inspections Act requires maintained schools to promote the well-being of pupils (c.40, Part 3, Section 38). Yet many children do not find school a pleasant place to be: a substantial minority of pupils report being unhappy at school, while nearly 50% of 14-year-olds report having been bullied within the previous year (DCSF, 2008a).

The majority of secondary schools aim to cover much of this ground through explicit teaching in Personal, Social, Health and Economic (PSHE) education lessons (Formby et al., 2011; Ofsted, 2013). Current guidance from the Department for Education states that PSHE is "an important and necessary part of all pupils' education" and that all schools should teach it, yet there is no statutory curriculum (DfE, 2013a). The Office for Standards in Education (Ofsted) recently found that PSHE provision was not good enough in a substantial proportion of schools (Ofsted, 2013). Thus although this area is deemed to be important by government, schools, and parents, provision is patchy and often inadequate.

Yet a number of programmes which aim to promote psychological wellbeing and life skills have been developed and trialled by psychologists (Payton et al. 2008). Here there appears to be a substantial disconnect between research and practice: these programmes have been developed in many different contexts and countries, but few of them have been delivered at scale under ordinary and sustainable conditions (Calear & Christensen, 2010). This matters because the effectiveness of an intervention in the real world may be very different to that seen when it is delivered by its developers to few students in a controlled setting (Malti, Ribeaud, & Eisner, 2011; Weisz & Jensen, 2001). This gap between research and practice was the motivation for introducing the Penn Resiliency Program (PRP) into UK schools as the UK Resilience Programme (UKRP). Schools were already scheduling time (through PSHE lessons) for the development of noncognitive skills and promoting wellbeing, but in many cases could have been making better use of this time (Formby et al., 2011; Ofsted, 2013).

I designed and ran a controlled trial of an evidence-based intervention to promote wellbeing and non-cognitive skills, in a scaled-up, real-world setting. At the start of the project there were almost no instances of attempts to evaluate similar programmes in the real world, and even now there are very few (Calear & Christensen, 2010). I also take measures from psychology and introduce them to an audience of economists and policy analysts. This is the first contribution of this thesis.

Another contribution is the adoption of methods from different disciplines, and in particular, evaluating an intervention through a controlled trial which used an arbitrary method of assigning pupils to intervention and control groups. The main challenge of any evaluation is to establish causation: what are the effects of an intervention? The main impediment to this is selection bias: people and institutions who choose to engage in certain interventions may not be the same as those who choose not to, so a simple comparison of their respective outcomes is not meaningful and they do not provide convincing counterfactuals of one another (Ch. 2, Angrist & Pischke, 2009). In the medical field, thousands of randomised controlled trials (RCTs) have been carried out since the first published RCT appeared in 1948, and this method forms the basis of most evidence-based medical practice (Stolberg, Norman & Trop, 2004) For example, the Centre for Evidence Based Medicine guidelines place RCTs and systematic reviews of

RCTs at the top of hierarchy of evidence, meaning that results from these studies should be given more weight than results from studies using other methods, and the National Institute for Health and Care Excellence uses a similar hierarchy in developing public health guidance (Phillips et al., 2009; NICE, 2006). In economics, randomised trials of social policies have been used periodically since the 1960s, but in the past two decades economics researchers have made increasing use of field experiments, aiming at applying randomisation to naturally occurring settings and constituting a middle way between laboratory experiments and observational data from the real world (Levitt & List, 2009). In particular, the Abdul Latif Jameel Poverty Action Lab (J-PAL) was established at the Economics Department of MIT in 2003 with the aim of using RCTs to evaluate interventions and provide robust policy evidence. At the time of writing, they had tallied 409 evaluations in progress or completed, in 52 countries (J-PAL, 2013). Many economists now regard an idealised RCT design as the 'benchmark' against which to compare the validity of other identification methods, with the most convincing research designs using true random assignment (Angrist & Pischke, 2009).[4]

We can compare these traditions with that in education research. Most available evidence on education policy is obtained from non-experimental research (Bouguen & Gurgand, 2012).[5] There may be good reasons for this: if most education policy interventions change at the macro level, it may be difficult to design and implement a meaningful RCT. In these cases, natural or policy experiments could be used, where there is at least a well-defined control group (although we must judge its suitability; Angrist, 2003). More serious is the poor quality of a substantial proportion of the educational research published in the UK, a criticism levied by Hargreaves (1996) and investigated by Tooley and Darby (1998), who found that much research was highly partisan, not relevant to policy, and used questionable methods or failed to report methods altogether. Only 15 of 41 articles they looked at in detail could be described as examples of good practice. At the start of the UKRP trial in 2007, there had been very few randomised education trials conducted in the UK, or indeed elsewhere in Europe (Bouguen & Gurgand, 2012). Research commissioned by the Department for Education to evaluate major policies had tended to use (at best) matched control groups, examples

[4] Not all economists approve of the widespread use of RCTs without accompanying attempts at constructing causal models or uncovering the mechanisms of impact, e.g. Deaton (2009).

[5] Excluding the laboratory-based experiments of cognitive science and psychology: these tend to be explanatory trials (Schwartz & Lellouch, 1967), aiming at understanding of the mechanisms of cognition or how an intervention performs under ideal circumstances. This is quite a different goal from understanding the actual impact of policies implemented on the ground.

being the evaluation of the Education Maintenance Allowance and the Primary Behaviour and Attendance Pilot (Middleton et al., 2005; Hallam, Rhamie & Shaw, 2006).[6] Other methodological flaws encountered include the lack of any control groups; short or no follow-up periods; and use of outcomes which rely on participants' satisfaction with an intervention rather than impact on more tangible outcomes (e.g. Hallam et al., 2006 ch.6).

The UKRP trial, although not randomised, used within-school arbitrary assignment and therefore presented a substantial improvement over standard education policy evaluations. The Department for Children, Schools and Families (DCSF) funded the research partly for this reason.[7] Since then, education trials with random allocation have become more popular in the UK. In 2011 the Department for Education established the Education Endowment Foundation (EEF). The EEF aims to promote the educational achievement of pupils from disadvantaged background through funding innovative interventions. Any organisation can propose an intervention, but conditions of receiving funding for it are that the project will be set up as an RCT, and that EEF will appoint independent evaluators unaffiliated to the programme providers.[8] In August 2013 there were at least 34 EEF trials underway (EEF, 2013). A recent Cabinet Office policy paper promoted the use of RCTs in all areas of policy (Haynes, Service, Goldacre & Torgerson, 2012).

Thus there is some agreement about the usefulness of experiments in policy research, although the details of trials may differ substantially by discipline and the purpose of the research. For a start, there is an important distinction to be made between explanatory and pragmatic trials (Schwartz & Lellouch, 1967).[9] Explanatory trials show how an intervention performs under ideal conditions. Strict entry criteria will usually

---

6 The Department for Education (DfE) is currently the name for the government department responsible for setting school-level education policy in England. It has previously been known as the Department for Education and Employment (1995-2001); the Department for Education and Skills (DfES; 2001-2007); and the Department for Children, Schools and Families (DCSF; 2007-2010).

7 Most evaluations funded by DCSF at this time evaluated or were closely aligned with government policy. This intervention was independently funded, so the research design was a major reason for DCSF support of the evaluation. However, the department was also promoting an alternative social and emotional learning programme in schools at the time (SEAL; see Humphrey, Lendrum & Wigelsworth, 2010), so the subject matter of UKRP was of interest.

8 This follows a similar drive in the US to promote rigour in education research: the Education Sciences Reform Act passed by the US Congress in 2002 requires rigorous experimental or quasi-experimental techniques to be used for all education research which is federally funded.

9 I use the terminology of Schwartz and Lellouch (1967), but pragmatic trials are very similar in concept to 'social experiments': "full-scale policies or interventions in a social context that should be as close as possible to the conditions of a scaled up program" (Bouguen & Gurgand, 2012; the terminology of 'social experiments' is also used by Levitt & List, 2009).

ensure a homogenous population. They are carried out to advance scientific knowledge, and are concerned with identifying the mechanisms of impact as well as the outcome. By contrast, pragmatic trials are performed in naturalistic conditions with the aim of informing policy decisions such as choices between different interventions. The population involved will be heterogeneous and should be representative of the population as a whole – in sum, it should give an indication as to how an intervention would perform if applied in the real world (Roland & Torgerson, 1998). The distinction between explanatory and pragmatic trials is not dichotomous, but rather a multidimensional continuum between two extremes (Thorpe et al., 2009). Explanatory trials dominate in many fields (Zwarenstein, Oxman & PRACTIHC, 2006). This can be problematic: the lack of external validity of many of these trials may bias their results; they may be of little use to practitioners and policy makers because of their restricted scope; and the lack of information about real-world impacts may result in underuse of the associated interventions (Patsopoulos, 2011; Treweek & Zwarenstein, 2009; Rothwell, 2005).

Pragmatic trials are not without problems. Context may play an important role in the success of programme on the ground, which is a major reason why pragmatic trials may have different outcomes to explanatory trials, and failing to take this context into account will generate results of little use to the policymaker. For example, if a programme fails because of poor administrative organisation, this may not in itself reflect on the quality of the intervention or even on the ability of the intervention staff, but only on the effectiveness of the intervention in those particular circumstances. This is particularly a problem for interventions which are vague or ill-defined, as they may not be reproducible or suitable for this sort of evaluation (Deaton, 2009; Treweek & Zwarenstein, 2009; Bouguen & Gurgand, 2012). If circumstances are highly atypical, a pragmatic trial might have no more external validity than a tightly-controlled, laboratory-based explanatory study. Two ways of accounting for context may be to take measures of intervention implementation, adherence and quality; or to carry out qualitative work around the implementation. I report on measures of implementation adherence in the evaluation below.[10] In addition, pragmatic trials may fail to account for the mechanisms of impact, making them of limited use in understanding what really

[10] The UKRP evaluation also contained qualitative research elements, but I do not report on them in this thesis because I did not author them. See Challen, Machin, Noden and West (2009; 2010; 2011) for details of the qualitative findings.

matters to policy. Moreover, the population heterogeneity which makes pragmatic trials more 'realistic' than explanatory trials can pose problems in interpreting and extrapolating the effects, as variation in observed outcomes may partly reflect the heterogeneity of the population and may necessitate subgroup analysis (Patsopoulos, 2011).

The standard evaluation technique in the field of intervention psychology is to use small-scale explanatory controlled trials, usually with randomised condition assignment. For example, all the PRP trials reviewed in Brunwasser, Gillham and Kim (2009) use small samples and maintain good control of the intervention, although not all use randomisation. These tend to have good internal validity – overcoming the selection problem through randomised assignment; overcoming other biases through blinding – but, partly because of the constraints put on the implementation, may not reflect what would happen if the intervention was implemented in the real world. This is partly what creates the scaling up issues mentioned above. The UKRP was implemented in order to test PRP on the ground in real-world conditions. The evaluation used a trial method not common in education research at the time, but borrowed from public health and economics literature, with the aim of making the research results both credible and applicable. It therefore contributes to the literature as a pragmatic education trial.

However, this thesis covers more than just the UKRP trial evaluation, and trials are not the only way to gain evidence. Using data obtained in the course of the evaluation I investigate the predictors and persistence of behavioural incidents, which I interpret as a measure of noncognitive skills. The method used here is purely descriptive, using observational panel data to describe relationships between variables and test whether hypotheses are consistent with the data. Basic descriptive work is critical to gaining an understanding of the nature of a variable such as behaviour incidents, and can form the basis for future experimental work (Borghans, Duckworth, Heckman & ter Weel, 2008). This is particularly true for the behaviour incidents data I have, since I know of no other published statistics on incidents for any school in England, with the Department for Education and other organisations relying primarily on proxy measures of behaviour such as teacher reports, Ofsted inspections (which in turn rely on reports and a few days of observations), and exclusion statistics (e.g. DfE, 2012a). In my final empirical chapter, I use quasi-exogenous lesson scheduling as a natural experiment to uncover the causal impact of time of day and day of the week on behaviour incidents. Natural experiments are perhaps the most pragmatic of pragmatic trials: even the events or constraints giving rise to the quasi-exogenous allocation are a product of the everyday functioning of the system, and participants are likely to accept them as such. In this example, even the data collection was routine. Thus what natural experiments can lack in terms of the credibility of their ‘random’ allocation – and researchers must work hard to convince their audience that the mechanism is exogenous, as I do – they may make up for with their uncontrived settings and execution (Angrist & Pischke, 2009). Thus this thesis uses three main research methods – a controlled trial; descriptive relationships between variables; and a natural experiment – to investigate the emotional health, friendships, and behaviour of pupils in English secondary schools.

There are two major limitations to the research I present here. First, the lack of randomisation in the trial component of this thesis may undermine the credibility of the results. The trial arbitrarily assigned classes to the intervention, and this has produced intervention and control groups which are similar on a range of outcomes. True randomisation may not have made any difference to the results obtained, but it could have made them more convincing (Oliver et al., 2010). Second, I had hoped to include measures of pupils’ academic attainment in national Key Stage 3 tests in English, maths and science. The UKRP intervention pupils would have sat these tests in May 2010, so the results could have been used as a measure of academic attainment at two-year postintervention, and could be included in the chapters on behaviour. However, Key Stage 3 tests were abolished in 2008 (BBC, 2008), so these data were not available. The next point at which students sit nationally-graded exams is at GCSE, which for the intervention pupils would have been in May 2012. Since these data were not be available until April 2013 this did not give me enough time to include this outcome in the analysis, or in the analysis of pupils’ behaviour. However, it will be available for future work.

Research questions

The thesis as a whole is about related themes: mental health and wellbeing, behaviour, and educational outcomes in the context of secondary schools. These are addressed through the main research questions for each empirical chapter as below:

Chapter 2: Impact of the UK Resilience Programme on mental health symptoms, absence and academic attainment

1. What impact did the UK Resilience Programme have on pupils' symptoms of depression and anxiety, behaviour scores, absence from school and academic attainment?

2. Was there heterogeneity in intervention impact by pupil or workshop characteristics?

Chapter 3: Impact of the UK Resilience Programme on pupil popularity

3. What impact did the UK Resilience Programme have on pupils' popularity?
4. Was there heterogeneity in intervention impact by pupil or workshop characteristics?

Chapter 4: Pupil behaviour in secondary schools

5. Do all pupils misbehave?
6. Do demographic characteristics predict behaviour?
7. Are there different dimensions of behaviour incidents?
8. Is (bad) behaviour persistent? Is behaviour context specific?

Chapter 5: Day and time patterns in behaviour at secondary school

9. Is pupil's behaviour affected by the time of day and day of the week?

Overview of thesis

Chapter 1 describes the background to the UK Resilience Programme, and provides information on the intervention, which uses techniques based on cognitive behavioural therapy (CBT). The evaluation design is discussed, including the recruitment of participants and the method used to assign classes of pupils to the intervention. It sets the context for the evaluation results reported in Chapters 2 and 3.

The first empirical paper (Chapter 2) is the main evaluation paper for the UK Resilience Programme, presenting the impact of the intervention on symptoms of depression and anxiety, poor behaviour, absence from school, and academic attainment. I find a small impact on depression symptoms scores and pupil absence from school, which do not last beyond the end of the academic year in which the programme took place. I also find a small impact on academic attainment, which lasts until two years after the end of the intervention. I find no impact on anxiety scores or behaviour at any point. There appears to be some heterogeneity in programme impact: in particular, pupils who had worse depression scores or lower academic attainment at baseline experienced greater intervention-induced reductions in depressive symptoms than other pupils. I also found some difference in programme impact by the quality of workshop implementation, measured by hours scheduled and number of pupils in the class, with smaller workshop groups scheduled for more hours having a greater impact on some outcomes. However, given the cost of the intervention and the small average impacts, it is not clear that offering this sort of programme universally offers value for money.

Contributions to the literature This evaluation is one of a very few scaled up interventions in this area, making a significant contribution to the psychology literature. My use of a range of outcome measures means that this trial also contributes to the education literature, using a relatively high quality research design. Economists have only recently started paying attention to the role of noncognitive skills in life outcomes, and this and the subsequent papers provide some information as to how they might be developed and to what effect. I also provide clear cost information, to enable comparison of the cost-effectiveness of this intervention with that of similar interventions.

Chapter 3 assesses the impact of the intervention on pupils' friends. Questionnaires administered in summer 2008, 2009 and 2010 asked pupils to list their friends, and I can use the number of friendship nominations a pupil receives to estimate the impact of the intervention on social capital. I find a small average impact at postintervention equivalent to about one half of an extra friendship nomination for pupils in the intervention group, or to a shift up 3 percentiles in the popularity distribution. This is driven by increased nominations as a more distant friend, with little or no change seen in the number of close friend nominations intervention pupils receive. There is no impact at one-year or two-year follow-up. The results suggest that programmes such as UKRP can have an impact on pupils' social skills and popularity, but the lack of a lasting impact suggests that social capital can decay just as it can be developed, and schools might need to maintain the context in order to see lasting changes in social cohesion.

Contributions to the literature There are very few scaled-up wellbeing intervention with pupils' popularity as an outcome, particularly in universal samples. I use a robust measure of popularity: others' reports of friendships, unrestricted by gender or other characteristics such as age or school attended; with pupils given the opportunity to list a large number of friends. My results are informative as to the development and decay of social capital.

Chapter 4 investigates the determinants and persistence of an indicator of noncognitive skills – behaviour incidents at school. Poor behaviour in schools is a major challenge to the effectiveness of schooling. The data I have on behaviour incidents represents a substantial improvement over previous attempts to measure pupil behaviour and noncognitive skills more generally, as it uses administrative data and is likely to be more reliable than either self- or teacher-reports. I find that demographic characteristics are strong predictors of the number of incidents per pupil, but that they do not explain much of the overall variance. Different types of incident represent different dimensions of noncognitive characteristics, and the distribution is highly skewed, with less than 10% of pupils responsible for more than half of incidents. The rate of incidents varies according to the context within the school: by time of day, day of the week, and the subject being studied, suggesting that behaviour responds strongly to context and that schools could influence behaviour by modifying the environment pupils find themselves in. However, a pupils' behaviour rank is remarkably persistent over these different contexts. Behaviour is also persistent over a period of three years, as is the rank order of behaviour over this time. This suggests that the tendency to misbehave is a stable noncognitive trait, and targeting pupils with problematic behaviour for inclusion in interventions may be effective and cost-effective.

Contributions to the literature The data I use to examine pupils' behaviour represents a substantial improvement over other measures, which usually rely on proxies (such as absence or exclusion) or on teachers' global judgements of behaviour. I can therefore accurately describe the distribution of a noncognitive trait in a universal population, a contribution to the economics literature which has already shown school behaviour to be related to later outcomes. In particular, my ability to observe behaviour in different contexts within the school allows me to make inferences about the malleability and rank order stability of the tendency to misbehave. The ability to accurately measure behaviour also makes a substantial contribution to the education and psychology literature on misbehaviour in schools.

Chapter 5 uses the same behaviour incident data to examine the impact of time of day and day of the week on behaviour, using the arbitrariness of lesson scheduling as a natural experiment. There is a sizeable literature on time of day and day of the week effects on many areas of human activity, but this is the first paper to show these patterns in schools, and my identification is not subject to many of the usual problems such as endogenous selection into activities or schedules, or variation through the week in activities. I find that Mondays have the highest rate of incidents, followed by Tuesdays, with substantially lower rates on Wednesdays, Thursdays and Fridays. The first lesson of the day has the lowest rate of incidents, The rate then rises through the day to reach a peak during lesson 5 (the last lesson of the day). I investigate mechanisms, concluding that these patterns are not due to selective reporting or misreporting, and are not due to endogenous timetabling. Rather, they are likely to be related to pupils' subjective wellbeing, particularly tiredness and boredom. Pupil absence and lateness; teacher wellbeing; and teacher absence may also contribute. Poor behaviour can be a serious problem for schools, and I suggest ways of improving behaviour through modifying schedules. Specifically, subjects critical to academic success (such as English and maths) and subjects with higher incident rates, should be scheduled during low-incident time slots. This could reduce the total number of incidents as well as improving attainment in the subjects which make most difference to pupils' overall attainment. Since timetable modification would be almost costless, this could be a very cost effective strategy for improving behaviour in schools.

Contributions to the literature I contribute to the literature on time of day and day of the week effects in psychology, economics and industrial organisation, using a natural experiment which provides a more plausibly exogenous framework than in the majority of settings studied. In addition, I suggest a way of improving the efficiency of schooling by rearranging timetables, a policy which could be very cost effective, contributing to education research.

Chapter 6 reiterates the main findings from the individual chapters, and outlines research contributions, policy implications, and limitations of the research and directions for future work.

Chapter 1: The UK Resilience Programme Evaluation

In this chapter I describe the UK Resilience Programme: the motivation for the project; the project background, set up and implementation; and the evaluation design. I also present descriptive results on the sample, attrition, programme adherence, and the cost of the intervention, as well as discussing the appropriateness of the measures used. This chapter serves to explain the context, evaluation design and measures used for the evaluation results reported in Chapters 2 and 3.

Introduction

There are high rates of mental illness in young people in the United Kingdom. In 2004, 11.5% of young people aged 11-16 had a clinically diagnosed mental health disorder (Green, McGinnity, Meltzer, Ford, & Goodman, 2005). Some authors have also found evidence of a recent increase in the rate of emotional problems through time (Collishaw, Maughan, Goodman, & Pickles, 2004; Rutter & Smith, 1995), and children's wellbeing appears to decline with age as they become teenagers (Rees, Goswami, Pople, Bradshaw, Keung & Main, 2013). Mental illness during adolescence is associated with impaired functioning; academic and interpersonal difficulties; poor health behaviours; poor labour market outcomes; and suicide (Covey, Glassman, & Stetner, 1998; Garrison, Jackson, Addy, McKeown, & Waller, 1991; Ialongo, Edelsohn, Werthamer-Larsson, Crockett, & Kellam, 1996; Bardone, Moffitt, Caspi, Dickson, Stanton & Silva, 1998; Fergusson & Woodward, 2002; Fergusson, Horwood & Ridder, 2005; Stansfeld, Clark, Rodgers, Caldwell & Power, 2011). But less severe and sub-clinical levels of symptoms may also interfere with functioning, and are associated with similar levels of impairment as diagnosable conditions (Gotlib, Lewinsohn, & Seeley, 1995).

Psychological disorders in childhood and adolescence are associated with mental disorders in adulthood (Kim-Cohen, Caspi, Moffitt, Harrington, Milne & Poulton, 2003; Reinherz, Paradis, Giaconia, Stashwick & Fitzmaurice, 2003; Roza, Hofstra, van der Ende & Verhulst, 2003; Fergusson et al., 2005; Stansfeld et al., 2011). For example, each episode of depression that an individual experiences predicts increased likelihood of recurrence, suggesting that preventing a first occurrence of depression could prevent a large part of the burden in later life (Harrington, Fudge, Rutter, Pickles, & Hill, 1990). Mental illness in adulthood accounts for 40% of all disability in the UK, and the estimated cost of anxiety and depression is £12 billion a year, including loss of earnings and expenditure on welfare benefits (Layard & Mental Health Policy Group, 2006). In adulthood, poor mental health is one of the best predictors of unhappiness, better than income, marital status or employment status (Layard, Clark, & Senik, 2012). Given the economic and social burden of mental illness in adulthood, prevention in childhood could be an important means to improved mental health over the lifespan.

Schools could provide a more effective access point to mental health services for young people than clinics, because of near-universal participation in education (Masia-Warner, Nangle, & Hansen, 2006). Even schools which strongly prioritise academic attainment should be motivated to promote access to mental health services, as psychological distress in adolescence has a significantly negative impact on educational achievement (Wilson & Marcotte, 1996; Fergusson & Woodward, 2000, 2002; Rucklidge & Tannock, 2001; Försterling & Binser, 2002; Biederman et al., 2004; Shahar, Henrich, Winokur, Blatt, Kuperminc & Leadbeater, 2006; Rothon, Head, Clark, Klineberg, Cattell & Stansfeld, 2009). Schools may even be able to prevent the incidence of mental health problems, through either delivering or providing access to prevention programmes.

A number of programmes have been developed to promote psychological wellbeing, social skills and positive behaviours, and help prevent or treat mild to moderate symptoms of depression, anxiety and other mental illnesses. Many of these interventions are suitable for use in schools (Durlak, Weissberg, Dymnicki, Taylor & Schellinger, 2011; Payton et al. 2008). However, relatively few of these programmes have been delivered at scale, using regular school staff and under ordinary conditions (Calear & Christensen, 2010). This is important: programmes which are effective in small samples with a high degree of control from the developers may not work at scale, and there may be considerable difficulties in disseminating programmes (Malti, Ribeaud, & Eisner, 2011; Weisz & Jensen, 2001). In addition, these small scale evaluations often only report programme impacts on a limited range of psychological outcomes, rather than the broader educational outcomes such as school attendance and academic attainment which are also of interest to schools and policymakers.

The evaluation of large-scale, real-world implementation of prevention programmes is an important next step in research on mental illness. For example, in the literature on depression prevention, the large majority of studies are efficacy studies, in which interventions are delivered by researchers or highly trained providers unaffiliated with the schools or community organizations for which these programmes are intended.

The few pragmatic studies, which use ordinary school staff to deliver interventions, have yielded smaller effects (Stice, Shaw, Bohon, Marti, & Rohde, 2009). A large-scale evaluation of a universal depression prevention program in 50 schools in Australia found no significant short- or long-term benefits on students' depressive symptoms (Sawyer, Harchak, et al., 2010; Sawyer, Pfeiffer, et al., 2010). Thus at least in the field of depression prevention, larger pragmatic trials may be an important direction for research.

Most secondary schools in England aim to develop pupils' noncognitive and life skills through explicit teaching in Personal, Social, Health and Economic (PSHE) education lessons (Formby et al., 2011; Ofsted, 2013). However, PSHE is not a statutory subject, meaning that there is no defined curriculum which schools have to cover and no prescriptions about how to teach it.[11] The content and delivery methods of PSHE vary, but the curriculum typically covers: emotional and physical health and well-being; sex and relationships; drugs, alcohol, and tobacco; personal finance; careers; and study skills (Formby et al., 2011; Ofsted, 2013). The quality of PSHE provision also varies, with a recent Ofsted report claiming that it was not good enough in a sizeable proportion of schools, and that this was associated with poor curricula and untrained staff (Ofsted, 2013). At its worst, inept attempts to teach sensitive and controversial issues resulted in the provision of inaccurate information and damage to pupils' emotional well-being, while other teachers omitted certain topics altogether because of a lack of confidence in their ability to teach it. The report recommended that schools should ensure that staff teaching PSHE received specialist training and ongoing support and CPD (Ofsted, 2013). In line with this, a mapping study carried out in 2009-10 found that the most effective PSHE lessons were taught by well-qualified staff, yet only 45% of surveyed secondary schools had any members of staff who held the national PSHE education qualification. Most staff said that it would be difficult to obtain funding and day release for PSHE training. The authors point out that at least 90% of staff teaching PSHE do not have a specialist qualification, and that this is quite unlike provision for other subjects for which the majority of teachers will be specialists at secondary level (Formby et al., 2011). Indeed, it is not currently possible to train as a specialist secondary teacher in PSHE (DfE, 2013d), so specialist training must come

[11] PSHE lessons are non-statutory, although sex and relationships education and careers education are statutory (DfE, 2013b; DfE, 2013c), and schools have a duty to prepare pupils for adult life (Ofsted, 2013).

after the period of initial teacher training in another subject.[12] Most PSHE teachers are therefore trained to teach a different subject, and may still spend most of their time doing this. A previous attempt to introduce a more structured PSHE curriculum through the Social and Emotional Aspects of Learning programme (developed from 2003) were not well thought out or consistently applied, and an evaluation of the programme in secondary schools found it had no impact on students (Humphrey, Lendrum & Wigelsworth, 2010). The Department for Education asserts that “Children can benefit enormously from high-quality Personal, Social, Health and Economic (PSHE) education” (DfE, 2010b), yet given the current quality of provision in many schools it may be the case that most pupil do not benefit much.

Thus we have a situation in which programme developers promote small-scale, unsustainable noncognitive interventions which schools could not replicate; while schools are obviously willing to teach noncognitive skills, and have scheduled time for this, yet often cannot find effective means of doing so. Faced with this, and a desire to promote wellbeing in schoolchildren, during 2006 Richard Layard of the Centre for Economic Performance conducted a search for an evidence-based wellbeing and life skills programme which could be used with a universal population of adolescents and delivered in UK schools. To be effective, interventions need to be sufficiently long; evidence based; integrated; consistently well taught; well-defined and highly-structured (Humphrey et al, 2010; Durlak et al, 2011). There were therefore a number of requirements that the programme had to fulfil: first, it must have been robustly evaluated in a universal population with promising results. Second, school staff should be able to teach it. This had two further consequences: to be taught by staff who were not specialists in mental health or life skills, there must be training available for them to be able to teach it effectively, and the programme would have to be sufficiently detailed and replicable to be used by a range of providers, not just programme developers or experts in the field. Third, the intervention should be intensive and long enough to be able to have an impact, but short enough to be fitted into the school curriculum. Fourth, it should be acceptable and appropriate for use in a universal population. And fifth, the programme must be commercially available. These requirements were meant to ensure

[12] The Ofsted review recommended that all initial teacher training courses should include specialist PSHE training, but even this would not amount to expertise in this area. It is possible to complete initial teacher training in citizenship.

that the programme had some evidence basis, and that it could be used widely in English schools.

The most feasible programme appeared to be the Penn Resiliency Program for Children and Adolescents (PRP), a group intervention developed by a team of psychologists at the University of Pennsylvania and trialled multiple times, but never at scale. The evidence on PRP is now mixed: Brunwasser, Gillham and Kim (2009) present a meta-analytic review of PRP’s effect on depressive symptoms from 17 studies, finding an average effect size of 0.11-0.21 (postintervention to 12-month follow-up) with considerable variation between studies. Most of these studies were conducted in the US, but five were implemented in China, Australia, Scotland and England. The ages of participants in the studies ranged from 8 to 18. The largest PRP study reports results for only 697 students. The current study improves on this with a larger sample than for the 17 past PRP studies combined. The sample is also larger than in many similar pragmatic studies (e.g. Calear, Christensen, Mackinnon, Griffiths, & O'Kearney, 2009; Malti et al, 2011; Stallard et al., 2012), with implementation conditions approaching ‘real world’ conditions, and a follow-up period of two years postintervention.

I evaluated the effects of a UK adaptation of PRP (UKRP) on 11-year-old pupils in 22 comprehensive secondary schools. Below I describe the context and participants; the design of the trial; the level of attrition; and descriptive statistics for the outcome measures. I also present indicators of programme dosage, and provide estimates of the cost of the intervention. In Chapter 2 I present the results for the programme impact on the main outcome measures, and in Chapter 3 I present the impact on pupils’ popularity. Although I am writing this first chapter primarily to describe the evaluation set up for chapters 2 and 3, some of the material is also relevant to chapters 3 and 4: the schools for which I have behaviour data are drawn from the same sample, and the data were collected in the context of the UKRP trial.

Context and recruitment

In 2006 Richard Layard of the Centre for Economic Performance at the London School of Economics (CEP) and Geoff Mulgan of the Young Foundation set up the Local Wellbeing Project, with the aim of understanding how local government agencies could promote wellbeing at a local level. They worked in conjunction with the Improvement and Development Agency (IDeA, now the Local Government Association), an organisation representing local government, to recruit Local

Authorities (LAs)[13] to the project. At this stage the project consisted of a number of proposed research strands and nine possible interventions. LAs could choose from among these which ones they wanted to be involved in. The chief executive and other staff of South Tyneside LA were most interested in the proposed resilience programme for school-aged children. LA staff were motivated in particular by a perception that pupils in their region did well at school up to the end of primary school, but started to fall behind the national average once at secondary school, performing poorly at age 16 and beyond. South Tyneside is a small, deprived, predominantly white working class metropolitan borough in the North East of England.

In order to test the programme at scale, in a variety of contexts and with heterogeneous populations, it was necessary to recruit other LAs to create a sample which was geographically dispersed and demographically varied. The heterogeneity of the sample was important to ensure that the trial would function as a pragmatic study (Rothwell, 2005). Having found one LA (South Tyneside) which wanted to run the intervention, the project sponsors approached five other LAs with differing demographic characteristics. Of these, Hertfordshire and Manchester LAs were recruited: see Table 1.1 for LA characteristics. Although taken individually, each of these LAs has a population which is substantially different from the average for England as a whole, taken together they are broadly representative in terms of ethnicity and income, and cover three different geographic regions. For example, in England as whole in 2001 91% of people were of white ethnic origin, while 4.6% were of Asian origin; in these three LAs 91% were white and 4.4% were Asian (ONS, 2003). The schools which were recruited from within these LAs were more deprived than the average for England, however. For more information on the characteristics of the evaluation sample and the populations they were drawn from see below.

The three LAs used a number of different funding sources to support the programme, including the Neighbourhood Renewal Fund and children's mental health budgets. Having recruited the three LAs, CEP proposed an evaluation to the Department for Children, Schools and Families (DCSF, now the Department for Education, DfE; the government department responsible for school-level education policy in England), which agreed to fund the research. The intervention was therefore organised as a collaboration between the three LAs, which funded and implemented the programme;

[13] Local Authorities are local governments responsible for delivery of certain public services, including a number of state-funded schools in an area.

the Penn research team, which provided training and support on a largely commercial basis; DCSF, which funded the evaluation; and staff at CEP who initiated and helped organise the programme, and designed, set up and carried out the evaluation.[14]

Schools

In December 2006 the LAs began to recruit schools from among their maintained (state funded) secondary schools. PRP was designed for use with children and adolescents aged 10-14, and it was agreed that UKRP would be trialled with 11-year-olds starting secondary school. Local authorities were paying the licensing and training costs of the programme; schools were expected to absorb the opportunity costs of teaching the programme.[15] LAs could only fund a limited number of places at the training. Schools enjoy substantial autonomy in England, subject to national education policies and LA oversight, and were under no obligation to participate in the project. South Tyneside invited all 9 of its secondary schools to sign up, of which 7 did so. Manchester invited all 22 of its secondary schools, of which 9 signed up. Hertfordshire invited a small group of schools in a single town, chosen because it was typical of the county as a whole on a range of factors, and because the six secondary schools in the town were geographically close and frequently worked together. Five of these six schools opted to take up the programme, with the remaining place taken by a school a few miles away and invited as a substitute.

All of the schools involved in the evaluation were state-funded, comprehensive (not academically selective), and accepted children from the age of 11 until (at least) the end of compulsory education at 16.[16] LAs funded the training for school staff on the understanding that schools would participate in the evaluation. This meant agreeing to survey staff and pupils as necessary. However, it also meant delivering the programme in a standardised way: teaching the curriculum to Year 7 pupils; teaching the full

[14] Staff at CEP included: Richard Layard, who initiated the programme and proposed the evaluation; John MacIntosh, who organised the early stages of the intervention; and Stephen Machin, who advised on the evaluation design and contributed some analysis and a small amount of text to the first evaluation report only (Challen et al., 2009). I designed and organised the evaluation, including data collections; secured funding for the evaluation; carried out the analyses for and wrote the DfE evaluation reports, and did the same for all other related papers mentioned in the preface. I also helped to organise the implementation of the programme and training from March 2007 until July 2008. Members of the Penn team also provided advice on the research design.

[15] This meant replacing 18 hours of PSHE lessons for the relevant classes, and (in effect) doubling staffing and rooms for these 18 hours.

[16] Some staff were trained who were not based in schools. These were LA staff who worked with children, such as counsellors and staff from children’s homes. Staff using UKRP outside of schools could do so however they wished, and these workshops do not form part of the evaluation.

curriculum (rather than picking out specific lessons); scheduling 18 hours of lessons; and teaching in classes of no more than 15 pupils. Without these requirements schools could have provided any part of the materials as they wished; this would have resulted in an 'intervention' which was ill-defined and not replicable, and arguably not suited to be trialled (Bouguen & Gurgand, 2012).

Intervention facilitators

The implementation and teaching of the intervention was undertaken by school staff rather than university researchers or mental health professionals. This was important, as the purpose of the study was to test whether the intervention could be delivered in the real world, with the staff and facilities ordinarily available to schools. Intervention teachers ('facilitators') were unlikely to be familiar with the intervention, with CBT, or with similar programmes prior to being involved with the UKRP, and required training. This greatly limited the number of facilitators and students who could be involved in the intervention.

Participating schools were allocated a number of training places by their LA, and were asked to find appropriate staff. Selection procedures varied by school, as did facilitator characteristics. In some schools senior managers offered places to the individuals they thought would be most appropriate, while in others all staff were invited to apply with senior managers selecting from amongst the applicants or randomly drawing out names to participate. A number of schools did not fill all their training places, so these were offered to other UKRP schools, or to LA staff who worked with children.

There were 73 facilitators who taught workshops which feature in this evaluation: 75% were female, and most were school staff, with the largest occupation group being teachers (Table 1.2). Teachers would ordinarily teach specific academic lessons such as English, mathematics, humanities etc. Other school-based staff included learning mentors, teaching assistants, and school nurses, while staff not based in schools were ordinarily employed by their local authorities in roles relating to children and families.

Students

All students who started in Year 7 at a UKRP school in September 2007 (the 2007 cohort) were intended to be included in the evaluation. In 13 schools the students who had started in Year 7 in September 2006 (the 2006 cohort) were also included in the evaluation, and were all in the control group. This extra yeargroup was included when the school was planning to include all their 2007 cohort students in workshops; the 2006 cohort therefore provided a control group for these pupils. In practice not all of these schools managed to include the full 2007 cohort in workshops, but the 2006 cohort was surveyed anyway to provide an additional control group (see Table 1.3).

All students within the 2007 cohort (grade) in participating schools were surveyed at baseline and at follow-up points, unless they were absent, had left the school, or if they did not want to complete the questionnaires or their parents had withdrawn them. The same was true for the 2006 cohort, except that they did not complete a baseline psychological survey.

It was decided to use the intervention with the Year 7 cohort for a number of reasons. First, PRP was designed for students aged 10-14 (Brunwasser et al., 2009), which would roughly correspond to pupils in Years 6-9 at English schools. Second, the programme developers recommended trialling the programme with 11-year-old children because they were likely to be old enough to have the necessary self-awareness to understand and use the skills, but were still young enough to have the plasticity to learn new ways of thinking and behaving (Thomas, 2012). Third, the project sponsors wished to use the programme in secondary schools, due to a perception that secondary schools were less likely than primary schools to be focusing on social and emotional learning.[17] Fourth, the transition between primary and secondary schools can be a source of stress for pupils, so providing extra support in the year immediately post-transition (Year 7) could help pupils when they need it most (Graham & Hill, 2002; McGee, Ward, Gibbons & Harlow, 2004; Galton, Gray & Ruddock, 1999). Fifth, pupils sat (now abolished) national tests at the end of Year 9 so schools were more likely to be willing to provide the time for UKRP in Year 7 or Year 8. Thus the Year 7 cohort was chosen based on a perception of need, appropriateness and feasibility.

Statistical analysis

I use three main methodological approaches in this thesis: a controlled trial with regression analysis; descriptive regression analysis with extensive use of control variables to understand the factors associated with the outcome of interest; and a natural experiment. Here I give a brief overview of the approach adopted for each chapter, but the precise statistical analysis used is detailed in the chapter itself.

[17] This view is backed up by more recent research, see Formby et al. (2011).

Chapters 2 and 3 evaluate the programme impact on a range of outcome variables, and adopt the methodological approach of a pragmatic non-randomised controlled trial, taking the intervention assignment to be arbitrary, and comparing outcomes for pupils who had participated in the intervention with those who had not. The impact of the intervention is calculated using regression analysis. Even with credible random assignment, using regression to control for covariates is still useful in evaluation. In particular, if intervention assignment is conditionally random, so that pupils are assigned based on their school and class membership and these groups have different probabilities of being treated, a raw comparison of pupil outcomes by treatment status may be confounded with the characteristics of schools with different treatment probabilities. Controlling for pupil and school characteristics will help to adjust to this, and I use specifications with school fixed effects throughout this thesis in order to account for stable differences between schools and between pupils at different schools. In addition, pupil characteristics may be related to the outcomes of interest. Even if they are not correlated with intervention assignment, including them will help to reduce the residual variance and increase the precision of the estimates (Ch. 2, Angrist & Pischke, 2009). In the UKRP experiment, intervention assignment is correlated with certain pupil characteristics (particularly prior academic attainment), so it is particularly important to include controls.

In Chapter 2, looking at the impact of the intervention on depression and anxiety symptom scores, behaviour scores, absence and attainment, I use a difference-in-differences methodology. This allows me to take into account the baseline level of the outcome variables and focus on the intervention impact on the change in these variables. I compare specifications which include pupil or school fixed effects, to account for stable differences between groups. As a pragmatic trial, I have included a number of outcomes of interest and importance to policy makers, academics and participants: although depression and anxiety scores could be difficult to interpret in terms of real impact, academic attainment and absence are easily understood outcomes of obvious importance to participants.

In Chapter 3 I examine the programme impact on the number of friend nominations a pupil receives. Here I do not have a baseline, so cannot look at the *change* in friend nominations following the programme and I cannot use difference-in-differences. Instead, I compare the outcomes between the intervention and control group at each point in time (simple differences), and include specifications with controls for pupil characteristics and school fixed effects. I can show that the intervention and control groups looked similar at baseline on a range of demographic factors, but I cannot show that the outcome variables were similar at baseline because I do not have this information.

Chapter 4, ‘Pupil behaviour in secondary schools’, is an extended descriptive chapter, exploring the incidence, distribution and stability of behaviour incidents in secondary schools using detailed panel data. I also examine how pupil characteristics are related to incidents, and test the consistency of my hypotheses with the data.

Chapter 5, ‘Day and time patterns in behaviour at secondary schools’, uses the arbitrary assignment of subjects and teachers to lesson slots as a natural experiment, observing the impact on pupils’ behaviour. Because of a lack of good academic outcome data I cannot show that this has an impact on academic attainment, but it is a suggestive piece of evidence which could inform an RCT to test such a policy.

Thus in three chapters I use two different strategies to understand causality (a trial and a natural experiment), and in one chapter I use descriptive methods to understand the nature of behaviour in schools. All four papers use regression analysis of different kinds. See the individual chapters for more detail on the statistical analysis used.

The Intervention: The UK Resilience Programme

Chapters 2 and 3 evaluate a single intervention on different outcomes. Here I describe the intervention and features of the controlled trial.

The Penn Resiliency Program

The Penn Resiliency Program (PRP) is a curriculum developed by a team of psychologists at the University of Pennsylvania (see Brunwasser et al., 2009, for a summary of the programme and previous PRP research). Its original aim was to prevent adolescent depression through techniques from cognitive behavioural therapy (CBT), but it now has a broader remit of building resilience and promoting realistic thinking, adaptive coping skills and social problem-solving in children (Reivich & Gillham, 2008). The UK Resilience Programme is the large scale UK implementation of the PRP.

PRP teaches participants to understand the link between thoughts and feelings, to develop coping strategies, and to practise them in the safe environment of the classroom. The curriculum includes a range of activities, including case studies, role plays, investigation and talking games, and worksheets. The programme is based on Ellis's *Activating-Belief-Consequences* model, which postulates that beliefs about events mediate their impact on emotions and behaviour (e.g. Ellis, 1962, 1977; Beck, 1967). The foundation ABC skill involves learning to separate out the facts of an activating event A, from the B, our beliefs about or interpretations of it. The key point of the model is that it is not activating events which cause how we feel and how we behave (the Cs); it is activating events *mediated by* our beliefs about them. An example of this is given in Figure 1.2 below.

PRP builds on this foundation skill to develop pupils' understanding of their thinking style and how this affects how they feel and what they do. The aims are accuracy and flexibility. If pupils can think more accurately and flexibly about difficult situations, they may be able to solve problems effectively. Using the example in Figure 1.2, a more accurate belief might be "I failed the exam because of the fractions section", leading to a different C: a less negative emotion and a decision to get more help with fractions. This interpretation allows the individual to accurately recognise where they have performed poorly and need help, but be less likely to think "I'm stupid" and give up on maths. By the end of the programme pupils should have developed a range of skills which they can use when their Cs (how they feel and behave) are hindering them. In this way, pupils are encouraged to identify and challenge unrealistic beliefs, to employ evidence to make more accurate appraisals of situations and others' behaviour, and to use effective coping mechanisms when faced with adversity. Participants also learn techniques for positive social behaviour, assertiveness, negotiation, decision making, and relaxation (Gillham, Reivich & Jaycox, 2008a, 2008b; Reivich & Gillham, 2008; Challen & Bailey, 2012).[18]

The UK Resilience Programme

The UK Resilience Programme (UKRP) is the first large-scale implementation of PRP.[19] It was first delivered as an 18-hour universal intervention in the academic year 2007-08. In subsequent years the programme was expanded, and it is now used in more than 85 schools and in non-school settings across the UK. However, only the first year of workshops, those conducted during 2007-08 in secondary schools, were involved in the evaluation and are the subject of this research. Approximately 2,000 students participated in UKPRP workshops in the pilot year, mostly new Year 7 students at the 22 evaluation schools. Schools were asked to ensure that UKRP was taught by trained

[18] See Gillham & Reivich (2007) for more information about PRP and related research.
[19] UKRP is also known as the UK Penn Resilience Programme (UKPRP).

staff; that intervention group pupils received the full programme over 18 hours of lessons; and that it was taught in classes of no more than 15 pupils. This last requirement meant that UKRP lessons required twice the number of rooms and staff ordinarily assigned to a class (usually one teacher and one classroom for 30 students). The PRP curriculum required minor adaptations for spelling, vocabulary and references to the UK context. Table 1.4 sets out the table of contents for the UKRP intervention.[20]

Resilience

The aim of the intervention was to promote positive coping strategies for school-age children. An important concept here is that of resilience: the ability to make positive responses and adjustments when faced with risk or adversity (Luthar, 2003, p4). The concept of resilience requires that an individual is coping and 'doing ok', and that they have faced some form of adversity. In the absence of adversity, we cannot know if an individual would be resilient or not (Coleman & Hagell, 2007). Teaching specific coping skills is one way in which UKRP could have an impact on pupils' outcomes, by adding to the protective factors pupils can draw on when facing difficult life events.[21] If students adopted and used the CBT-based emotional regulation, cognitive reframing and social problem-solving skills, this could enable them to deal better with any adversities and achieve better outcomes. In this case, I would expect any intervention impact on any of the outcomes to be mediated through use of the skills taught in the programme. However, there are other ways in which UKRP could have an impact. The small group size and subject matter could promote closer relationships between pupils and with the class teacher. Social relationships are important protective factors in dealing with adversities (Costello, Swendsen, Rose & Dierker, 2008; Hall-Lande, Eisenberg, Christenson & Neumark-Sztainer, 2007; Ladd, 1990; Lavy & Sand, 2012). UKRP pupils started workshops shortly after making the transition from primary to secondary school, and this transition can be difficult for some pupils (Graham & Hill, 2002; McGee, Ward, Gibbons & Harlow, 2004; Galton, Gray & Ruddock, 1999). Pupils might therefore benefit from additional opportunities to forge social bonds. Thus even if pupils did not use the UKRP skills, they might still prove more resilient and have better outcomes. In addition, if pupils enjoyed the lessons, this could improve their emotional

[20] For more details on the intervention content see Chapter 8 of Challen et al. (2009).

[21] Risk factors are conditions or attributes associated with a greater likelihood of negative outcomes; protective factors have the opposite effect, reducing the likelihood of undesirable outcomes in the presence of risk (Jessor, Turbin & Costa, 1998).

wellbeing directly, and possibly also their attendance at school.[22] These three possible mechanisms are not mutually exclusive. Coleman and Hagell (2007, Ch. 1) suggest three main strategies for promoting resilience in adolescence: reducing exposure to risk so that fewer negative events occur; mitigating the impact of negative events; and developing positive experiences and protective factors. UKRP might be seen to act mainly through the latter two strategies: mitigating risk through developing coping skills, and promoting protective factors through social relationships and positive experiences. Note also that not all of the skills taught by UKRP are coping skills formulated to deal with negative situations and events. Some, such as the prosocial, goal-setting and decision-making skills, aim to promote new positive outcomes rather than focusing on mitigating negative outcomes, and these could be included under the third strategy. It is not clear that most members of a universal sample of school children will face anything other than very minor adversities between the ages of 11 and 14, and even if they did, many will already possess protective factors to enable them to weather these successfully (such as supportive families). However, many children will still face a number of risks and adversities, albeit minor ones, and UKRP is not a ‘whole school’ programme, does not involve parents, and does not directly aim to change the world a child is in outside of the classroom.

Training

Since UKRP was to be taught by school staff, teacher training was essential for ensuring programme quality (Gillham, Brunwasser & Freres, 2008). Once selected, future facilitators registered and completed an online positive psychology program called Resilience Online.[23] This introduced them to the principles of CBT, and encouraged reflection on their own emotional responses and behaviour. Facilitators were trained in Philadelphia in the United States from 23rd July to 3rd August 2007 (the first two weeks of the summer holidays), receiving 10 days of training. Training was provided by members of the Penn research team and others they had trained, most of whom had backgrounds in psychology or education. The first week of training focused on learning about CBT and adult-level CBT skills, with days 6-10 spent familiarising trainees with the curriculum and practising delivering lessons. Further support was

[22] Pupils cite boredom as a major reason for truanting, while parents cite bullying and poor relationships with teachers (Malcolm, Wilson, Davidson & Kirk, 2003). Improving school on either of these dimensions could therefore increase attendance rates.
[23] For further information about Resilience Online see Abbott, Klein, Hamilton and Rosenthal (2009).

offered once facilitators had started teaching the programme, through participation in nine one-hour support conference calls with a PRP trainer and 10 other facilitators.

This project was intended to be a scaled-up, real-world intervention, trialling a programme which had previously only been implemented in small samples and with high levels of control from the developers. To truly reflect ordinary conditions, the programme must also be sustainable. Flying staff to the USA for training is not a sustainable feature, and all subsequent training has been conducted in the UK. However, there is no difference in the training materials or hours of training received in this first year of implementation and in subsequent years, and the decision to site the original training in the USA was due to financial reasons – unusually favourable exchange rates made this the cheaper option at the time. It therefore seems reasonable to believe that the first year of training was very similar to that received once the programme became embedded in subsequent years. As evidence of the practicality and sustainability of the intervention there are now 85 schools teaching it in the UK with over 800 teachers trained at 10 training courses. About 250 of these places were entirely funded by schools, with the remainder jointly funded by schools with LAs, public health organisations, or special grants, suggesting that the programme is affordable.

Intervention assignment

One major flaw of the trial is that allocation to intervention or control condition was not random. Schools and teachers are very resistant to randomising individual pupils, and timetable constraints on both teachers and pupils made this difficult. Randomising whole classes faced similar difficulties. LAs were also motivated by a need for cost-effectiveness: given the limited number of staff who could be trained in each school, they felt it was important to let schools teach as many workshops as possible. Moreover, the LAs involved were funding their own project: they were not keen on randomisation (they were not keen on having a control group at all), and it was not possible to insist: if they did not want to do the evaluation they could run the programme alone. By contrast, schools now applying for projects to the Education Endowment Foundation (EEF) will only receive funding on condition that they agree to participate in an RCT, or the project will not be funded.[24] This lack of control over the LAs meant that we had to reach a compromise on evaluation design, and it was agreed

[24] Even here, most EEF projects randomise at the school, class or yeargroup level, partly because of resistance by schools to randomise pupils in the same class to different treatments. This is despite an often critical lack of power to detect feasible effect sizes.

that schools would allocate classes to intervention arbitrarily, or use adjacent year groups as intervention and control groups. I describe below how this was carried out.

Within each school, classes of students from the relevant cohort were assigned to UKRP or a control condition. Schools were responsible for the assignment. In order for a class to be able to have UKRP workshops there had to be enough classrooms available at the right time (when the class was scheduled to have the lesson UKRP would replace e.g. PSHE), and there had to be a trained facilitator available to teach it at that time. Facilitators who were subject teachers had a busy timetable of other lessons to teach, leaving little flexibility in when they were available to deliver the intervention. Non-teaching staff could be more flexible, but they made up a minority of facilitators. There was also a very limited number of trained staff per school, usually only two. This meant that there were relatively few classes which even *could* have participated in UKRP. Most schools assigned as many classes as possible to the intervention based on these timetable constraints.

Intervention assignment was therefore not random, but, conditional on class membership, it was arbitrary and largely unrelated to student characteristics, particularly mental health. Students were assigned to classes and timetables were finalised by July 2007, before students joined these schools in September 2007. As a result, schools had limited information about students when they planned class membership and timetables: they would have known basic demographic and academic information, but would have had little information about mental health, for example. It would therefore have been difficult for teachers to select individual students for workshops based on perceived psychological need. However, the original assignment of students to classes could have been related to student characteristics, so using the class as the unit of condition assignment could result in differences in average student characteristics between the intervention and control groups. I will control for many student characteristics in regressions. In 15 schools the lesson to be replaced by UKRP was taught in mixed ability groups. These will have been designed to have a mix of pupils in terms of their observable characteristics such as gender, ethnicity, free school meals eligibility, and prior attainment.[25] This resulted in treatment and control groups which were largely balanced on pupil characteristics in these schools. In 3 schools the lesson which UKRP

[25] Most schools use software packages to assign new pupils to registration groups which are mixed in terms of ability and demographic characteristics, and also mix up pupils from different primary schools. However, teaching may take place in these groups or in lesson groups, which may be streamed or setted based on academic ability in that subject. Policies on grouping pupils vary by school.

replaced was taught in ability groups, but the schools were able to include classes of different ability levels in the intervention, so their intervention and control groups are balanced on academic attainment.[26] In the remaining 4 schools the lesson was taught in ability groups, and higher ability groups were selected for the UKRP intervention, leading to a mismatch in academic ability between the intervention and control groups (though there are no major differences in other characteristics between the two groups within each school). Of these 4 schools, 2 made a deliberate choice to include only the most able pupils in UKRP workshops, as they felt they would be better able to understand the programme and could spare the time away from other lessons. In the other 2 schools the classes which fitted the timetable slots just happened to be classes with more able pupils.[27] Because of these patterns of assignment the intervention group has slightly higher academic attainment than the control group in most samples. I include controls for prior academic attainment and demographic characteristics in my analyses in order to account for these differences. Since assignment into the intervention or control group was by class, all regressions use standard errors clustered by class grouping. Intervention assignment was therefore not random, but, conditional on class membership, it was largely arbitrary and unrelated to student characteristics.

In the absence of deliberate manipulation of treatment status it was hypothesised that this assignment could be arbitrary or 'as-if' random, allowing an accurate identification of the treatment effect through comparing control and treatment group outcomes. However, I know that two schools deliberately selected some students into the UKRP workshops because of concerns about their emotional well-being or behaviour, withdrawing them from their timetabled classes. I cannot be sure of the exact number, but there were approximately 45 intervention students selected in this way, most of them in workshops which started later in the year when schools knew more about their pupils. I include robustness checks excluding these pupils.

Additional control group

Hertfordshire and South Tyneside wanted their schools to provide the UKRP intervention to all their new Year 7 pupils who started the school in September 2007

[26] For example, one large school assigned new pupils into one of three broad ability bands when they joined the school, and all lessons were taught in classes formed from within these bands. The school selected one class from each band to participate in UKRP, so the intervention and control groups match on demographic characteristics and prior attainment.
[27] Note that not all the classes assigned to intervention in these two schools were high ability classes: several were average, but the presence of one or two high ability classes raised the average attainment levels in the intervention group.

(the 2007 cohort), and provided enough training places for this. This would have meant that there were no control group pupils available in the 2007 cohort, so it was proposed that they should use pupils who were already at the school as the control group. Pupils who started in Year 7 at these schools in September 2006 will generally have been very similar to those starting Year 7 in the 2007 cohort, and these were deemed to be an appropriate control group. The crucial assumption here was that adjacent cohorts in the same school are formed quasi-exogenously, and so can be viewed as the counterfactuals of one another when one cohort receives the intervention and another does not. Cohort membership is almost always determined by date of birth: over 99% of pupils in this sample are in the cohort predicted by their date of birth, and only 4 of 6,510 pupils (0.06%) appearing in the psychological dataset moved cohort during the evaluation. In addition, cohort membership is very unlikely to be manipulated for the sake of an intervention taking up only 2% of a student's Year 7 timetable, so it seems reasonable to treat cohorts as having been formed exogenously.[28] In general, the year-above cohort in the UKRP data does have similar characteristics to the workshop cohort within each school. However, it is important to compare pupils when they are the same age and at the same stage in their school careers, as these are factors which could have a sizeable impact on their psychological health and their friendship groups. Students in the 2006 cohort surveyed at the end of Year 8 in 2008 should therefore compared with 2007 cohort pupils surveyed at the end of Year 8 in 2009, for instance. The 2006 cohort is available in 13 schools as an additional control group. However, because of the timing of the evaluation set up it was not possible to obtain baseline measures for these pupils, as this would have required surveying them in September 2006, before schools had been recruited (see Table 1.5). This means that the data on this yeargroup cannot be used if the analysis uses difference-in-differences to assess the intervention impact, as this requires a baseline measure. However, I can still use this data if I am using differences in outcomes at a point in time, without taking account of the baseline value of the outcome.

In the event, only 7 schools included the entire 2007 cohort in the intervention, as scheduling workshops proved difficult. Thus I have data on psychological measures for 22 schools in total: 9 schools have data for only the 2007 cohort, and have a control

[28] Indeed, cohort membership is probably more plausibly exogenous than class group membership. The only cases in which adjacent cohorts within schools appear to be significantly different from each other occur when schools are in serious danger of closing down due to poor academic attainment and therefore quite quickly become seriously undersubscribed within a few years.

group within this cohort; 7 schools included the whole of the 2007 cohort in workshops, using the 2006 cohort as their control group and surveying both cohorts; and 6 schools surveyed both cohorts, and had control groups in both cohorts (see Table 1.5).[29] The 2007 cohort is split roughly equally into control and intervention students, while all students in the 2006 cohort are in the control group.

Spillovers

It is possible that by having intervention and control groups within the same school, we risk spillovers of the intervention into the control group (or vice versa). For instance, if intervention pupils spent much of their time with control group pupils, or if trained UKRP teachers decided to teach the curriculum to control group pupils, this would mean that the control group was partially treated. Alternatively, if the impact of the intervention is greater the more pupils are involved, having control group pupils in adjacent classes could diminish the possible impact. This is a version of the Manski reflection problem (Manski, 1993), and is a further threat to the internal validity of the experiment, although it would bias downwards the estimate of programme impact so it may be less of a concern.

Workshop groups

All students in the intervention group in the participating schools participated in the UK Resilience Programme workshops unless they no longer attended the school or if they moved to another (control) class for unrelated reasons. The intervention was incorporated into the school curriculum during the normal school day. Schools differed in the proportion of this new cohort they were able to include in the intervention, with the proportion ranging between 11% and 100% of the cohort (see Table 1.7). Seven schools included the entire cohort. The smallest number of workshop classes in a given school was two, and the largest number was 17. There was also considerable variation in the size of schools. The smallest school cohort (year group) contained about 60 students, and the largest 300, and 18 schools contained between 115 and 240 students. Because of these two factors, the number and proportion of workshop students and control group students differed between schools, and I control for school at baseline in many of our analyses to account for this and other differences between students at different schools. Despite this variation between schools, overall there are

[29] Since I had to organise a survey of pupils at specific times to obtain the psychological data, availability of this data was highly constrained. Attainment, absence, behaviour and demographic data were more likely to be routinely collected in school or national databases, so I have such data for more cohorts and more time points.

approximately 50% of students from the 2007 cohort in the control group and 50% in the intervention group.

Scheduling classes of 15 students usually meant doubling the staff and rooms needed for each lesson, which could be a challenge. As a result, most pilot schools ran fewer workshops groups than planned, while some increased the group size. Most schools chose to schedule UKRP during PSHE, citizenship or Learning to Learn lessons, as these best fitted the aims and content of the programme. Schools had been asked to find 18 hours in their students' class timetables for each series of workshops, but the length of each session and the time between sessions depended on the school's timetable. The majority of schools (17 of 22) had lesson slots of one hour duration, and so split the programme across 18 sessions. The frequency of sessions also varied by school. Eleven schools scheduled workshops fortnightly, with 10 scheduling them weekly and one school having three lessons every fortnight. When scheduled for one hour a week, UKRP lasted about 18 weeks, or half an academic year; scheduled fortnightly it would last for a full academic year.[30]

Alternative provision

Since schools had to make room for UKRP workshops within an already full curriculum, control group pupils received some lessons that intervention pupils did not. In most cases this was be 18 hours of the standard Year 7 PSHE curriculum, but some schools have displaced other lessons such as English, science or maths (see Table 1.8). One school was reorganising its timetable when beginning the project and created a new UKRP slot, meaning that there is no direct comparison in the control group. Class sizes for UKRP groups were not recommended to be larger than 15; in most cases the class size for the alternative treatment was around 30. It is therefore not possible to disentangle the effects of smaller classes from the impact of the curriculum. However, as the evaluation was designed as a pragmatic trial this is the appropriate comparison to make: the impact of UKRP against the standard school provision in this area.

Measures and data

There were three main sources of data: the psychological questionnaires administered between 2007 and 2010; schools' databases; and the National Pupil Database (NPD), which collects a range of data for all pupils at state schools in England.

---

[30] See Challen et al (2009), for more information on how UKRP lessons were scheduled.

Data availability therefore depends on when it was possible to survey pupils; what data each school was collecting and storing at each point for their own purposes (this varied between schools); and what data were standardly collected for the NPD at the time.

Data obtained from pupil and teacher surveys

Depressive symptoms. The primary outcome of the trial was symptoms of depression, and this is the outcome measure that was used in the power calculations. Depressive symptoms were measured using the Children's Depression Inventory (CDI; Kovacs, 2003), a self-reported symptom checklist. The scale is based on the Beck Depression Inventory (Beck, Ward, Mendelson, Mock & Erbaugh, 1961), adapted for children. It has been shown to be valid and reliable in measuring the severity of depressive symptoms (Kovacs, 2003; Reynolds, 1992), and it is the depression measure used in most previous PRP research (Brunwasser et al., 2009). Symptoms relate to negative mood, interpersonal problems, ineffectiveness, anhedonia, and negative self-esteem. Children indicate if they have experienced these symptoms over the previous two weeks. Each item is scored 0, 1 or 2 depending on the severity of the response: 0 indicating no symptoms of depression on that item according to the child's response; and 2 indicating strong symptoms. Item scores are then summed to create a total score, with higher scores indicating more or more severe symptoms. The full version contains 27 items; I omitted item 9 on suicidal ideation to ensure the acceptability of the scale for participants, so scores range from 0 to 52.[31] In tables of descriptive statistics I present the raw CDI score, while for the regression analyses I standardised the score to have a mean of 0 and a standard deviation of 1, using the baseline mean and standard deviation of the full sample (combined intervention and control groups). This allows for easier interpretation of the results in terms of effect sizes.

Anxiety symptoms. Anxiety was measured with the Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds and Richmond, 1985), a self-reported checklist of symptoms relating to physiological anxiety, worry, oversensitivity, social concerns, and concentration problems. The 28 items which measure anxiety are scored 0 or 1 (no/yes

---

[31] If more than 10% of items are unanswered then the assessment is considered invalid. When up to 10% of items are missing these scores can be replaced by the mean of the nonmissing items in order to create a total score. See the CDI Technical Manual for details on development and scoring (Kovacs, 2003).

responses to items), giving a score range of 0-28.[32] Higher scores indicate more severe symptoms, and the scale has been shown to have good reliability and validity (Reynolds and Richmond, 1985). As for the CDI score, I present raw RCMAS scores as descriptive statistics, but use the standardised variable in regressions to allow for easier interpretation of effect sizes.

Behaviour. Overall behaviour was measured using the self-report and teacher-report versions of the Goodman Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). The SDQ total difficulties score comprises 20 items, each scored 0, 1 or 2 according to the perceived severity of the symptom. This gives a minimum possible score of 0 and a maximum of 40, with higher scores indicating more (and more severe) symptoms. There are four 5-item subscales which sum to give the total score: emotional symptoms, conduct problems, hyperactivity/inattention, and peer relationship problems.[33] In addition, there is a 5-item prosocial scale.

Friends. Data on pupils' friendships was obtained from one page of the evaluation questionnaire, which asked pupils to list who their good friends were, along with each friend's form group and school to aid us in matching their names to codes. There were 24 blank lines to fill in names. A team of research assistants matched these names to codes using class lists from participating schools, and school lists for all other schools in each LA (obtained from the NPD). These codes can be used to match pupils into the NPD, and thereby obtain access to demographic and attainment data for all matched pupils listed as friends, not just those in the evaluation sample at UKRP schools. The codes are also used to match listed friends to evaluation data and generate the outcome measures used in Chapter 3: the number of times each pupil is listed as a friend by others; and the number of friends each pupil lists. See Chapter 3 for more details on these measures.

Other demographic characteristics. The pupil questionnaire booklet included several questions on who pupils lived with and on proxies for family functioning. These data are pupil-reported and somewhat noisy, and I make limited use of these variables in this thesis.

Use of UKRP skills. In order to measure the acceptability of the programme,

---

[32] The full 37-item scale includes a 9-item 'lie scale' which detects responses motivated by social desirability; this is not included in the calculation of the total anxiety score used as an outcome measure. See the RCMAS Manual for details on development and scoring (Reynolds and Richmond, 1985).

[33] The assessment is valid if at least 3 items of each of the four difficulties subscales have been completed.

UKRP facilitators and pupils were asked to fill in satisfaction surveys at the end of the workshops. These included questions on whether pupils used the skills taught in the intervention (See Chapter 5 of Challen et al., 2009, for details.) I use these data in examining mechanisms for change in the evaluation papers.

Data obtained from school or national databases

Absence from school. Annual absence from school is measured as the fraction of school sessions for which students were absent, as an authorised or unauthorised absence. A full school day consists of a morning and an afternoon session, so there are two sessions per day and at least one lesson per session. The data on absence was obtained from the NPD, which collects this data for pupils at every state-funded school in England. This reduces attrition for three reasons: first, I can obtain follow-up data for pupils even if they have left schools in the evaluation sample, provided they still attend a state school in England; second, pupils do not choose whether this data is collected or not; and third, data will be available whether or not pupils attend school.[34] I can therefore expect very low attrition and excellent coverage of the relevant population. However, for the programme evaluation in Chapter 2 I use a pupil's absence rate when they were in Year 6 (the last year of primary school) as the baseline value.[35] National collection of absence data in primary schools began in 2006-7, when the 2007 cohort was in Year 6, and the first year of data collection was imperfect. As a result there is no data available for 7 schools in one of the three participating regions (South Tyneside). Moreover, I cannot use the 2006 cohort as an additional control group because when they were in Year 6 (2005-06) the NPD did not collect absence data from primary schools. This limits the sample to 15 schools for this outcome measure.

Academic attainment. There are two sources for academic attainment data. Baseline academic data consists of the results of national tests which students sat at the end of primary school (Key Stage 2 tests). These are marked externally to the school, so results are comparable across students. I obtained these results from the NPD, and can therefore obtain results for many other cohorts of pupils as this data has been collected nationally for some time. Later years of academic data (postintervention and follow-ups)

---

[34] Most missing data on the other outcome measures is due to pupil absence or refusal to complete questionnaires; attrition as a result of pupil mobility is relatively minor in comparison.

[35] Intervention students began UKRP shortly after joining secondary schools, and absence data is only available by term, so if I want to use difference-in-differences methodology, Year 6 data is the latest baseline data I can use which will come from a period prior to programme participation.

are based on within-school teacher assessments which I obtained from school databases, rather than standardised tests. Teacher assessments use the same measurement scale as national tests. Only 20 of the 22 schools involved provided internal academic data of this kind.[36] For the evaluation, it is particularly important to have within-school control groups on this measure, as it is likely that standards will vary slightly across schools when assessing students (although teachers have clear guidelines on how to do this). I am able to supplement data from the 2007 cohort with data on students from the 2006 cohort, who had largely the same teachers and the same assessments. Because of these additional control groups, the intervention group used in the analysis of the impact on academic attainment is about the same size as those included in the analysis of the depression and anxiety outcomes, while the control group is much larger.

For Key Stage 2 tests, both the fine grained test score and a global teacher assessment are reported for each subject in the NPD. I use the test score as the primary indicator of academic attainment. If the score is missing, which accounts for approximately 2% of entries (e.g. because the pupil was absent on the day of the test), I replace the missing score with the level of the teacher assessment for each subject. Later assessments are entirely based on teacher assessments of pupils' working level within the secondary school. When using academic attainment as an outcome variable, I calculate the combined measure of academic attainment by separately standardising students' scores in each of English, maths and science, then taking the mean of these three scores.

The 20 schools which provided academic attainment data also provided data from developed ability tests which pupils sat when they entered the school (usually in September of Year 7). All 20 schools used one of two tests: the Cognitive Abilities Test (GL assessment, 2012); or MidYIS (CEM, 2013). I use this in the chapters on pupil behaviour as an additional control for academic ability.

Behaviour incident counts. Many schools were motivated to be involved in the UKRP because of concerns around pupils' behaviour. Although I attempted to measure pupils' behaviour through pupil- and teacher-reports on the Goodman SDQ, it is not clear that these will be a reliable indicator of behaviour for the majority of students, given the lack of reliability of self-reports, particularly for children (Dunning, Heath, &

---

[36] The remaining two schools refused on the grounds that it was too much work for their staff to extract the data.

Suls, 2004), and because the SDQ is primarily designed to screen for psychopathology so may not have good discrimination in the rest of the population (Goodman, 2001). I therefore attempted to collect information on pupils' behaviour from school databases. Schools are not legally obliged to record behaviour incidents in this way (DfE, 2012b), but many schools use software to record behaviour incidents, detentions, or other data relating to pupil behaviour. I managed to obtain incident count data from 7 out of 22 schools. Two schools refused to provide any data at all (including academic data); the remaining 13 schools all recorded behaviour in a systematic way, but either felt that the data was too sensitive and did not want to provide it, or had problems with outputting the data in a format I could use. Of the 7 schools which provided data, 4 of them were able to provide it with dates attached to incidents, and I use this data in Chapters 4 and 5. There was not enough data provided for the relevant cohorts to be able to carry out analyses of the intervention impact.

Pupil characteristics. I obtained information on pupils' demographic characteristics from the NPD. Including these as control variables should help to control for selection bias in intervention allocation, and improve the precision of estimates when the outcome variable is related to these characteristics. The values of many of these variables will change through time, so I use the baseline value from a pupil's first year at secondary school in all regressions.[37] Characteristics available from the NPD include gender; entitlement to free school meals (FSM); special educational needs (SEN) status; ethnic background; month and year of birth; and deprivation score of a pupil's neighbourhood. FSM functions as an indicator of poverty. SEN status is largely defined internally by schools, and can be applied to pupils with a range of characteristics including emotional and behavioural disorders; specific learning disabilities such as dyslexia; physical impairments; or generally low academic ability or attainment. It could be interpreted as indicative of the existence of some impediment to academic attainment. The threshold for being considered to have SEN appears to vary by school, so it is important to view it in context (Keslair, Maurin & McNally, 2011). There were over 40 different ethnic groups recorded for pupils in my sample. Since the large majority of pupils are of white British ethnic origin, with most other ethnic groups very small, I constructed five broad ethnic categories which were most relevant to the sample:

[37] When a value is not available for that year I use the value closest in time. It may seem unlikely that later changes in free school meals status could be endogenous to the programme, but changes in special educational needs status could be.

white; Bangladeshi and Pakistani; other Asian; black; and mixed or other. The source for the information on ethnic origin was pupils' parents in the majority of cases (89%). The neighbourhood deprivation score (IDACI score) is defined as the percentage of children in a pupil's postcode area living in poverty. All of these characteristics are associated with differences in academic attainment during compulsory schooling: girls generally perform slightly better than boys; pupils with SEN and FSM have lower attainment than those without; white pupils tend to have lower attainment than those from other backgrounds, particularly once FSM status is taken into account; and pupils who are younger within the cohort (born in summer) perform worse than older, autumn-born children (DfE, 2013e; Crawford, Dearden & Greaves, 2011). These characteristics are also associated with the incidence of mental health disorders, with girls experiencing more emotional disorders but fewer conduct disorders than boys; low income and poorer neighbourhoods associated with higher rates of disorders; white ethnic origin associated with an equal or greater incidence of disorders than other ethnic backgrounds; and SEN associated with a greater incidence of disorders (Chapter 4, Green, McGinnity, Meltzer, Ford & Goodman, 2005). Thus although these measures of demographic background are relatively crude, they are likely to be relevant to the outcomes I am examining.

Programme dosage. I collected information on workshop scheduling and class registers from school databases and individual facilitators' records. From these I calculated the number of hours scheduled for each UKRP class; the number of hours of classes actually attended by each student (available for most pupils); and the scheduled class size (the number of students assigned to the class). I use these to describe how schools managed to schedule the workshops, and as a proxy for intervention quality or intensity in estimating the programme impact.

Procedure

The trial was funded by the Department for Children, Schools and Families (DCSF, now the Department for Education), responsible for school-level education policy in England. Ethical approval was granted by DCSF. There was no attempt made at blinding UKRP participants (pupils or teachers) to condition allocation. However, control group students may not have been aware that students in the same school were receiving different lessons, and intervention group pupils may not have been aware that the questionnaires they were being asked to answer were associated with UKRP (it was an ordinary timetabled lesson). Consent for participation in the evaluation was sought from both parents and students. First, schools wrote to parents giving information about the intervention and evaluation and offering them an opt-out for the evaluation, using text agreed by DCSF. Very few parents chose to opt out. I cannot be sure of the exact numbers because schools managed this process and did not always report reasons for questionnaire non-response. However, I know that the withdrawal rate by parents was low: in 2007-08 fewer than 5 students were withdrawn by their parents; at 2-year follow-up this had increased to about 15 students. Students could themselves choose not to fill in questionnaires, and could decide on the day of the survey. There were many more withdrawals by students than by parents, up to about 5% of the cohort at any one time. However, the large majority of students chose to complete the inventories, as evidenced by the low attrition rate. As a result, there was very little selection generated by opting out of the evaluation, and the major reasons for incomplete questionnaires were student absence and mobility across schools. It is difficult to estimate the exact attrition rates because school registers were not perfectly accurate, particularly at baseline.[38] I calculate attrition statistics below based on the sample of students for whom I have baseline data, i.e. students who I know were in one of the UKRP schools at baseline and were available and willing to complete a survey.

Data collection

Data on the psychological state of pupils were collected through paper questionnaire booklets containing the outcome measures, plus some demographic questions. Pupils completed surveys in classrooms during normal lesson times, supervised by a teacher not involved with UKRP. Students had roughly one hour to complete the questionnaires. Once finished, students placed their questionnaires in envelopes and sealed them, and the envelopes were returned to me at CEP. Students were informed beforehand that their responses were confidential, but that if the researchers were worried about them they would contact the school. I obtained a named child protection contact for each cohort in each school prior to starting data collection. After obtaining the data, I contacted schools about students who had either: (1) scored 19 or higher on the CDI (indicating high levels of depressive symptoms; Kovacs 2003 pp. 66-70); (2) scored 20 or higher on the RCMAS (high anxiety symptoms; Reynolds & Richmond, 1985 p. 9); or (3) made comments about bullying, violence or other

[38] Pupils were new to these schools at baseline, and school attendance had not necessarily been finalised at that point. For example, in September 2007 several students were listed as being on roll at more than one school in the sample, while other students listed as being on roll never actually attended the school.

potential child protection issues. Schools responded to these concerns according to their own policies. Data on pupils' behaviour were also collected through surveys filled in by school staff, usually pupils' form tutors. Teachers could complete these whenever was convenient for them, but the surveying month was the same as for the pupil questionnaires.

Schools administered postintervention questionnaires to students (both intervention and control) within two weeks of finishing a set of workshops. Most schools began teaching UKRP in September 2007, and all had started by January 2008. This meant that the time between the baseline assessments and the start of workshops was short for the majority of the intervention group. However, 7 schools which finished their first series of workshops early in the academic year (February-March 2008) started a second set of workshops which lasted until June 2008. Students in these workshop groups therefore experienced a longer gap between the baseline measure in the autumn of 2007 and the start of workshops in February-March 2008. Because of these differences in starting date, duration, and timing, UKRP workshops finished at different times and so students completed the postintervention measures between February and June 2008. I know the dates when questionnaires were completed, and I include a survey month dummy in regressions using survey data outcomes in order to control for any seasonal, age or timing effects.

Power calculations

The unit of assignment to intervention was the class group. The required sample size was calculated to detect a postintervention effect size of 0.2 with power of 0.8, $\alpha=0.05$ (two-tailed), and an intra-class correlation in CDI scores of 0.06 and an average class size of 30. The effect size was based on the effect sizes from previous PRP trials (Brunwasser, Gillham & Kim, 2009) and the expectation that a pragmatic trial would have a smaller impact; estimates of the intra-class correlation came from pilot surveys in a non-UKRP school and with eight schools in the 2006 cohort. This suggested a need for 75 classes of 30 students, or 2,250 students in total. The size of school varied substantially, but the mean number of students per cohort was 179 (median 180), equivalent to a 6-form entry school (see Table 1.7). With 6 clusters per school, and a roughly equal split of intervention and control students within schools, this would require 13 schools for adequate power. However, some schools chose to include the whole of the 2007 cohort in workshops, effectively reducing the number of clusters within these schools to 2. Taking this into account I get an average number of clusters per school of 5. Since some schools also did not split the 2007 cohort equally into intervention and control groups, this will further reduce the power. To take account of this I assume that there are only four clusters per school, meaning that I would need 19 schools to detect an effect size of 19. I then calculated the same figures with the change in depression score as the outcome. This had an even lower intra-class correlation coefficient of 0.03, meaning that I would need only 13 schools for adequate power.[39] Thus using either methodology, assuming whole-school attrition was low, I would have adequate power to detect an effect size of 0.2 or larger. However, the study may have been underpowered to detect effect sizes smaller than this.

I calculated power based on the CDI score alone, as this was my primary outcome. Performing power calculations on primary outcomes only (i.e. not considering power for secondary outcomes) is standard practice (for an example from a similar trial see Stallard et al., 2012), and indeed is recommended as best practice in the CONSORT 2010 guidelines (Moher et al., 2010). One reason for this is that with multiple outcome variables the chances of obtaining a ‘favourable’ result on at least one outcome become very high (Bland, 2009). This would mean that even small samples would appear to have sufficient power according to standard benchmarks, yet we would still be unsure whether the intervention had any effect on each individual outcome, and trials would usually be underpowered on the outcomes of primary interest and concern. Note that because this would result in a smaller sample size, power would be reduced for both primary and secondary outcomes, and so identifying the true intervention impact on *any* outcome would be made harder. Failing to define a single primary outcome in this way also leaves room for the possibility of changing the primary outcome and the focus of the trial once outcome data are available, resulting in misleadingly optimistic reported results (Chan, Hrobjartsson, Haahr, Gøtzsche & Altman, 2004; Chan, Krleza-Jerić, Schmid & Altman, 2004). Since the primary aim of the project was to promote pupils’ wellbeing, taking the CDI score as the primary outcome and recruiting enough schools and pupils to have sufficient power to detect an impact on this measure should be appropriate.[40]

---

[39] When looking at the change in the depression score, I have to exclude individuals with no baseline data. Seven schools have no control group with baseline data. However, because of the very low intraclass correlation coefficient in the change I could exclude them from analyses and still have sufficient power for the outcomes of interest.

[40] It would also be possible to require a sample size to provide sufficient power to find an impact on *all* outcomes, primary and secondary. However, this would be very demanding in terms of cost, and would only be appropriate if the programme were only of interest if it could have an impact on all outcomes.

I did not adjust the p-values for multiple testing, despite intending to look at impact heterogeneity with respect to demographic characteristics, and having a number of secondary outcomes to examine as well. One major reason for this is that in a pragmatic trial we will often be less concerned about type I error: if comparing a new intervention against current practice, our main concern is to avoid rejecting a more effective programme. If two programmes are in fact equally effective, we should not be too concerned if one appears to be better, as the two are interchangeable for practical purposes and it will not matter if one gets used more as a result (Schwartz & Lellouch, 1967).[41] Another reason is the messy control design, with different control groups available for different schools, and different proportions of the cohort included in interventions – these factors make it difficult to accurately calculate the required power.[42]

Descriptive results

I give an account of the intervention impact on the major outcome variables in Chapter 2, and on pupils' friends in Chapter 3. Here I present information on the sample; attrition; workshop adherence and quality; and intervention cost; as these factors are relevant to both evaluation chapters.

Sample

Table 1.9 presents the demographic characteristics of my sample of schools, comparing it to the full population of pupils at state secondary schools in England as a whole, and to the school population in LA areas from which the schools are drawn.

Since the purpose of this project was to improve pupil wellbeing, requiring sufficient power for all outcomes would not match the scope of the intervention.

[41] This only applies to effectiveness – one reason it might not apply is cost: if UKRP is more expensive than ordinary PSHE lessons, then we have a different reason to want to know whether they are equally effective or not.

[42] It would be possible to perform power calculations for the analyses carried out in Chapter 4 and Chapter 5. However, I have not seen power calculations performed for descriptive analyses or natural experiments. Insufficient power in these cases would produce problems with false negatives, not false positives, and it is the latter I am primarily concerned with when considering biases which would undermine my conclusions. (By contrast, when evaluating an intervention which has had a substantial impact in the past, we should be concerned about both false negatives and false positives.) Since in descriptive statistics such as are presented in Chapter 4 I do not pretend to enumerate *all* the pupil characteristics which have an impact on behaviour, only those for which I have available data and which have the most sizeable impacts on behaviour, I do not face major problems of bias. In Chapter 5 I find a sizeable impact of day of the week and time of day on behaviour, and the statistical significance of these would only be biased downwards by insufficient power. Since again I do not claim to be able to enumerate all the scheduling factors which could contribute to behaviour patterns, a lack of power would not undermine my results. For these reasons I do not report power calculations for these chapters.

Note that the LA population figures shown in Table 1.1 may differ from the statistics shown here: Table 1.1 presents figures on the whole population resident in the LA, while Table 1.9 only shows figures for children of a specific age range in state secondary schools. The demographic characteristics of pupils in the three LAs vary substantially, and the characteristics of pupils in my sample largely reflect these differences. The pupils in my sample are much more likely to be eligible for free school meals than the average for England, and live in areas with a larger fraction of deprived children. They are also more likely to be deemed to have special educational needs by their schools. In two regions pupils are much more likely to be from a white ethnic background than the average for England, but overall the sample is more likely to be from an ethnic minority background than the national average because of the large number of pupils from Manchester. Attainment at Key Stage 2 does not appear very different between my sample and England as a whole. There are three reasons why my sample is more deprived than the average for England: two of three LAs are above averagely deprived; the schools which chose to participate in the intervention are more deprived than the average for their LA; and schools from the two deprived LAs were larger than those from Hertfordshire. However, an important feature of the sample is its diversity: the three areas schools are situated in are very different, so for example, pupils in Hertfordshire are substantially less likely to be eligible for FSM or live in a deprived area than the national average. Thus although the sample over-represents deprived children, it includes pupils from a range of backgrounds as intended in the recruitment strategy.

Attrition

Figure 1.1 presents a flowchart of the recruitment and retention of schools and pupils in the evaluation, the main results of which are reported in Chapter 2. The attrition figures presented here are based on the observations used for analysing the depression symptoms score; other outcomes will have different data availability.[43] There were four main points for the collection of psychological measures: at baseline (September-October 2007); postintervention (between February and June 2008, depending on the school); 1-year follow-up (June 2009); and 2-year follow-up (June 2010).[44] All 22 schools which undertook baseline assessments in autumn 2007 remained

[43] Since the analyses on the depression symptoms score exclude any observations without a baseline, the 2006 control cohort is excluded from this diagram.
[44] Some questionnaires were completed late, due to student absence or poor school organisation. The

in the trial until 2-year follow-up.

Table 1.10 presents information on attrition for different outcome measures. The depression and anxiety scores are included together here because their data availability overlaps almost exactly. In descriptive statistics and regression analyses I include any student for whom I have sufficient data, i.e. at least one measure at baseline and at one follow-up point when using difference-in-differences, and at least one follow-up measure when using simple differences. The number of observations is therefore limited by non-response or unavailability of data but not by any other selection on pupils included in regressions. The first rows of Table 1.10 give the source of each outcome; the number of students who have a baseline measure available; the number and percentage of students who have a baseline measure and follow-up data in at least one period (i.e. the number included in analysis); and the number of schools for which I have data. I use the number of students for whom I have a baseline measure as the basis on which to calculate attrition for each outcome. Here we see that I have data for 22 schools available for the depression, anxiety and teacher-reported behaviour scores, thus attrition on these outcomes was at the student level. Response rates were generally very high on these measures: I am able to include at least 97% of students who had a baseline measure in the analyses. There is also very little attrition for absence from school and academic attainment, conditional on having baseline measures. However, I do not have information on absence for one entire region due to reporting problems: this leaves data for only 15 of 22 schools, or 67% of pupils in the 2007 cohort. Likewise, two schools refused to provide any academic attainment data, meaning that I am missing this outcome for about 350 students, or about 10% of the 2007 cohort.[45]

The four lower panels in Table 1.10 present results from regressions with dummy variables for availability of data at each time point on the left hand side, and a dummy variable for treatment on the right hand side, with and without school fixed effects. This is to check for differential attrition between the intervention and control groups. Here we see that intervention pupils are slightly less likely to have missing data, particularly for the psychological outcomes and for academic attainment. These differences

large majority were completed on time: at baseline, 89% of questionnaires were completed by the end of October. Note that for students in UKRP groups which completed the curriculum earlier in the year, the 1-year follow-up point would take place more than one year after postintervention. Thus 1-year follow-up actually represents time periods of 12-16 months follow-up, and 2-year follow-up 24-30 months. I control for month of survey to take account of these differences.

[45] I have not included attrition statistics for the friendship measures here, because there was no baseline available for them. However, they were collected as part of the psychological questionnaire and response rates will be very similar to those reported here for the depression symptoms score.

becomes statistically insignificant once I control for pupils' academic attainment, suggesting that higher attaining pupils are less likely to have missing data, and intervention pupils are more likely to be high attaining. Attrition for the questionnaire data (depression, anxiety and behaviour) increased through time. This was due to schools becoming less willing to chase up questionnaires through the course of the project. By contrast, attrition for the academic data appears to decrease through time, largely because most schools will assess students at the end of Year 9 at the time of the two-year follow-up (and in previous cohorts students would have sat Key Stage 3 exams at this point), while assessment timing varies by school prior to this.

Programme adherence, completion and attendance

This is a pragmatic trial, so I do not exclude any pupils, workshops, facilitators, or schools from the main analysis simply because the intervention was not properly implemented: I want to know what impact the intervention has under normal conditions. However, given the acknowledged problems around scaling-up interventions (Malti et al., 2011; Weisz & Jensen, 2001), any measures which help us to understand *why* a programme works or not could be very useful. Thus measures of programme quality are relevant and could inform interpretation of the results. I was not able to get an accurate measure of curriculum completion, or of teacher absences, but I do have information on the amount of time schools allocated to the programme for all workshops, and on student absence for most workshop groups. I also have data on class size. Combining these measures I can give an indication of programme quality or dosage.

Almost all students in the intervention group in the participating schools participated in the UK Resilience Programme workshops unless they left the school or changed classes for other reasons, as the programme itself was incorporated into the school curriculum and was not optional.[46] Only in a very small number of cases were students allowed to opt out of lessons. There was therefore very little attrition from the intervention. However, not all classes will have covered the full curriculum, because of teacher absence, teaching speed, or difficulties with student comprehension.

Table 1.11 and Table 1.12 give details of the class size and duration of UKRP workshops. These indicate how closely schools stuck to the programme developers' guidance that 18 hours of class should be available for UKRP, in classes of no more

[46] Although note that participation in the evaluation was optional.

than 15 students, taught by trained staff. The amount of time scheduled for workshops ranged between 8 and 25 hours. Of the 12 groups with less than 8 hours available, 10 of these were in one school with poor planning; one group started UKRP but had abandon it because of timetable changes which meant that the facilitator was no longer available; and one was taught by a facilitator who did not like the programme and did not want to finish it. Over 70% of workshop groups were scheduled for 16-18 hours. When slightly less than 18 hours was available this was usually because of external constraints such as public holidays, training days, school trips or other special activities which interrupted the usual timetable. Of the workshops with more time available, in some cases this was due to extra time being available in the schedule, but more often time was made available in compensation for slow progress, usually for pupils with low academic attainment.

Apart from the time schools allocated to workshops, students may not have received the full course due to their absence from school or due to changes of class or school. The workshops were scheduled during ordinary school days, so we would expect attendance to be high as students of this age must attend school. I have data available on the attendance of 1,680 students (86% of intervention students in the evaluation) from 127 of 146 workshop groups (87% of workshop groups). From this sample, I find that 75% of students attended less than 18 hours of workshops, although 66% of students received at least 16 hours. Of those receiving less than 16 hours of workshops, half were in groups which had fewer than 16 hours scheduled, so that any individual absences would have resulted in low programme hours.

The size of UKRP workshop groups also varied. Since most secondary school classes have approximately 20-30 students they usually had to be split in two to form workshop groups of 10-15. Group size ranged between 5 and 26, with 81% of workshop students in a group of 15 students or fewer. Of the students in groups of more than 16, more than half of these were accounted for by one school. Clearly, if absence rates were high, effective class size would have been lower than the class size based on enrolment numbers.

Overall, only 19% of students actually attended at least 18 hours of lessons in classes of 15 or fewer (Table 1.13). However, the majority of students were in classes of approximately the right size, and the majority received 16-18 hours of workshops, broadly in line with the developers' recommendations. Programme hours compare favourably with previous trials of the PRP curriculum. For example, Gillham et al. (2006) report 45% of students attending at least 88% of sessions, while in this study 66% attend this amount. The major reason for this may be that the programme was scheduled during normal school lessons, for which most students have high attendance rates. Very low programme dosage was largely due to schools not providing enough time for the lessons, or due to teachers stopping the lessons or deliberately manipulating the treatment assignment.

In all analyses I use an intention-to-treat approach and do not account for the size of workshop groups, or the time students received.[47] This is important: the aim was to test the impact of the programme as it would be on the ground, and problems of timetabling and absences are likely to reduce the impact of interventions rolled out in this way (Roland & Torgerson, 1998; Angrist & Pischke, 2009). However, I do show results which use these factors as measures of programme quality or intensity to see whether they are associated with programme impact.

Table 1.14 presents data from the pupil satisfaction survey on pupils' self-reports of whether they used the UKRP skills: 49% of pupils in the evaluation sample who responded said that they did. It is not clear whether pupils' self-reports on this are reliable, but I use this measure in the evaluation chapters to look at whether it is associated with programme impact, which could provide indicative evidence of the mechanisms of intervention impact.

Facilitators and workshop groups

Most facilitators were not well informed about the programme, and few schools had worked out exactly how they would schedule UKRP lessons before sending staff to the training in July 2007. Partly as a result of this, the majority of facilitators (70%) taught only one or two workshop groups in this first year of the programme (Table 1.2).[48] 90% of workshop groups were taught by a single facilitator, but some did not feel confident teaching workshops alone, and so team teaching was used. This was often the case with support staff who were not used to working with larger groups of pupils. It is not clear that this would affect the quality of the intervention, but it clearly has an impact on cost effectiveness if trained staff teach only a small number of classes and do

[47] That is, I treat pupils who were assigned to the intervention group as having received the intervention, and pupils assigned to the control group as being control group pupils, regardless of whether or not they actually received the intervention.

not teach alone. There was only one untrained facilitator who taught one workshop group.

Appropriateness of measures

The trial was intended to be a pragmatic trial of a resilience programme in a universal population. Ideally, the main outcome measure should relate to the purpose of the intervention; should be practically meaningful to participants; and suitable for use in this population (Thorpe et al., 2009). However, the measures used in the evaluation may not fully satisfy these criteria. First, it is not clear how to measure resilience. As discussed above, to be able to observe resilience we need to be able to see how individuals are coping (or performing, or feeling), and we need to know that they have faced some adversity. Without knowing about adversities, we cannot know if an individual has been resilient or not, as it is possible that they are coping well simply because they have not faced anything particularly challenging (Coleman & Hagell, 2007). I have good data on a number of different measures of emotion, behaviour and performance, so I can observe the part of resilience relating to outcomes. However, I do not have good information on the challenges which pupils have faced, and this data is difficult to obtain even for adults, with self-reported negative life events an inadequate proxy measure (Uher & McGuffin, 2010). I therefore cannot distinguish between good outcomes for pupils who have responded well to adversity, and outcomes for pupils who have simply not faced many challenges. One way I address this is by looking at outcomes by subgroups of pupils: if pupils with SEN, low attainment, or who are eligible for FSM are more likely to be facing challenges than other pupils, looking at the outcomes for these pupils specifically can be seen as a way of asking whether the intervention has had an impact on pupils who have faced adversity.

One alternative would be to ask pupils to complete scales which ask them to report on their own resilience. However, the evidence on the meaningfulness of these scales is not convincing. A review identified only one scale which had been adequately tested for use with adolescents and which appeared to have acceptable psychometric properties (Ahern, Kiehl, Sole & Byers, 2006). Moreover, the concurrent validity of this measure was tested against other self-reported factors such as life satisfaction and depression: there was no attempt to judge it against more objective measures of coping

[48] Some school-based facilitators were trained but taught no workshops at all in the first year, and they are not included in this evaluation.

or effective functioning in the face of adversity. Since children are not likely to be better witnesses to their own behaviour than adults are about theirs (Dunning, Heath & Suls, 2004), it is not clear to me that using this measure would represent an improvement over measurement of outcomes, particularly given a positive response bias towards giving socially desirable answers (e.g. Reynolds and Richmond, 1985).

Given this, I am left with the measurement of pupils' outcomes: assuming that arbitrary assignment to intervention worked correctly, we can assume that pupils in the intervention and control groups were just as likely to have experienced adverse events, and so we can hypothesise that differences in outcomes indicate improved resilience in the intervention group. However, there are other potential problems with the outcome measures available, notably that they may not be suitable for identifying change throughout the population. Most previous research on the PRP curriculum has used the Children's Depression Inventory score as the primary outcome (Brunwasser et al., 2009), and this was the reason for choosing it as the primary outcome of this study. But it has some limitations as an outcome measure in a universal population. First, the CDI was designed to assess symptoms of depression, so primarily measures deviations from wellbeing, with no attempt to assess an 'upside' of positive wellbeing. Most pupils of this age in a universal population do not have many symptoms of depression (Green et al., 2005), and as a result the distribution of scores is highly skewed (see Figure 1.3). One challenge this presents is a non-normal distribution in the outcome variable (skewness = 1.32 at baseline); this can be addressed through a square root transformation (baseline skewness = -0.02), and results are robust to using this alternative outcome variable.

More critically, I find a strong ceiling effect, with 11% scoring 0 or 1 on the CDI at baseline, and 54% scoring 7 or lower, out of a maximum possible score of 52. Thus although this might be a good outcome measure for pupils who either have high scores at baseline or who are at risk of developing these later on, and can certainly discriminate between pupils with symptoms of depression and those without (Kovacs 2003) it cannot discriminate between (for example) pupils who are not depressed but not happy, and pupils who are not depressed and happy. The same is true of the RCMAS (anxiety) score and the Goodman SDQ scores: all these scores are skewed and face a strong ceiling effect (see Figure 1.3). Even the least skewed score (the pupil-reported SDQ score), may not be a particularly good measure of behaviour: its relationship to behaviour incident counts is relatively weak, and is in fact no better a predictor than the self-reported anxiety score (see Chapter 4). Moreover, 11 is the lower age limit recommended for its use, and its reliability may not be entirely satisfactory in younger children in a community sample (Muris, Meesters, Eijkelenboom & Vincken, 2004). The Goodman SDQ includes a prosocial scale which is not included in the calculation of the main SDQ score, and which measures positive aspects of prosocial behaviour. I use this as an additional evaluation measure, but this too suffers from stark ceiling effects: 26% of pupils get the top score at baseline on the teacher-reported scale, and 17% on the pupil-reported scale. Thus all four of the main psychological outcome variables may not be ideal for many pupils without any risk of at least moderate symptoms.

One way of compensating for this is by looking at more outcomes, such as academic attainment and absence from school (Figure 1.4). Absence is highly skewed, with most pupils having very low absence rates. In addition, I am missing baseline absence data for one whole region. The distribution of academic attainment in Key Stage 2 looks more normal, but there is a long left tail and a strong concentration of marks in the middle of the distribution, so it may still be the case that we are better able to discriminate between pupils at the bottom of the distribution than at the top. Later values of the academic attainment measure are based on teacher reports. These may be more subject to bias, to reporting differences between schools, and to a compression of the distribution – teachers may not be as willing to award extreme marks as would be obtained on tests and this will reduce my ability to discriminate between pupils' outcomes (Gibbons & Chevalier, 2008).

Chapter 3 looks at the impact of the programme on another outcome, and this may be more appropriate: pupils' friendship nominations. This measure could relate to resilience in two main ways: having more friends could indicate that a pupil is exposed to fewer negative peer interactions (reduction of risk); or by promoting positive social experiences both as a protective factor, and as being valuable and enjoyable in themselves. The main outcome (the number of times a pupil is listed as a friend) is skewed (skewness = 1.04 at end of Year 7), but does not present significant ceiling effects, and looks much more like a normal distribution. Indeed, it may present slight floor effects, as I cannot discriminate between pupils who do not have any friends and those who are actively disliked. At the positive end of the scale there is significant variation, and this variable does in fact have a long right tail. Thus this provides me with an outcome measure which is a good measure of social functioning and acceptability; and provides good discrimination between pupils at the positive end of the scale.

Table 1.15 presents linear correlation coefficients pairs of outcome variables. The pupil-reported outcome measures are highly correlated, but correlations between these and the other outcomes are weaker. In particular, absence and friendship nominations are only weakly associated with the other outcomes.[49] Thus even if the main outcome measures are not entirely suitable for identifying change in this population, in combination they may cover all parts of the distribution and may be identifying different aspects of good functioning. Ideally, outcomes should also be practically meaningful to participants. The CDI, RCMAS and SDQ scores may not fulfil this criterion: it is not clear how to value changes in these scores, or what they would mean in practice. However, changes in academic attainment, absence from school, and number of friends are arguably comprehensible and important outcomes to participants: it is clear that an improvement of two half days in absence rates means that pupils will attend one day extra of school per year, for instance. Whether there is any impact on these outcomes, and whether these impacts are of a magnitude to be practically significant, are different questions – the important thing here is that participants should be clear what these outcomes mean. One strength of the study is that I have a range of outcome measures covering different domains, and that these are obtained from different sources such as pupil-reported, teacher-reported and routinely collected information from databases. If primarily interested in subjective wellbeing, pupil-reported data may be the best source, as individuals are the ultimate judges of how they feel. However, for other factors such as absence or behaviour, individuals, particularly children, are poor judges, and it may be better to use database information. Overall, the use of data covering a range of outcomes and from a variety of sources allows me to explore impact in a number of areas and promotes confidence in any observed outcomes.

Limitations and extensions

There are a number of limitations to the design of this evaluation. The most serious is probably the lack of randomisation. Even if the arbitrary assignment of classes to intervention was truly arbitrary (or conditionally random), and even if the results obtained were identical to those which would be obtained after randomisation,

---

[49] Table 1.15 reports linear correlation coefficients, but using Spearman rank correlation coefficients to avoid imposing linearity generates almost identical results.

the advantage of externally-imposed randomisation is its credibility: if well carried out, randomised designs constitute the most believable identification strategy to uncover programme impact (Angrist & Pischke, 2009). With non-random assignment we can always ask if there was in fact some bias in the allocation to intervention which is generating the observed outcomes. Related to this is the messy control design, which makes it difficult to accurately calculate power: there are two different control groups, one of which (the 2006 cohort) lacks the baseline data required for certain analyses. This is a result of substantial variation in the number of pupils schools were able to provide workshops for, which appears to be related to how committed schools were to the intervention, or how organised schools were. Chaotic schools either managed to teach only a very few workshops, or taught more by using large classes, reduced workshop hours, and some untrained staff. Nevertheless, this probably reflects how schools would use the programme on the ground, and the existence of a control group within every school should allow me to control for the unobserved differences between schools. This may be a problem when using difference-in-differences, because the only control group available for 7 schools is the 2006 cohort, and these pupils cannot be included because they do not have a baseline. However, excluding these schools does not change the main results (Challen, Machin & Gillham, forthcoming). There was little choice over the set-up, given the degree of control which the programme implementers had. Nevertheless, the design is messy and this makes any results more difficult to believe.

I have discussed above the possible inadequacies of the outcome measures. Given the same intervention and the same population I am not convinced I could have found better measures. However, there are certain circumstances which have combined to make the available data further suboptimal. First, I was not able to collect baseline measures for a large part of the control group, because the evaluation set up occurred too late.[50] Second, I was planning to use Key Stage 3 tests for a two-year follow-up measure of academic attainment in English, maths and science for all schools. Like Key Stage 2 tests, Key Stage 3 tests were national exams set and marked externally to the school, but taken at age 14 when pupils were at the end of Year 9. They are likely to be superior to teacher assessments because they are based on tests sat under strict conditions, and because of the uniformity of the exams and the anonymous marking, standards are unlikely to vary between schools and across pupils (Gibbons & Chevalier,

[50] I started working on the evaluation in March 2007.

2008). However, in October 2008, the Department for Children, Schools and Families announced the abolition of Key Stage 3 tests, which means that none of the 2007 cohort took these tests (BBC, 2008). As a result I have to rely on teacher reports for academic outcomes at this point. The next point at which pupils standardly take national exams is at the end of compulsory education at age 16, when pupils sit their GCSE exams. This happens at the end of Year 11, or at about 4-year follow-up, and these measures are available through the NPD. However, results were only available in March 2013, which was not soon enough to include them in the analyses. Thus I am unable to include any academic follow-up measures from national exams, and have to rely on teacher reports.

Other outcomes of interest include the proportion of students staying in education beyond compulsory education (post 16); and labour market outcomes. Data on students staying on at school collected through the NPD would tell me which students had stayed on into further education. However, information on whether the 2007 cohort has stayed on to sixth form has only just become available. A level results will be available for much of the 2007 cohort in April 2015. It is currently not possible to link the identifiers used in the education databases into those for later labour force outcomes. This is unfortunate, especially since evaluations of other noncognitive interventions have found longer run impacts on academic attainment and employment, despite having no longer run impact on measured IQ (e.g. Heckman, Stixrud, & Urzua, 2006). However, looking at longer run educational outcomes at GCSE, staying on beyond 16, and A level will be possible, and I am prevented from doing so only because of the age of the pupils involved in the intervention.

One limitation of data available from the NPD is the paucity of information about family background. The circumstances in which a child lives, and particularly the quality of their relationships with their family, are likely to be related to their wellbeing (Rees, Goswami, Pople, Bradshaw, Keung & Main, 2013). I included some questions on family background on the psychological questionnaire in order to obtain more information on who pupils lived with (parents, step-parents, siblings), how often they ate together as a family, and whether the adults in the household worked. However, I do not have enough data for most of these outcomes to include them as controls in regressions without a substantial reduction in sample size, although I do use the variables on whether pupils live with their parents in Chapter 4. Moreover, these variables are pupil reported, and so may be noisy indicators of family characteristics.

Strengths to the study include the range of outcome measures, covering different domains and obtained from different sources; the large sample size; the very low attrition; and the realistic implementation in schools. The follow-up period is also relatively long compared to similar studies (e.g. Stallard et al., 2012). There is a need for more pragmatic studies in the literature on wellbeing and mental health promotion interventions for children and adolescents. Following the guidelines for classifying a trial detailed in Thorpe et al. (2009), the UKRP intervention does appear to be highly pragmatic. All pupils in the relevant cohorts at participating schools are included, regardless of their characteristics, and 91% of pupils of this age attend this type of school (state maintained secondary schools; DCSF, 2008b; 2009b). The control group received the 'usual practice' lessons, taught in the usual way. Programme compliance was only loosely and unobtrusively monitored through class registers which teachers usually kept routinely for their own records, and there were no attempts to intervene if schools or facilitators deviated from the suggested dosage or tried to modify the programme. I use intention-to-treat analysis, taking pupil's intervention assignment as their final assignment regardless of switches. In other ways the trial is less pragmatic: teachers were trained to use the intervention, although the training period lasted days not years, and they were educational practitioners rather than mental health professionals. Moreover, a critical part of the training was the emphasis on teachers adapting the teaching style to suit their pupils – there was no suggestion that there was only one way of teaching the curriculum. I did use a 2-year follow-up measure which all pupils were assessed on, but there was no further intervention after the first year. Moreover, much of the data comes from databases and it was not necessary to have contact with participants to obtain this. However, the primary outcome (the depression symptoms score on the Children's Depression Inventory) of the trial was chosen for research reasons rather than being of particular usefulness or meaning to study participants. Secondary outcomes such as absence, academic attainment, behaviour and friendships are arguably of more importance to participants.

Cost of the intervention

Table 1.16 presents estimates of the cost of the intervention per pupil. The 2007 training course was substantially more expensive than the training offered from 2009 onwards, primarily because in 2007 and 2008 trainees came from geographically dispersed regions so accommodation had to be provided.[51] From 2009 training sessions were provided locally and trainees commuted each day. Note that since the training course is a large fixed cost, the number of students a trained facilitator goes on to teach will be a major factor in the cost effectiveness of the intervention. Long run costs per pupil therefore converge to the variable cost rate, and are very close to this figure after a facilitator has taught 10 workshop groups (Figure 1.5). The variable cost is determined by the cost of printing and other classroom resources; and the cost of halving class sizes for the duration of the programme – this in turn depends on staff salaries. Because of this, poor selection of facilitators which results in training people who only teach a few groups vastly increases the per pupil cost of the intervention. I will use these cost figures in estimating the cost effectiveness of the intervention in Chapters 2 and 3.

The UK Resilience Programme: Current practice

The feasibility and sustainability of the intervention is partly what makes this a pragmatic trial. As mentioned above, the intervention was offered to subsequent cohorts of teachers and pupils, even though this research followed only the first cohort. Since this first cohort there have been 10 training courses for facilitators held in the UK, more than 800 teachers have been trained, and UKRP is now used in more than 85 schools. About 250 of these places were entirely funded by schools, with the remainder jointly funded by schools with LAs, public health organisations, or special grants. Several details of the implementation, particularly the training, have been modified since the first cohort. Flying staff to the USA for training was not sustainable, and subsequent courses have been held in the UK. Training has also been shortened to five days, preserving the number of hours by lengthening each day. The first year’s training took place during the summer holidays, but courses are now scheduled during term time. Since funding for the programme now comes from schools directly, they tend to think more carefully before paying for it how many workshops each teacher can teach. School organisation and the selection of facilitators have probably improved as a result (the training period is no longer a holiday). See Challen and Bailey (2012) for more information about current practice. A further field experiment will being in September

[51] Note that the greater cost in 2007 was not due to the training occurring in the USA: costs were similar in 2008. In 2007, most trainees would have had to be accommodated in hotels even if the training had been held in the UK, and exchange rates and the cost of UK accommodation made going to the US was the cheaper option. From 2009, accommodation would only be needed for delegates from outside of the region.

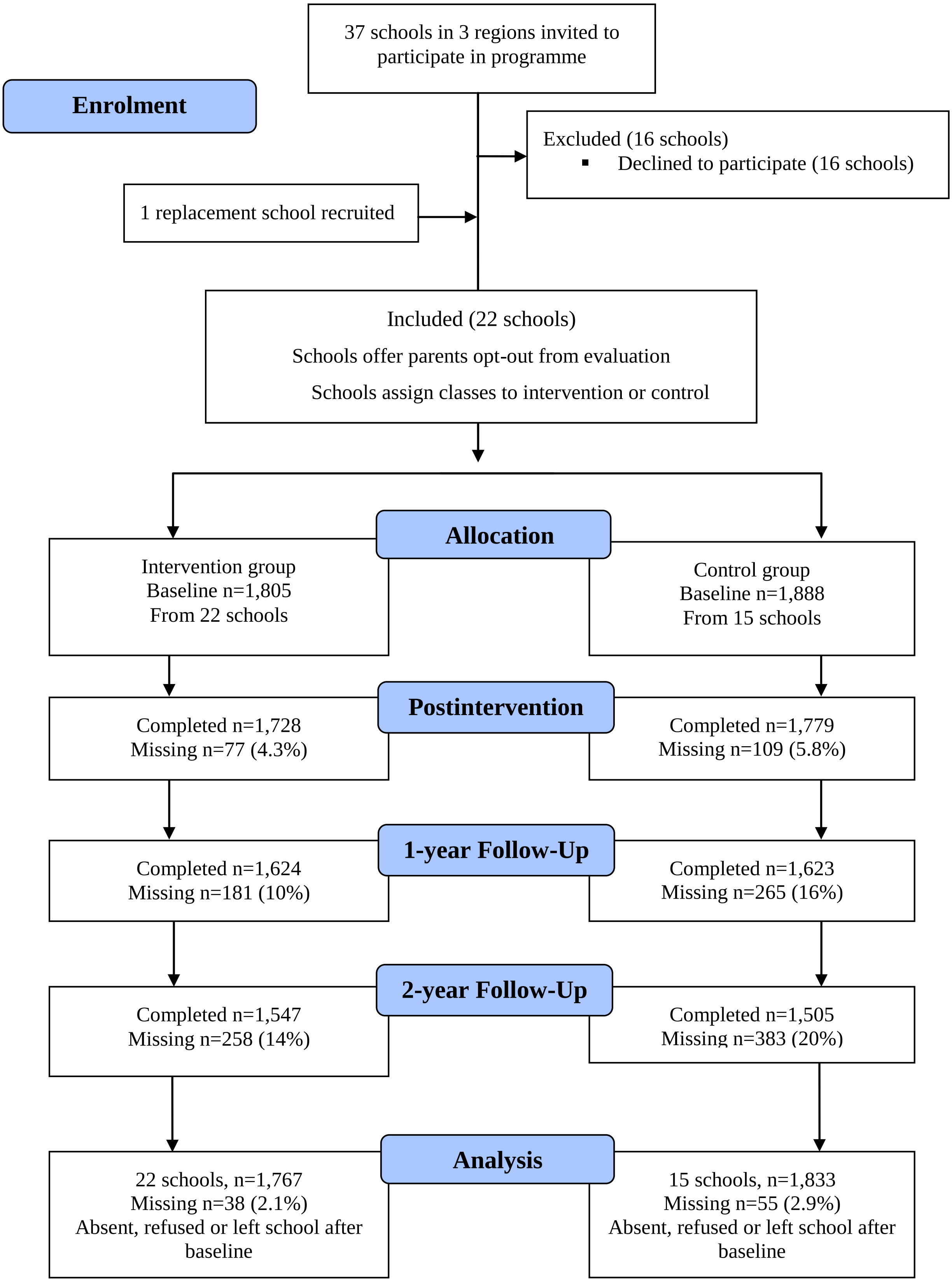
2013, trialling a full PSHE curriculum over four years, the first component of which is UKRP (EEF, 2013).

Conclusion

In this chapter I have described the background and set up of the programme and evaluation, and described the motivation for piloting UKRP and running a pragmatic trial. I also provided descriptive statistics of the LAs and schools involved relative to all state schools in England, and explained the recruitment strategy. The sample of schools I obtained data from is more deprived than the average for England, but pupils appear to have a similar standard of academic attainment at entry. I gave a description of the intervention and its aims, and provided an overview of the data I will use in the next four chapters. I also presented some indicators of implementation fidelity: intervention dosage was generally high, with most pupils receiving a large part of the programme, in groups of roughly the recommended size. I also gave estimates of the cost of the intervention, in the first year and in the longer run. In Chapter 2 I report the evaluation impact on symptoms of depression, anxiety, behaviour scores, absence and attainment. In Chapter 3 I present the evaluation impact on pupil’s popularity.

Chapter 1: Figures and tables

Figure 1.1: Flowchart of the recruitment and retention of participants in the evaluation



Notes: The data counts in this table refer to observations with valid measures on the depression symptoms (CDI) score.

Figure 1.2: The ABC model
Figure taken from Challen and Bailey (2012), courtesy of How To Thrive
http://www.howtothrive.org/

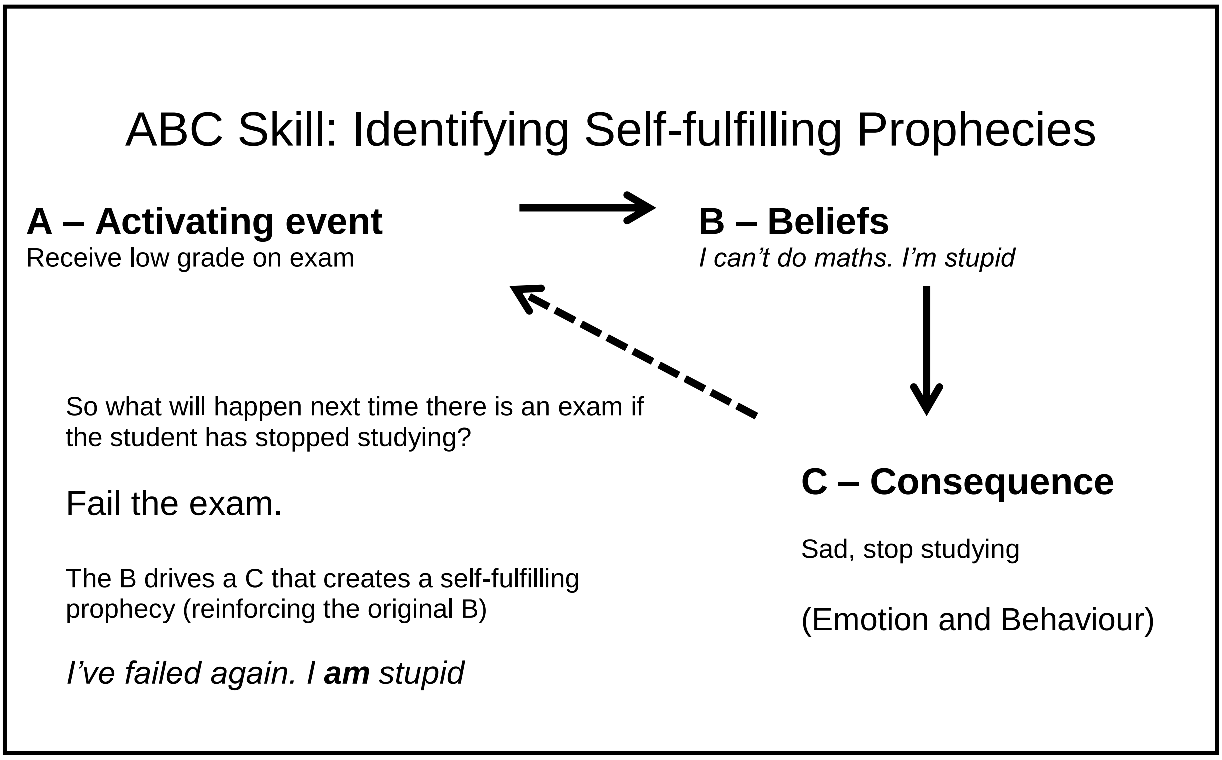


Figure 1.3: Distributions of the main psychological outcome variables at baseline
All pupils in 2007 cohort surveyed in September 2007. Baseline scores are presented for simplicity and clarity; distributions of each variable are very similar at other time periods.

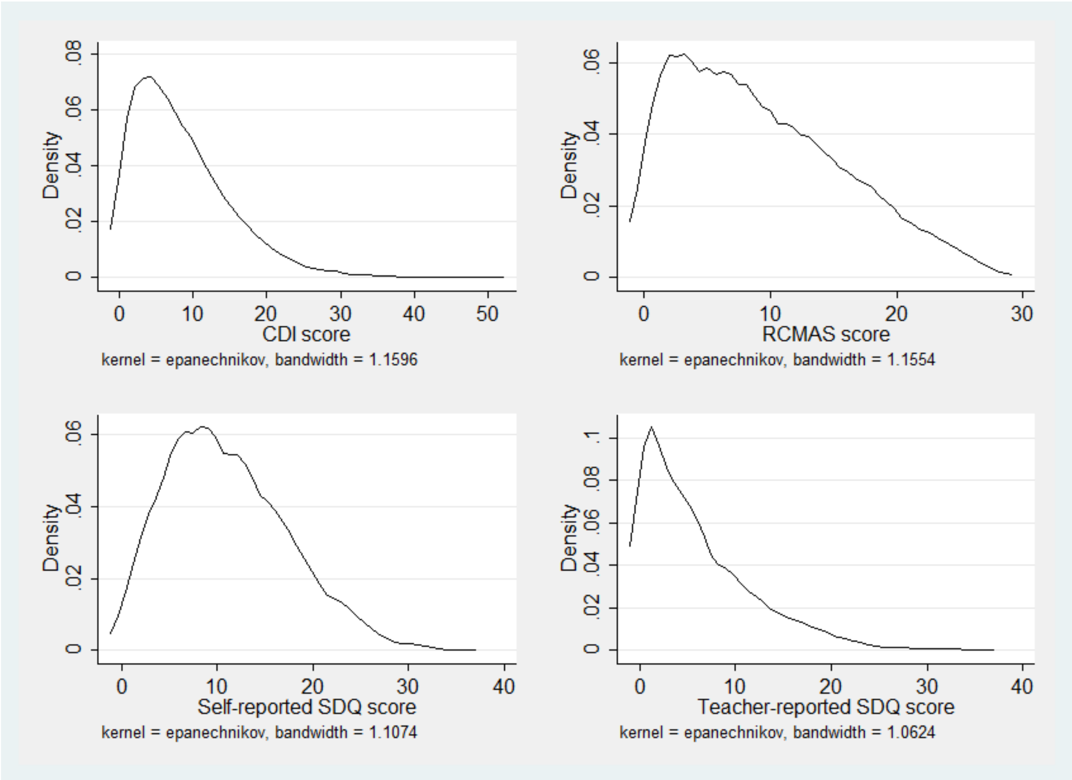


Figure 1.4: Distributions of the other outcome variables at baseline
Data from 2007 cohort. Absence rate is for academic year 2006-07; academic attainment for Key Stage 2 tests May 2007; friendship nominations received for end of Year 7 in June 2008. Distributions of each variable are very similar at other time periods.

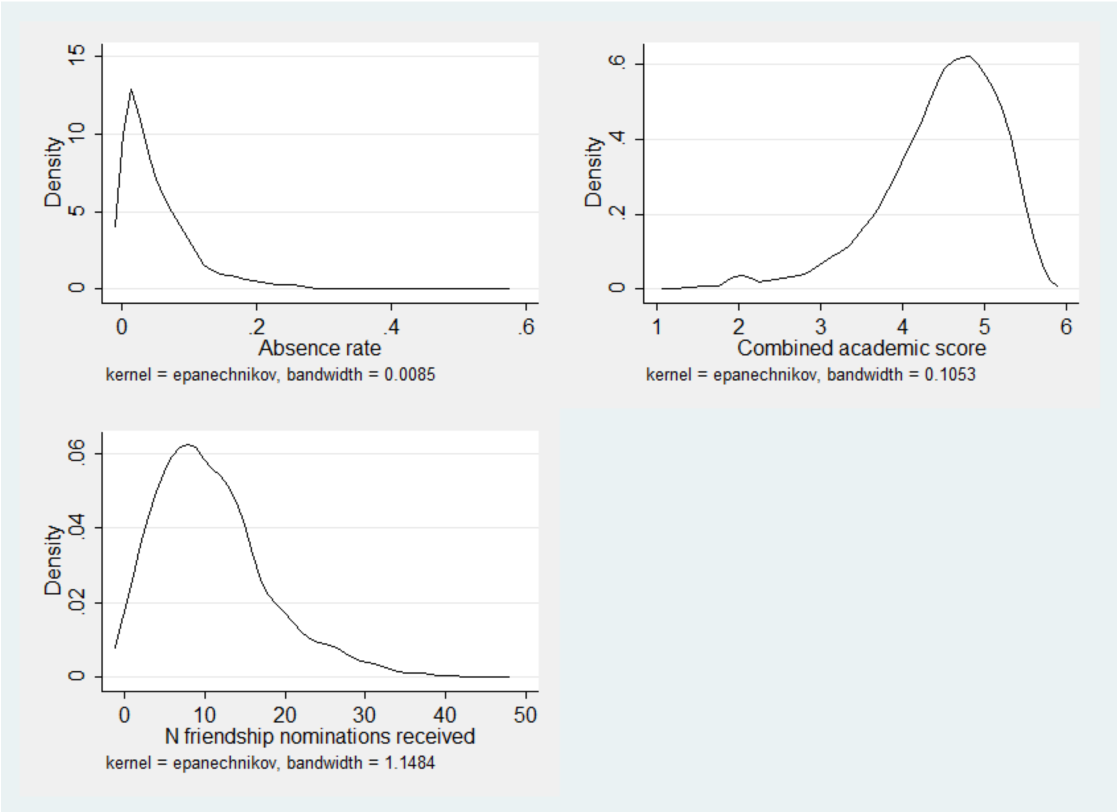
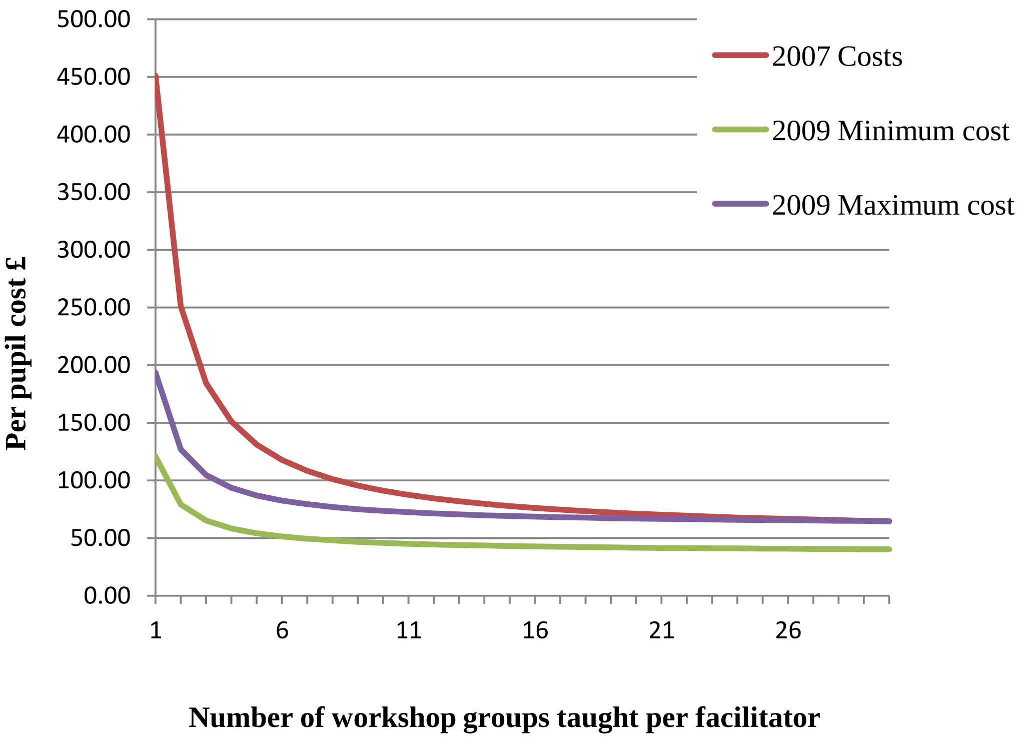


Figure 1.5: Per pupil intervention costs by number of workshop groups taught



Notes: The per-pupil cost of the UKRP intervention depends on the number of workshop groups taught by their facilitator, because of the large fixed cost of training. The figure therefore shows the per-pupil cost under a range of scenarios. I have not attempted to discount the upfront investment in teacher training and workshop delivery by the cost of capital: doing this would mean that facilitators who taught 10 workshop groups in the first year after training would be more cost-effective than teachers who taught the same number of groups over a longer time period (all else being equal) because of the time dimension of money.

Table 1.1: Participating Local Authorities

	South Tyneside	Hertfordshire	Manchester
Description	Metropolitan borough in North East of England	Large non-metropolitan county outside Greater London	Densely populated urban borough in North West of England
Deprivation rank (of 149)	39	139	4
% White ethnic origin	97%	94%	81%
% Asian ethnic origin	2%	3%	9%
% Other and mixed ethnic origin	1%	3%	10%

Data sources: Deprivation rank in 2010 (DCLG, 2011); ethnicity from 2001 census (ONS, 2003).

Table 1.2: Number of UKRP workshop groups taught by facilitator

	Number	%
All facilitators	***73***	**100%**
Female	*55*	75%
School staff	*62*	85%
Subject teachers	*43*	59%
Taught 1 UKRP group	*22*	30%
Taught 2 UKRP groups	*29*	40%
Taught 3 UKRP groups	*11*	15%
Taught 4 UKRP groups	*5*	7%
Taught 5 UKRP groups	*4*	5%
Taught 6 UKRP groups	*2*	3%
All UKRP groups	***146***	**100%**
Taught by:		
1 facilitator	*131*	90%
2 facilitators	*11*	8%
3 facilitators	*4*	3%

Notes: These tables describe the workshop groups, students and facilitators involved in this evaluation and which took place during 2007-08. Other facilitators were trained to teach UKRP who ran workshops outside of mainstream schools which are therefore not included here.

Table 1.3: Control groups available

Number of schools	No 2006 cohort control group	With a 2006 cohort control group	**Total**
No 2007 cohort control group	0	7	**7**
With a 2007 control group	9	6	**15**
Total	**9**	**13**	**22**

Table 1.4: PRP Curriculum Contents
Figure adapted from Challen and Bailey (2012)

Cognitive components	**Skills and coping strategies**
Emotion awareness ABC: the link between thoughts and feelings/actions Identifying thinking styles Examining alternatives beliefs and looking for evidence Thinking flexibly and accurately Putting it into perspective (de-catastrophising) Challenging negative beliefs quickly, in the moment (Real Time Resilience)	***Problem-solving and Social Skills*** Assertive Communication Negotiation Creative problem-solving Decision making Overcoming Procrastination Social Skills
	Coping and Calming Skills Emotion regulation & control Relaxation Distraction (Changing the Channel)

Table 1.5: Data available by cohort and measure

Number of schools for which data available at each point

		2007 cohort	2006 (control) cohort
Questionnaire measures	Baseline	22	0
	End of Year 7	22	13
	End of Year 8	22	13
	End of Year 9	22	13
	At least one follow-up point	22	13
Friendship measures	Baseline	0	0
	End of Year 7	22	0
	End of Year 8	22	13
	End of Year 9	22	13
	At least one follow-up point	22	13
Academic attainment data	Baseline	22	16
	End of Year 7	19	16
	End of Year 8	19	15
	End of Year 9	19	15
	At least one follow-up point	20	17
Absence data	Baseline (Year 6)	15	0
	Year 7	22	13
	Year 8	22	13
	Year 9	22	13
	At least one follow-up year	22	13

Notes: The questionnaire measures include the depression and anxiety symptom scores, and the self- and teacher-reported Goodman SDQ scores. The friendship measures were also collected via the same questionnaire, but the question was not introduced until June 2008. Academic attainment data is obtained from the National Pupil Database (baseline – Key Stage 2 test data) and from schools’ own databases (other time points). Absence data is from the National Pupil Database. The time points indicated will vary by cohort, e.g. the 2006 cohort will reach the end of Year 7 in July 2007, while the 2007 cohort will reach it in July 2008.

Table 1.6: Behaviour incident data by cohort
Number of schools for which data available by year group (grade) and cohort

Cohort (year starts school)		2009	2008	2007	2006	2005	2004	2003	2002	2001
Behaviour incident data	Year 7	2	2	2	2	1				
	Year 8	1	2	3	2	1	1			
	Year 9		1	4	3	1	1	1		
	Year 10			2	3	2	1	1	1	
	Year 11				2	2	2	1	1	1

Table 1.7: Number of UKRP students and workshops by school and LA (2007 cohort only)

LA	School	Number of workshops by school	Number of workshops by LA	Approx. size of cohort by school (# students)	Approx. size of cohort by LA (# students)	Number of intervention students by school	Percentage of cohort in intervention by school	Number of intervention students by LA	Percentage of cohort in intervention by LA	Percentage of sample in intervention
Herts	1	6	57	60	920	60	100	770	84	
	2	10		120		120	100			
	3	14		200		200	100			
	4	4		180		60	33			
	5	10		180		150	83			
	6	13		180		180	100			
Manchester	7	4	38	240	1655	60	25	515	31	
	8	3		60		33	55			
	9	4		190		57	30			
	10	6		300		75	25			
	11	6		170		85	50			
	12	4		195		59	30			
	13	4		180		60	33			
	14	5		140		56	40			
	15	2		180		31	17			
South Tyneside	16	4	51	215	1370	60	28	677	49	
	17	2		280		31	11			
	18	2		205		31	15			
	19	9		115		115	100			
	20	7		205		90	44			
	21	17		200		200	100			
	22	10		150		150	100			
TOTAL		**146**	**146**	**3945**	**3945**	**1962**		**1962**		**50**

Notes: Cohort size and number of intervention students are approximations due to student mobility and inaccuracies in school databases.

Table 1.8: Lessons replaced by UKRP workshops (control group alternative treatment)

	Number of schools
PSHE, citizenship, Learning 2 Learn, thinking skills, or pastoral lessons	16
Religious Studies	2
English, science, maths or modern languages	3
UKRP designated slot	1
All Schools	**22**

Learning 2 Learn lessons aim to teach pupils how to approach learning, covering self-awareness, organising and planning, reflection, self-assessment etc. See Formby et al. (2011) for details of different PSHE delivery models and curricula.

Table 1.9: Sample characteristics and population characteristics

			Hertfordshire		Manchester		South Tyneside	
	England average	My sample	All pupils in LA	My sample	All pupils in LA	My sample	All pupils in LA	My sample
Male	51%	53%	50%	51%	49%	54%	51%	51%
Special Educational Needs	20%	27%	16%	23%	24%	31%	19%	25%
Free school meals eligible	13%	28%	6%	10%	37%	41%	16%	23%
IDACI score: % children in neighbourhood who are deprived	21%	37%	13%	17%	44%	51%	28%	32%
White ethnic background	84%	73%	88%	89%	59%	45%	96%	96%
Attained level 4 in Key Stage 2 English (2007)	80%	77%	85%	82%	74%	71%	81%	82%
Attained level 4 in Key Stage 2 maths (2007)	77%	76%	82%	80%	74%	71%	78%	79%
Attained level 4 in Key Stage 2 science (2007)	88%	87%	91%	91%	84%	83%	88%	90%

Notes: Demographic characteristics for LAs and for England refer to pupils aged 11-15 in state-funded secondary schools in January 2008 (source: DCSF, 2008b). IDACI scores for 2007 (source: DCLG, 2008). IDACI scores given for Hertfordshire LA only apply to Dacorum district. Key Stage 2 data gives the percentage of pupils in the region obtaining the level specified who sat the exams in 2007, i.e. who were aged 11 and at the end of primary school in May 2007. These are the same cohort as the 2007 cohort in this evaluation. The statistics for 'my sample' refer to the full 2007 cohort.

Table 1.10: Attrition

		Depression/anxiety		Teacher-reported behaviour		Attendance at school		Academic attainment	
Data source		Pupil questionnaire		Teacher questionnaire		National databases		School and national databases	
N pupils with a baseline		3,693		3,570		2,448		6,592	
N pupils with a baseline and at least one follow-up period		3,600		3,522		2,427		6,575	
% pupils with a baseline and at least one follow-up		97.5%		98.7%		99.1%		99.7%	
Number of schools		22		22		15		20	
No follow-up data for any period	Coefficient on 'treated'	-0.008	-0.021*	-0.002	-0.006	-0.002	-0.002	0.054	-0.008
	SE	(0.007)	(0.011)	(0.004)	(0.007)	(0.001)	(0.001)	(0.036)	(0.008)
	p	0.254	0.052	0.672	0.397	0.149	0.151	0.138	0.303
	N missing information	93		48		2		560	
	% missing information	2.5%		1.3%		0.1%		8.5%	
No post follow-up	Coefficient on 'treated'	-0.015	-0.025*	-0.033	-0.040	-0.002*	-0.002*	0.015	-0.086***
	SE	(0.011)	(0.013)	(0.020)	(0.025)	(0.001)	(0.001)	(0.041)	(0.031)
	p	0.161	0.067	0.102	0.121	0.079	0.081	0.718	0.005
	N missing information	186		147		3		973	
	% missing information	5.0%		4.1%		0.1%		14.8%	
No 1-year follow-up	Coefficient on 'treated'	-0.040**	-0.033*	-0.052	-0.058	-0.007	-0.000	0.030	-0.057**
	SE	(0.016)	(0.019)	(0.034)	(0.038)	(0.004)	(0.005)	(0.040)	(0.023)
	p	0.015	0.084	0.127	0.130	0.119	0.936	0.459	0.014
	N missing information	446		460		20		877	
	% missing information	12.1%		12.9%		0.8%		13.3%	
No 2-year follow-up	Coefficient on 'treated'	-0.060***	-0.027	-0.138***	-0.077	-0.008	0.004	0.029	-0.010
	SE	(0.019)	(0.023)	(0.045)	(0.055)	(0.006)	(0.008)	(0.043)	(0.010)
	p	0.002	0.233	0.002	0.161	0.188	0.640	0.492	0.276
	N missing information	641		750		50		741	
	% missing information	17.4%		21.0%		2.0%		11.2%	
	Including School FE	No	Yes	No	Yes	No	Yes	No	Yes

Notes: Table presents information on attrition for each outcome separately. The value of the coefficient on ‘treated’ is from a regression of a missing dummy (set to 1 when an observation is missing the relevant outcome measure in a particular time period) on ‘treated’ (a dummy indicating treatment assignment), clustered at the level of treatment assignment (class). The second column under each heading includes school fixed effects. The base from which attrition is calculated is all students for whom I have baseline information on that outcome.

Table 1.11: Descriptive statistics for UKRP implementation by workshop group

	Number of students in UKRP group	Scheduled time available for intervention (hours)	Fraction of full programme scheduled (18 hours)
Mode	15	18	1
Range	[5,26]	[8,25]	[0.44,1.39]
10th percentile	10	15	0.83
25th percentile	11	17	0.94
Median	14	18	1
75th percentile	15	18	1
90th percentile	16	19	1.06
N workshops	146	146	146

Table 1.12: Descriptive statistics for UKRP implementation by intervention students

	Number of students in UKRP group	Scheduled time available for intervention (hours)	Fraction of 18-hour programme scheduled	UKRP hours actually received	Fraction of full 18 hours actually received
Mean	13.84	17.20	0.96	15.75	0.88
Mode	15	18	1	17	0.94
Range	[5,26]	[8,25]	[0.44,1.39]	[0.83,25]	[0.05,1.39]
10th percentile	11	14	0.79	12	0.67
25th percentile	12	16	0.89	14.29	0.82
Median	14	18	1	16.67	0.93
75th percentile	15	18	1	17.95	1.00
90th percentile	16	19	1.06	18	1.00
N pupils	1951	1951	1951	1680	1680

Table 1.13: Summary of UKRP workshop size, scheduling, and attendance

% of students in classes of 15 or fewer	83%
% of students in classes of 16 or fewer	94%
% of students in classes scheduled for 18 hours or more	55%
% of students in classes scheduled for 17 hours or more	74%
% of students in classes of 15 or fewer, scheduled for 18 hours or more	46%
% of students in classes of 16 or fewer, scheduled for 17 hours or more	73%
% of students in classes of 15 or fewer, actually attending 18 hours or more	19%
% of students in classes of 16 or fewer, actually attending 17 hours or more	45%

Table 1.14: Use of skills

Do you use any of the skills that you learnt in the [UKRP] classes?

	N	% of respondents	% of pupils with outcome data
Yes	676	48.9	38.3
No	706	51.1	40.0
Total with data	1,382	100.0	
No data	385		21.8
TOTAL	1,767		100.0

Table gives the numbers and percentages of pupils who are in the evaluation sample (outcome: CDI score) for whom I have data on their self-reported use of the UKRP skills. See Challen et al. (2009) p.23 for more details on the survey.

Table 1.15: Linear correlation coefficients between pairs of outcome variables at baseline

		Depression score	Anxiety score	Pupil-reported behaviour score	Teacher-reported behaviour score	Absence rate	Combined academic score
Anxiety score	coefficient	0.73					
	p-value	0.000					
	N	3,645					
Pupil-reported behaviour score	coefficient	0.73	0.77				
	p-value	0.000	0.000				
	N	3,645	3,607				
Teacher-reported behaviour score	coefficient	0.28	0.19	0.31			
	p-value	0.000	0.000	0.000			
	N	3,082	3,037	3,038			
Absence rate	coefficient	0.09	0.03	0.08	0.16		
	p-value	0.000	0.140	0.000	0.000		
	N	2,275	2,241	2,240	1,759		
Combined academic score	coefficient	-0.24	-0.19	-0.24	-0.35	-0.19	
	p-value	0.000	0.000	0.000	0.000	0.000	
	N	3,029	2,993	2,998	2,520	2,221	
N friend nominations	coefficient	-0.16	-0.16	-0.16	-0.25	-0.17	0.25
	p-value	0.000	0.000	0.000	0.000	0.000	0.000
	N	3,645	3,592	3,591	3,047	2,382	5,234

Notes: Coefficients are linear correlation coefficients between pairs of the seven outcome variables. Here I use the combined academic score, but correlations are similar between each component factor (English, maths and science) and the other outcomes.

Table 1.16: Estimated costs of the intervention

		Cost in 2007	Min cost 2009 and later	Max cost 2009 and later
Fixed costs (per facilitator)	Training course	6000.00	900.00	1200.00
	Lesson cover for facilitator training	0.00	350.00	800.00
Variable costs (per pupil)	Printing and resources	15.00	15.00	15.00
	Implicit cost of doubling staff for UKRP lessons	36.19	22.53	45.30
Facilitator teaches 45 students	TOTAL per pupil: training and printing only	148.33	35.00	41.67
	TOTAL per pupil: all costs, including staff time	184.53	65.31	104.75
Facilitator teaches 225 students	TOTAL per pupil: training and printing only	41.67	19.00	20.33
	TOTAL per pupil: all costs, including staff time	77.86	43.09	69.19

Notes: Table presents estimated per pupil costs of the intervention under different scenarios. The 2007 training programme was not representative of later courses, while costs for 2009 and later reflect the longer-run cost of providing the intervention to students. This varies for two reasons: first, there are high fixed costs to training provision, so courses with more facilitators will cost less per facilitator trained. I have provided the minimum and maximum current costs. Second, the opportunity cost of staff time varies by role. The maximum cost reported here is for covering a teacher to attend five days of training; the minimum for covering a learning support assistant. Similar figures are provided for the implicit cost of doubling staff for UKRP lesson. This is equivalent to employing an extra member of staff for 1 of 25 weekly lessons for half a year (18 weeks of the programme): (1/25)*(1/2)*staff annual salary or 2% extra annual staff costs per workshop group. The cost in 2007 for doubling staff is based on the 60/40 split of teachers/other staff observed in the data. No staff cover was needed for training in 2007 because it took place during the summer holidays. With 15 students per group, teaching 45 students would be equivalent to teaching 3 workshop groups, and 225 students would be 15 workshop groups. Note that I have not attempted to discount the upfront investment in teacher training and workshop delivery by the cost of capital: doing this would mean that facilitators who taught 10 workshop groups in the first year after training would be more cost-effective than teachers who taught the same number of groups over a longer time period (all else being equal) because of the time dimension of money.

Chapter 2: Impact of the UK Resilience Programme on mental health symptoms, absence and academic attainment

Introduction

In this chapter I present the main results for the UKRP evaluation. The background to the trial and details of the intervention are described in Chapter 1, as are the recruitment of LAs, schools and teachers. Here I present the intervention impacts on depression and anxiety symptom scores, pupil- and teacher-reported behaviour scores, absence from school and academic attainment. This allows me to assess the outcome of the programme on both subjective and objective measures of student outcomes. I collected data at baseline, at post intervention, one-year follow-up and two-year follow-up, so am able to follow most students up to 33 months after the baseline measure. I estimate the programme impact on a number of outcomes which are reliably measured, over a significant time period, using a panel dataset and student fixed effects. This study is the largest PRP evaluation to date, and one of the largest evaluations of similar programmes. The study was designed as a pragmatic trial, taught by teachers and other school staff within timetabled lessons in mainstream schools. The study therefore contributes substantially to the evidence on the impact of this and similar programmes under real world conditions, robustly measuring a number of outcomes to assess programme impact.

The depression symptoms score was the primary outcome, and is the outcome on which most previous research has focused. I found a small average impact on symptoms of depression and on absence from school, neither of which persists beyond postintervention. I also found a small impact on academic attainment which was still present at the one-year and two-year follow-up points. There was some variation in intervention impact by workshop quality and by pupil characteristics, and this varied by the outcome measure used. I found no impact on anxiety symptoms or on self- or teacher-reported behaviour. I present a cost-effectiveness analysis. I conclude that although the intervention had an impact on important outcomes, given that these effects were small and the cost of the intervention was relatively high, it is not clear that offering it universally would offer value for money. However, this would partly depend on how schools organised the programme.

This paper addresses the following questions:

1. What impact did the UK Resilience Programme have on pupils' symptoms of depression and anxiety, behaviour scores, absence from school and academic attainment?
2. Was there heterogeneity in intervention impact by workshop organisation or pupil characteristics?

Since the programme set up, general approach, population, power calculations and data are described in detail in Chapter 1, I start by briefly outlining the evaluation design, then describing the statistical method. I then present baseline information on the outcome variables and other pupil characteristics, followed by the evaluation results. Then follows the discussion section, an analysis of cost effectiveness, conclusions and policy implications.

Evaluation Design and Data

The trial was designed as a pragmatic non-randomised controlled trial, with classes of pupils as the unit of intervention assignment. Evaluation participants were students starting in Year 7 at a UKRP school in September 2007. All students within this cohort (grade) in participating schools were surveyed at baseline and at follow-up points, unless they were absent, had left the school, or if they refused to complete the questionnaires or their parents had withdrawn them. There were no exclusion criteria apart from belonging to the correct cohort in each school.

I outlined the reasons for choosing each outcome variable in Chapter 1 ('Measures and data' and 'Appropriateness of measures'). The outcome variables came from different sources, and as a result I have a slightly different sample for each outcome. In order to maximise power and internal validity, I have included as many observations as possible in the assessment of programme impact on each outcome, which results in samples which do not fully overlap. However, I also present robustness checks using only pupils for whom I have data on all the main outcomes.

Statistical analysis

I adopt a difference-in-differences approach to data analysis, assessing the change in the outcome variable in the intervention group relative to the change in the control group. Analyses of the intervention impact were conducted using an intention-to-treat approach, and included all students with baseline data and data for at least one follow-up period (post intervention, one-year follow-up, or two-year follow-up).[52] Ordinary Least Squares regression in Stata 12 (StataCorp, 2011) was used to estimate the impact of the intervention at each follow-up period, by including the outcome at follow-up and at baseline on the left hand side, and regressing these on treatment assignment, and treatment assignment*post. I present several specifications, building up the regressions by sequentially including more control variables, in order to test the robustness of my results. This is particularly important given the non-random condition assignment. I include controls for: the day of the week and month the questionnaires were completed (if the outcome data came from questionnaires); student characteristics (gender; FSM; SEN; prior academic attainment; and ethnicity); and the school students attended at baseline. All control variables were entered as dummies. These characteristics were included because of their availability and association with the outcome variables (Chapter 1, 'Pupil characteristics'). I also split the sample by student and workshop characteristics to examine impact heterogeneity. Intervention impacts at different time periods are assessed in the same regression, using interactions of treated*time.

For each outcome measure I therefore present the following analyses.

(1) $$y_{it} = \beta_1 TREAT_i * \gamma_t + TREAT_i + \gamma_t + \varepsilon_{it}$$

(2) $$y_{it} = \beta_1 TREAT_i * \gamma_t + TREAT_i + \beta_2 X_i + \gamma_t + \pi_{it}$$

(3) $$y_{it} = \beta_1 TREAT_i * \gamma_t + TREAT_i + \beta_2 X_i + \gamma_t + s_i + \gamma_t * s_i + \nu_{it}$$

(4) $$y_{it} = \alpha_i + \beta_1 TREAT_i * \gamma_t + \gamma_t * s_i + \eta_{it}$$

Where for each student i in school s at time t, y_{it} is the outcome of interest (at baseline or follow-up); $TREAT_i$ is a treatment assignment dummy; γ_t is a set of four time dummies (postintervention; one-year follow-up; two-year follow-up; these are relative to baseline which is the excluded category); X_i is a vector of student characteristics; s_i is a school fixed effect; $\gamma_t * s_i$ is an interaction of the school a pupil

[52] In practice this means restricting the sample to pupils in the 2007 cohort for all outcomes other than academic attainment, for which I have data for many students in the 2006 cohort too.

attended at baseline with time period (a school-specific time dummy); α_i is a student fixed effect; and $\varepsilon_{it}, \pi_{it}$, ν_{it} and η_{it} are unobserved individual random effects. In each case β_1, the coefficient on treated*post, estimates the impact of the intervention. When controlling for the school a student attends I use the school they attended at baseline, as subsequent school moves could be endogenously determined. When the outcome is obtained from a questionnaire, the day of the week and the month the questionnaire were completed are also included as dummies in all four equations.

Equation (4) represents my preferred specification, and I use this to go on to explore heterogeneity in intervention impact by pupil and workshop characteristics. Taking impact heterogeneity by pupil gender as an example, these regressions will be of the following form:

$$\text{(5)} \qquad \begin{aligned} y_{it} &= \alpha_i + \beta_2 TREAT_i * \gamma_t * boy + \beta_3 TREAT_i * \gamma_t * girl \\ &+ \gamma_t * s_i + girl_i * \gamma_t + \omega_{it} \end{aligned}$$

Where β_2 will give the average intervention impact on boys and β_3 the average impact on girls. It is not possible to include a dummy for gender in this specification, because the pupil fixed effects causes any variable which does not change through time to drop out. However, it is important to include the interaction of gender with time to account for possible differences in the trends of these two groups through time. This highlights an important identifying assumption for difference-in-differences: I have to assume that the trends (changes) in the intervention and control groups would have been the same in the absence of the intervention (Angrist & Pischke, 2009). When condition assignment is truly random this is a plausible assumption to make; in this trial it was arbitrary but not random so there may be more doubt. All regressions have standard errors clustered at the level of class grouping, which is the level of intervention assignment.[53]

[53] We may still encounter another problem due to serial correlation through time as well as within clusters: with a panel, it is likely that observations will be correlated with previous observations on the same unit. Clustering the standard errors at class level rather than class*time should take account of this, as it allows unrestricted residual correlation within clusters, including across time (Angrist & Pischke, 2009).

I also present a simple test of mediation, following the procedure in Baron and Kenny (1986). This is to explore whether any programme impact on one variable is related to a programme impact on another, or whether the effects operate independently. For example, an impact on academic attainment could be associated with an impact on depressive symptoms. Using the same specifications (1)-(4) as detailed above, the steps involved are as follows:

Step 1: Regressing the dependent variable (e.g. academic attainment) on the independent variable (here, intervention assignment). This is to check whether intervention assign is a significant predictor of the dependent variable.

Step 2: Regressing the mediating variable on intervention assignment (the independent variable). If there is no significant association, then the mediator cannot be mediating the impact on the dependent variable.

Step 3: Regressing the dependent variable on both the independent variable and the mediator (in terms of the specifications (1)-(4) above, this would involve adding the mediator to each equation). If there is a significant relationship between the mediator and the dependent variable, and if the strength of the relationship between the independent variable and the dependent variable is greatly reduced (relative to Step 1), this may indicate that the mediator mediates the impact on the dependent variable. This cannot establish the direction of causation, but it can provide evidence as to whether a change in one variable is associated with a change in another.

Results

Chapter 1 presents information on attrition, and on programme attendance and completion. Here I present information on how well matched the intervention and control groups were at baseline, which provides evidence for the internal validity of the experiment. I then show regression estimates of the intervention impact on UKRP pupils.

Intervention and control groups

Table 2.1-Table 2.3 present mean scores at baseline for the treatment and control groups, including the p-values from mean-comparison tests with and without school fixed effects. The six main outcome variables have different sample sizes because of differential availability of data. I therefore present three separate tables to make clear the match between the treatment and control groups in each case. A good match at

baseline does not necessarily mean that the groups are the counterfactuals of one another: as discussed above, if the pre-existing trends in the outcome variables differed between the two groups this would undermine my identification strategy, even if baseline values were the same. However, a bad match at baseline might be seen as evidence that the two groups are not counterfactuals of each other.

The first three variables in the upper panel of Table 2.1 are the baseline values of the first four outcome variables, which were obtained from questionnaires: the depression, anxiety and two behaviour scores. The sample available for the depression, anxiety and behaviour outcomes is very similar since they come from the same source – the questionnaire booklet students filled in (although the teacher-reported behaviour scores come from the teacher questionnaire). The lower panel presents the values of demographic variables at baseline for this sample. The table includes all pupils who had a baseline measure on these variables, including those without follow-up data. All pupils appearing here are in the 2007 cohort.

The intervention group is not significantly different from the control group at baseline in terms of depression, anxiety, and behaviour symptom scores. However, the intervention group tends to score slightly higher (worse) on average than the control group. The groups are similar in terms of gender composition and the age of students. The proportion of students with special educational needs (SEN) looks the same, until I add school fixed effects. The reverse is true for entitlement to free school meals (FSM): students in the intervention group appear to be much less likely to be eligible for FSM, but adding school fixed effects makes this difference insignificant. Students in the intervention group had significantly lower prior attainment in English and maths.

There are two possible causes of differences between the intervention and control groups: selection into the intervention within schools; and schools with different student characteristics including different proportions of the cohort in workshops. The former would arise if, within schools, teachers targeted classes of pupils with particular characteristics for the intervention. Here we might not see a difference in the average levels of the variable, but we would expect to find a significant p-value once controlling for school fixed effect – this would be a within-school difference. The latter case is due to differences between schools, combined with differences in the proportion of pupils schools were able to include in workshops. For example, in this sample the proportion of pupils eligible for FSM by school ranges from 5% to 61%, while the proportion of the 2007 cohort included in workshops ranged from 11% to 100%. Since less deprived schools included a larger proportion of their cohort in the programme, the intervention group will be less deprived than the control group. With the FSM variable, we can see that the difference between intervention and control groups becomes insignificant once school fixed effects are added, suggesting that the between-school difference is responsible and schools were not more likely to include their non-FSM pupils in workshops than their FSM pupils. The reverse is true of SEN pupils and pupils with low attainment: schools were less likely to include pupils with SEN or low attainment in workshops. As discussed in Chapter 1, this was because four schools included higher ability classes of pupils in the intervention (two schools did this as a deliberate strategy; in two schools these just happened to be the classes which fitted the timetable). Once I control for school fixed effects there appear to be slightly more girls than boys in workshops (within school). This may be related to differences in academic attainment.

Thus I find no significant differences in the outcome variables at baseline (depression, anxiety and behaviour), but large and stable differences in academic attainment and SEN. I will control for these characteristics in the analyses, but these differences suggest that workshop assignment was not entirely arbitrary. Although the noncognitive outcome variables here are not significantly different between intervention and control groups, these may well be affected by cognitive ability (Borghans et al., 2008). This could prevent us from interpreting any differences in postintervention outcomes between the intervention and control groups as attributable to the intervention, rather than to other, correlated factors that differed between the two groups.

Table 2.2 presents descriptive statistics for the sample used when the absence rate is the outcome. The sample size is smaller than for the questionnaire outcomes because baseline data were missing for one region (7 schools), but attrition conditional on having a baseline was very low because these measures were obtained from the NPD (see Chapter 1, ‘Measures and data – absence from school’). The absence variable gives the fractions of sessions for which a student was absent during the previous academic year (September 2006 - June 2007), when they attended a different (primary) school. The patterns here are similar to those seen in the sample available for questionnaire outcomes described above (and reported in Table 2.1): the outcome variable is very similar between the intervention and control groups, as is age and FSM (once I control for school fixed effects). Here the fraction of students with SEN does not differ between the two groups, but the level of prior attainment does. Thus in this sample, despite similarities on the main outcome variable, there are sizeable differences in the composition of the sample in terms of gender and prior attainment, with the intervention group being more female and higher attaining.

Table 2.3 presents baseline statistics for the sample for which I have outcome data on academic attainment. Only 20 schools provided follow-up academic data, but because data came from school databases I was able to use a much larger sample of students for the control group by including data from students who were in previous cohorts at the same schools. As discussed in Chapter 1, the crucial assumption here is that adjacent cohorts in the same school are formed quasi-exogenously, and so can be viewed as the counterfactuals of one another when one cohort receives the intervention and another does not. Using this data provides me with a substantial within-school control group for each of the 20 schools with data.

As before, intervention group students' prior academic attainment is slightly higher than that of the control group, although this is only marginally significant for English and maths when school fixed effects are included. Using the extra control pupils evens up the demographic characteristics of the sample – FSM and SEN are both now very similar in the intervention and control groups. However, there is a larger fraction of girls in the intervention group. There are two measures of age provided here: the age of students in September 2007, and their age when they started secondary school ('baseline'). The former is clearly significantly different between the intervention and control groups, because I am using older cohorts to supplement the control group but not the intervention group. However, intervention and control group students were the same age at the start of secondary school.

Overall, the arbitrary assignment of classes to intervention or control groups within schools did not work that well: the intervention group is more female, and has higher prior academic attainment than the control group. I control for these differences in all my regressions, but they may indicate other, unobserved, differences between the two groups.[54] Moreover, an important assumption of difference-in-differences is that the trend in both groups would be the same through time in the absence of the intervention. Pupils with higher academic attainment at baseline also experience a higher growth rate in attainment to each follow-up point, with the difference becoming larger through time, so even small differences in baseline characteristics could result in biased estimates of

[54] Another possible difference between the two groups is the primary schools pupils attended, which could be associated with differential trends. However, secondary school classes are designed to mix up pupils from different primary schools, so this is not likely to be a problem over and above baseline differences in academic attainment.

programme impact on this measure. I present a number of robustness checks attempting to correct for these differences, but they may not be able to compensate for the non-random assignment.

Programme impact

Table 2.4 presents difference-in-differences estimates of the programme impact on students' depression and anxiety symptom scores and absence from school. All outcome variables have been standardised to have a mean of 0 and a standard deviation of 1, in order to facilitate interpretation in terms of standard deviations and thereby give an indication of effect size. The coefficient on Treated*Postintervention gives the average programme impact soon after the workshops finished; the coefficient on Treated*1-year follow-up gives the programme impact in June 2009, one year after the end of the intervention; and the coefficient on Treated*2-year follow-up gives the impact in June 2010, roughly two years after the end of the intervention. The match between the intervention and control groups in terms of the outcome variable at baseline is given by the coefficient on the 'Treated' variable. The descriptive statistics in Table 2.1 suggest that the intervention group had nonsignificantly higher depression and anxiety scores at baseline than the control group; in Table 2.4 this difference is statistically significant. However, these regressions are restricted to pupils for whom I have at least one follow-up data point and the demographic and attainment control variables, and excluding pupils without this information apparently makes the two groups more different.[55] As in Table 2.2, there is no significant difference between pupils in terms of their absence rates at baseline.

Each column represents a separate regression. The first column for each outcome variable includes only the variables reported, plus a time dummy (postintervention, 1-year follow-up or 2-year follow-up), and for the depression and anxiety scores, the day of the week and the month in which the questionnaires were completed.[56] The second column adds in pupil characteristics: dummies for gender, FSM, SEN, prior attainment in English and maths (five categories for each), and broad ethnic group (5 categories).

[55] I have restricted all regressions for a given outcome to have the same number of observations to facilitate comparisons between specifications. Allowing the number of observations to vary e.g. by including observations does not change the results. Additionally, in Table A2.1 I present the means and standard deviations of all outcome variables, including all pupils who had a baseline measure for that variable and at least one follow-up observation as described in the flow diagram in Chapter 1. This table does not exclude pupils without control variables.

[56] The day of the week can have a substantial impact on pupils' reported wellbeing and behaviour, see Chapter 5.

The third column adds in school fixed effects and school*time dummies, to account for school-specific trends. The fourth column uses pupil fixed effects, which is my preferred specification. Note that in this specification the pupil characteristics (which are set at the baseline value) and the school fixed effects drop out as they are fixed through time, but not the school*time dummies.

Based on these regressions I find a small programme impact on pupils' depression symptom scores at postintervention, of approximately 0.10-0.15 of a standard deviation, and no impact on depression scores at 1-year or 2-year follow-up. Remember that a higher depression symptoms scores indicates more or worse symptoms, so a negative coefficient on 'Treated*Postintervention' indicates an improvement in symptoms of depression. I find no impact at any point on anxiety symptom scores. There is a short-run intervention impact on pupils' absence from school, of around 0.14 SD, but this does not persist into the two follow-up periods. Note that while the psychological measures are assessed at a point in time, the absence rate is measured over the full academic year. Thus the baseline covers 2006-07, when pupils were at primary school; the equivalent of postintervention covers 2007-08, during which pupils participated in workshops; and the next two academic years are the equivalent measures for one-year and two-year follow-up.

Table 2.5 presents similar regressions for the next three outcomes: self- and teacher-reported behaviour, and academic attainment. The intervention and control groups are well matched on all three of these outcomes at baseline. There is no programme impact on self- or teacher-reported behaviour. Additional analyses on the prosocial subscales of the SDQ either show no impact at all (teacher-reported score) or even suggest a slight negative impact at 2-year follow-up (pupil-reported), although this is not robust to different specifications (Table A2.2; this scale is positively scored so that a positive coefficient would indicate an improvement).

There may be some impact on combined academic attainment in all three time periods, although the standard errors are large in the first two specifications and so although the magnitude of the coefficients does not change much they are not always statistically significant. The sample size here is larger than for the other outcomes because I was able to supplement the 2007 cohort data with data from the 2006 cohort, who will have had largely the same teachers and the same assessments. This leaves the size of the intervention group unchanged, but greatly increases the size of the control group and improves the demographic and academic 'match' between the two groups (see Table 2.3). It also means that I have a within-school control group for all 20 schools. Using these data, I find a small programme impact on combined academic attainment of 0.10 SD at postintervention; 0.08 SD at 1-year follow-up; and 0.13 SD at two-year follow-up. A higher score for academic attainment indicates higher attainment, so a positive coefficient represents an improvement.

Heterogeneity in programme impact

As discussed in Chapter 1 ('Programme adherence'), not all schools scheduled 18 hours for the workshops, and not all succeeded in forming classes of 15 students or fewer – most treated pupils were not in workshops which met these criteria. Yet workshop teachers felt that these were important elements in ensuring the success of the workshops (Challen et al., 2009, Chapter 6). Since I know the size of UKRP classes and the total time available for them, I can use this as a measure of treatment intensity or programme quality. Table 2.6 uses a similar specification to the pupil fixed effect regressions in Table 2.4 and Table 2.5, but splits the intervention group into students who were in 'high quality' and 'low quality' workshops, with the former defined as workshop groups which contained 16 or fewer pupils and which were scheduled for at least 17 hours.[57] The coefficient on 'Treated*Postintervention*High quality' measures the impact of the programme at postintervention relative to the control group; I also report the p-values of chi-squared tests of the equality of the coefficients on high and low quality workshops at each point so that they may be compared. Here we see that pupils in high quality workshops had significantly different (more negative) changes in depression symptoms scores in all three time periods relative to pupils in low quality workshop groups; the scores of the latter group did not differ significantly from those of the control group at any point. For anxiety, the groups had either equal outcomes, or the low quality groups had significantly worse outcomes than both the high quality intervention group and the control group, and we see a similar thing for the self-reported behaviour score. There were no significant differences in absence rates between high and low quality intervention groups. For academic attainment, the high quality

[57] Table A2.8 gives the numbers of observations, pupils and schools in workshops with these characteristics. Using other thresholds does not substantially change the results, nor does calculating z-scores for the two measures, summing them and defining a cut point along this. Note that I am using school- and teacher-level inputs to the workshops as the quality measure (the time and classes scheduled), and not the actual hours received or realised class sizes, as student absence would cause these to be endogenous.

intervention group had significantly greater changes in scores than the control group at all three follow-up points, and significantly greater changes than the low quality intervention group at postintervention and one-year follow-up.

I cannot treat these differences in workshop intensity as exogenous, as they were not imposed by the research design and are likely to be correlated with school characteristics: more organised schools, or schools which were more committed to the programme, might be more likely to ensure that enough time was available and that classes were small. I would therefore expect this measure to be correlated with overall school effectiveness. In addition, I would not interpret differences in outcomes related to this measure as necessarily due to smaller class sizes and more hours: they may be due to other factors (such as better teachers, or more priority given to the programme) which led to greater programme impact. However, there was also variation in workshop quality within schools: of the 22 schools in this sample, 8 schools had both high intensity and low intensity workshops scheduled, while 12 had only high intensity and 2 had only low intensity. Moreover, since I include school*time dummies in these specifications, this should help to account for the school-level predictors of workshop quality. However, pupils assigned to different quality workshop groups had differing levels of the outcome variables at baseline even after controlling for school, which suggests that even conditional on attendance at a particular school, assignment to high or low intensity workshops was not quasi-exogenous. For example, pupils in high quality workshops had higher depression and anxiety scores at baseline than those in low quality workshops and the control group, so some of the differential impact here could simply be due to mean reversion: high baseline scores are very likely to decline through time. The impact of high-intensity groups could therefore be confounded with the non-random assignment to these groups: since I cannot include a workshop quality*post term for the control group, I do not know what the counterfactual change would have been in the control group. By contrast, when I split the intervention group by gender as in equation (5) above, there are both girls and boys in the control group and by including an interaction of girl*post for control students I can account for the counterfactual trends in the control group.[58] Nevertheless, when academic attainment or

[58] Another way of putting this is that I cannot include all the necessary interactions in these regressions: to estimate the coefficient on treated*time*characteristics, I should also be including interaction terms for treated*time, time*characteristics, and each variable alone. This is possible when the characteristic is gender, but not when it is a characteristic only observed in the intervention group. Treated*time*characteristics may therefore be confounded with time*characteristics which is omitted.

the self-reported behaviour score are the outcomes, the two intervention groups are actually quite well matched on the outcome at baseline and also look similar to the control group, yet we still see significant differences in impact between the two groups. Thus there may be some heterogeneity in treatment effects associated with workshop characteristics.

Table 2.7 presents similar specifications to Table 2.6, but splits the intervention group into intervention pupils who reported using the UKRP skills which were taught in workshops, and those who reported not using the skills. About 49% of students answering the question said that they used the skills. Using this variable reduces my sample size, as I have excluded intervention students who did not complete the survey or who did not answer this question. This is clearly endogenous, as I am conditioning on an outcome (use of UKRP skills), and students who use the skills taught in lessons, or who say that they do, may well be different from the ones who do not. For instance, students who are in greater need of the skills may be more likely to use them, or students with better teachers. Nevertheless, this could be an interesting test of mediation: does the intervention impact work through the skills taught to pupils, which they then use? A chi-squared test of the equality of the coefficients on the intervention impact (treated*time) for pupils who reporting using the skills and those who did not cannot reject that they are the same for every outcome in every time period. However, the results for the depression score and absence at postintervention look different and are significantly different in specifications not using pupil fixed effects.

Note that a major limitation of the regressions on high and low quality workshops and on using skills is that I cannot include all the relevant interaction terms, because only the intervention group had workshops. This means that I cannot include a high quality*post interaction for the control group pupils, so the coefficient on treated*high quality*post may be confounded with this. By comparison, when I split the intervention group by demographic characteristics below, I can include an interaction of the characteristic*post for the control group too, which should account for trends for these groups.

Table 2.8 looks at heterogeneity of programme impact by student characteristics. These regressions use the pupil fixed effects specification, pooling all three follow-up periods (postintervention, one-year follow-up and two-year follow-up) and splitting the intervention group into two by pupils' baseline characteristics. Each box represents a separate regression. The coefficients on each intervention variable give the programme impact for that group relative to the control group; the p-values of chi-squared tests of the equality of the coefficients compare the two intervention coefficients. The box in the top row at the far left shows that students with baseline depression symptoms scores above the median experience a significantly larger decline in depression scores as a result of the intervention than those who scored at or below the median.[59] This is driven by a much larger effect for the high baseline group at postintervention (effect size = 0.28 SD, table not shown), while pupils who started below the median only show an impact of 0.06 at this point. At later time points the results are not significant and are not significantly different from one another. Pupils with SEN also experience a significantly larger decline in their depression symptom scores as a result of the intervention, although this difference is due to greater intervention impacts in all follow-up periods. There is no significant heterogeneity in impact by gender or FSM. There are no significant differences in programme impact by pupil characteristics on the other outcome variables when the three follow-up periods are pooled together.

Robustness checks

The outcome variables come from different data sources, and as a result of this there are different numbers of observations for each outcome. Since I use the largest possible sample for each outcome in the analyses above, the different samples overlap but are not perfectly contiguous. As a robustness check, Table A2.3 and Table A2.4 report the same regressions as Table 2.4 and Table 2.5, but use a sample of pupils who have data for all four of the main data sources: depression symptom scores (pupil questionnaire); teacher-reported behaviour scores (teacher questionnaire); absence (NPD); and academic attainment (school databases). This gives exactly the same sample for all four outcomes, and very nearly the same sample for the anxiety and pupil-reported behaviour scores. Note that although there are 1,879 pupils and 15 schools included in each regression, the number of observations may differ because they may have data missing at some follow-up points but not others. This leaves the results for the depression score, anxiety score, absence from school, and pupil- and teacher-reported behaviour scores unchanged, despite a large reduction in sample size. However, there is now no significant impact on academic attainment in the fourth column (preferred

[59] Since the regression contains an interaction of high baseline score*time, this effect is not likely to be simply due to mean reversion in the depression score.

specification), and the estimates of intervention impact are much more sensitive to the specification used (except for the two-year follow-up results). However, the coefficient on ‘treated’ is now much larger, meaning that the intervention group in this sample had about 0.1-0.2 SD higher attainment at baseline. This suggests that this sample may not be the most appropriate one in which to evaluate the impact on academic attainment.

Seven schools included their entire 2007 cohort in the workshops. This means that there is no control group for these schools for five out of the six outcomes, since baseline data were only available for the 2007 cohort (the exception is academic attainment, for which all schools have a control group). Table A2.5 and Table A2.6 present the same regressions but include only the sample of schools with a control group. The results are substantially unchanged, although some of the coefficients are smaller and less statistically significant.

Pupils with higher attainment at Key Stage 2 show greater progress through secondary school (DCSF, 2009d), and I see this pattern in my data. One problem with the estimates of the impact on academic attainment is that the intervention and control groups do not look entirely similar at baseline: the intervention group has slightly higher academic attainment on average. Since difference-in-differences requires a common trends assumption in order to identify the intervention impact, and since pupils with higher progress faster, the small intervention effects detected could be spurious (and could also explain why the intervention impact appears to get stronger with time). I therefore present three more robustness checks with academic attainment as the outcome to attempt to correct for this. The first panel of Table A2.9 displays the same specifications as were presented for academic attainment in Table 2.5, but includes dummies for the Key Stage 2 level sum and an interaction of these with time. Pupils could achieve levels 2-5 in Key Stage 2 tests at this time; a pupil who achieved level 4 in all three of English, maths and science would therefore have a level sum of 12. There are ten possible values for this sum, ranging from 6 to 15. Including these interactions in the regressions will control for trends in academic attainment by broad baseline score. This causes the coefficient on ‘treated’ to become significantly negative (but small), but the estimates for intervention impact remain largely unchanged.

The second panel of Table A2.9 uses the original specification, but excludes pupils from the four schools which assigned higher attaining sets of pupils to the intervention group (see Chapter 1 for details). This results in a reduction in the sample size, causes the coefficient on ‘treated’ to become very small and nonsignificant in all specifications, but gives very similar intervention estimates to those obtained with the full sample. Table A2.10 controls for baseline attainment in another way - by conditioning on baseline Key Stage 2 sum, and running regressions for three groups of pupils: those with a sum of 11 or below (30%); those with a sum of 12 or 13 (40%); and those with a sum of 14 or 15 (305). By restricting the sample by baseline attainment I make it more likely that the intervention and control groups will match, as indeed they do in these samples. However, I am also in effect looking at intervention heterogeneity. Based on these results I find intervention impact on academic attainment for pupils at postintervention or one-year follow-up who had low or average attainment at baseline (70^{th} percentile or below). However, all pupils may have seen some gain in the region of 0.08-0.11 at two-year follow-up, and pupils with high baseline attainment had gains in academic at postintervention and one-year follow-up of 0.19 and 0.09 SD respectively. These robustness checks suggest that even controlling for any average differences in baseline attainment between the intervention and control groups, I can still find statistically significant intervention impacts, although these gains are concentrated amongst pupils with high attainment at baseline.

Mediation

There are three outcomes on which I see some impact, depression symptom scores and absence (at postintervention), and academic attainment (in all time periods). It is possible that changes in each of these variables could have been mediated by changes in another. Table A2.7 presents estimates of mediation following the procedure outlined by Baron and Kenny (1986). Importantly, a variable can only be considered as a mediator if it is significantly related to the intervention variable itself, and using this highly restricted sample of observations, only the depression symptoms score is significantly related to the treatment variable.[60] Using the absence rate and academic attainment as the dependent variable, I find no change in the coefficient on the dependent variable when the depression score is included in the regression. Thus the changes in outcomes due to the intervention appear to be unrelated to each other, with no clear mediation. The same is true when depression is used as the outcome and these other two variables are used as mediators, but I have not reported the results because the coefficients on the mediators were not statistically significant in this sample. Splitting

[60] This sample is more restrictive than the one reported in Table A2.3 and Table A2.4 because the latter only require that each pupil appears in the regressions for all main outcomes; here I require that each time point for each pupil is available for both the dependent and the mediator variable.

the regression into changes across different time periods does not change these results. Moreover, the change in the depression symptoms score is not correlated with the change in the other outcome variables, with the exception of a weak but statistically significant correlation with the change in the absence rate from baseline to two-year follow-up. Changes in absence rates and academic attainment are weakly correlated at every point (Table not shown). Thus I find no evidence of mediation of one outcome by another.

Discussion

Here I discuss my results and place them in the context of the psychology and education literature. In particular, I consider whether the magnitudes of the estimated intervention impacts on different outcome variables are meaningful; whether the measurement is adequate for a trial of this nature; the possible mechanisms for impact; scaling up issues and pragmatic trials; and threats to internal and external validity of the trial.

Magnitude of impact

The primary outcome of the evaluation was the depression symptoms score, and on this I found a small impact at postintervention only. The results were robust to using different specifications and samples. To give an idea of the magnitude of this effect (about 0.15 of the baseline standard deviation, or 1 point on the CDI scale), it is equivalent to moving from the 50th percentile of the control group at this point (depression symptoms score = 6) to the 45th percentile. Alternatively, it is equivalent to 75% of the gap in average depression score between control group children eligible for free school meals and those not, when measured at post; and about 70% of the difference in average score between children who live with both of their parents and those who do not (i.e. they do not live with at least one of their parents).[61] Brunwasser et al. (2009) review 17 PRP studies, finding an average effect size on depressive symptoms (CDI score) of 0.11 at postintervention, with an average effect at 12-month follow-up of 0.20 (based on 10 studies; 4 of them in universal samples), and with considerable variation between studies.[62] However, they also found a greater impact in

[61] Demographic factors are good predictors of wellbeing as measured by the depression score and the other outcome variables, but do not explain much of the variance. As a result, the average wellbeing gap between e.g. students eligible for FSM and others is relatively small. This is a common finding in the wellbeing literature e.g. Layard (2006), p.267.

[62] The effect size measure used in the meta analysis is Glass's *d*, (Glass, McGaw, & Smith, 1981), with

targeted samples: a postintervention effect size of 0.14, and a 12-month follow-up impact of 0.22 (both statistically significant); while in universal samples the observed effect sizes were 0.06 (nonsignificant) at postintervention, and 0.19 (significant) at 12-month follow-up. They do not report follow-ups later than 12 months. Thus my postintervention results are larger than those found in previous universal PRP studies, while the 12-month follow-up results are smaller.

Reviews of the literature on depression prevention interventions have generally found larger effects for targeted programs used with children with a higher risk of depression than for universal interventions (Stice, Shaw, Bohon, Marti, & Rohde, 2009), although a recent review found evidence for both targeted and universal approaches (Merry et al., 2011). One reason for this could be that most students drawn from a universal sample already have good mental health and do not need the skills relating to emotional regulation. It is notable that when I split the sample by baseline symptoms of depression, the impact on depressive symptoms is driven by a reduction in scores for students who started with higher levels of symptoms (effect size = 0.28 SD). Pupils who started with a baseline score below the median have a nonsignificant effect size of 0.06 at postintervention, which is the same as that found by Brunwasser et al. The former might approximate the result I would expect to see in a targeted or indicated sample, but the cut point used (median at baseline = 7 CDI points) has no clinical or screening significance and is much lower than that used for targeted samples in the PRP research (Kovacs, 2003; Brunwasser et al., 2009). Unlike the PRP review I find no significant effects in either group in later follow-up periods, although using a higher cut point in my depression symptoms regressions does result in larger (more negative) effect sizes for the high-baseline group, and possibly also a longer duration of impact. Thus although the impact on depression was small overall, it may have had a more substantial impact in particular groups.

I can estimate the impact on pupils' absence using the control group's absence rate in 2007-08. This was 7.4%, which is 22.4 sessions, or 11.2 days per year (there are two sessions in a school day, morning and afternoon, and an average of 304 sessions for that academic year). The 0.14 SD improvement in the programme group represents a decrease in the absence rate of about 0.8 percentage points, or 11.5% of the control

---

Hedge's correction for small N used to generate the overall means (Hedges, 1981). I have estimated an effect size for my sample using regression analysis, so my results are not strictly comparable.

group's absence rate. Thus intervention group students attended 2.6 more sessions or 1.3 more school days over the 2007-08 academic year than control group students. Another way of looking at this is in terms of the gap in absence rates between FSM students and others in the control group. In 2007-08 this gap amounted to 10.8 sessions or 5.4 more days of absence for FSM students than for other students, more than one whole week of school. The programme impact is therefore equivalent to 24% of this gap. Whereas the psychological measures are assessed at a point in time, the absence measure is accumulated over the course of the year and so the short-run outcome (2007-8 absence) includes the time period when students were attending workshops. Students generally enjoyed the workshops, and some teachers commented that attendance was higher on UKRP days because students did not want to miss the lessons (see Challen et al., 2009, for more information on pupil and teacher satisfaction with the programme). A major reason cited by pupils for truanting is boredom, and many do so with the collusion of their parents, so it is plausible that making lessons more enjoyable would raise attendance (Malcolm, Wilson, Davidson & Kirk, 2003). If this is the case, it is possible that the effect observed could be entirely due to increased attendance on days with UKRP lessons. This is important both because one would expect the effect to disappear as soon as the workshops finished (which is what we observe), and because we might schools might place a lower value on higher attendance if this is only for a non-academic lesson. However, a session will usually consist of two or three lessons, so even if students were only motivated to attend school for a UKRP lesson it is likely that they attended others as well.

Despite the lack of a sustained impact on the depression scores and absence from school, the intervention impact on academic attainment lasts until two-year follow-up. To give some indication of the magnitude of this impact, in this sample the average difference between control group students entitled to free school meals and those who are not in the combined standardised scores in English, maths and science at post (end of Year 7) is 0.63 SD; it is 0.67 at one-year follow-up; and 0.79 at two-year follow-up. The programme impact is therefore about 12-17% of this gap at each point. It is also equal in size to the gender gap in attainment at two-year follow-up (gender gap=0.13 SD, with girls outperforming boys on average), 39% of the gender gap at post and 36% at one-year follow-up (gender gap=0.25 and 0.22 SD respectively). An alternative way of interpreting this is in terms of the change (average improvement) in combined academic attainment year by year. Between the baseline and the end of the first year (postintervention), average attainment in the control group increases by 0.60 SD, so the treatment effect at this point is an increase in the *change* in academic attainment of 16%; at 1-year follow-up attainment in the control group is about 0.96 SD higher than at baseline, so the treatment effect is an 8% improvement in the change in attainment since baseline; at 2-year follow-up the control group attainment is 1.62 SD higher than at baseline, so the treatment effect at this point represents a change in treatment group improvement since the baseline of about 8%. Table 2.9 also presents the impact at each point in terms of National Curriculum levels and percentile points.

What is interesting about these results is that the change in the *growth rate* in academic attainment for the treatment group happens entirely between the baseline and post intervention, with the improvement about 0.1 SD higher in the intervention group compared to the control group. This results in a difference in levels at post intervention, which is maintained in subsequent years, but with the same growth rate: between post intervention and 1-year follow-up there is no significant difference in the academic improvement of the treatment and control groups; and the same is true between the 1-year and 2-year follow-ups.[63] Thus the effect of the programme is to shift up the level of academic attainment in the short run, after which intervention group students continue to develop on a higher but parallel trajectory to the control group, with the result that they are still outperforming control group students by about 0.1 SD two years after the end of the programme.

In terms of Cohen's classification (Cohen, 1969), the effect size on academic attainment is small (about 0.1 SD). However, it is important to view effect sizes in context (Glass, McGaw & Smith, 1981), and in terms of educational interventions carried out in developed countries this could be a substantial effect, particularly if it persists (Coe, 2002; Higgins, Kokotsaki & Coe, 2012).[64]

I found no impact at all on anxiety symptom scores, despite the high correlation between anxiety and depression scores. I also found no impact on behaviour scores. This is perhaps not surprising, given that the intervention was designed primarily to

[63] I calculate this by running similar regressions as for Table 2.5, but adjusting the baseline each time so that the effect at one-year follow-up is calculated relative to post intervention, and two-year follow-up relative to one-year follow-up. In Table 2.5 the comparison is always the baseline level of academic attainment from summer 2007.

[64] Many reviews report much larger effect sizes in educational attainment. However, large effect sizes often come from differences in design, for example the impact is not the result of an intervention (e.g. a review of the impact of socioeconomic background on attainment); outcomes are not being compared with those of a control group, so do not test the intervention against ordinary provision; or proximal outcomes are used such as achievement on phonics tests rather than actual attainment in reading and English (Hattie, 2009).

combat internalising rather than externalising symptoms, and the behaviour measures were primarily designed as screening tools for disorders. However, schools and teachers involved in UKRP were often more concerned about behaviour problems than about internalising problems, possibly because the former had more impact on others in the school, so the programme may have less appeal to schools as a result of this.

Measurement

One reason for the small or no impacts on pupils' depression and anxiety symptoms and behaviour scores could be that students drawn from a universal sample have good mental health and social functioning and simply do not need the skills taught in the workshops. Certainly, conceived of as a depression prevention programme, pupils who are very unlikely to develop depressive symptoms during the three years of the trial are unlikely to see a clear benefit. However, in my satisfaction survey of students I found that students with high levels of baseline symptoms were less likely to say that they used the skills they learned in the workshops (Challen et al., 2009, Chapter 5). In addition, the skills which students said they used most often were the interpersonal skills (negotiation, assertiveness, and compromising), and the various in-the-moment emotion management techniques, which would have relevance to the lives of all children, not just children with poor mental health.

Another explanation might simply be the lack of sensitivity in the measures I am using. As discussed in Chapter 1, the depression inventory (the CDI minus one item) provides a scale ranging from 0 to 52, yet at baseline 50% of students score 7 or below, and only 8% scored 19 or above, which would indicate significant symptoms of depression. Because of this, there may be a ceiling effect: students without many symptoms do not have much room for (relative) improvement unless the control group students experience a very significant deterioration in their scores, and so on this measure I cannot discriminate between the wellbeing levels of a large proportion of the population. The same is true for the anxiety score, and for the absence rate: at baseline 8% of the sample have no absences at all, and more than 50% have an absence rate below 4%. Yet I find no evidence of heterogeneity in programme impact by baseline level of absence, or for the anxiety score (I find no impact on anxiety for any group at any time).

The measure of academic attainment is perhaps less subject to ceiling effects than the other measures, and is a routinely used and collected outcome easily understood by participants. It also measures general academic attainment in the three headline National

Curriculum subjects, rather than performance on specific tests or tasks.[65] However, it has other limitations. First, it is only available for 20 of 22 schools. Second, although the baseline measure is taken from national exams marked externally to the school, all three follow-up measures come from school databases and are based on teacher assessments, which may include exam marks but which are likely to be a reflection of work done over the course of the year. Standards are therefore perhaps more likely to vary between schools and between different teachers. Because these data are routinely collected by schools, I was able to obtain additional academic follow-up data for the 2006 cohort meaning that I have a within-school control group for all schools for this outcome. Yet there may be problems with using scores from different cohorts: grade inflation in Key Stage 2 and Key Stage 3 exams could superficially create differences in attainment between cohorts when in fact none exist. For example, 79% of pupils who took Key Stage 2 tests in 2006 achieved at least a level 4 in English, while 80% of pupils who took them in 2007 achieved this level (DCSF, 2006; 2007b). This would result in a difference in the *baselines* for the two cohorts: the attainment of the 2006 cohort would be biased downwards. But this would bias against finding an impact of the workshops: conditional on baseline grades, 2006 pupils should show a greater growth rate between baseline and follow-up periods. More importantly, many of the 2006 cohort sat Key Stage 3 tests before they were abolished, so some of the data from schools' databases may have reflected their grades on tests rather than teachers' estimates of their performance. Gibbons and Chevalier (2008) compare teacher assessments and national exam grades, finding that teachers' grades suffer from an aversion to extremes: higher ability students are graded lower by teachers than by tests, while lower ability students are graded higher by teachers than by tests. If this pattern of responses were the same in the data I have, we would expect the distribution for the 2007 cohort 2-year follow-up grades to be truncated relative to those for the 2006 cohort. Whether this had an impact on the estimated intervention impact would depend on the relative proportions of high-attaining and low-attaining pupils and the extent of the under- or over-grading by teachers: if the former dominate, this will bias downwards the estimate of intervention impact, because high ability students who sit exams in 2006 will be able to gain higher grades than high ability students who are assessed by

[65] By contrast, other education studies may use outcomes very specifically related to the programme being tested rather than general academic attainment (e.g. spelling tests to evaluate spelling programmes), or proximal outcomes of less visible impact in everyday life (e.g. attitudes towards education).

teachers in 2007. If low ability students dominate I would see the reverse. However, the two cohorts do not have significantly different academic attainment at baseline or 2-year follow-up, the standard deviations for each are not significantly different (and are slightly smaller for the 2006 cohort), and the distributions overlap almost exactly (Figure A2.1).[66]

Given the broader aims of UKRP to promote resilience in all children rather than preventing mental illness in only some, the depression and anxiety symptoms scores, and even the behaviour scores, may not be the best outcome measures. However, the use of two other outcomes – absence and attainment – which are only weakly correlated with each other allows me to look at many dimensions of wellbeing and good functioning for these children. The impacts on attainment, absence and the depression symptoms score were not correlated with each other and did not mediate the impact on each other, suggesting that different pupils experienced different gains from the intervention and that a range of pupils across the distribution may have benefitted. For example, the reductions in depressive symptoms were primarily seen in pupils with elevated symptoms at baseline and SEN pupils, while the gains in academic attainment were seen for pupils with a high baseline academic score and pupils without SEN (although these differences were only statistically significant at particular time points). There was no suggestion of heterogeneity in the impact on absence. Thus several different groups of pupils may have benefitted from the intervention but in different ways. In the absence of a reliable measure of resilience (see Chapter 1), using this combination of measures may be a second-best solution.

Mechanisms

As discussed in Chapter 1, possible mechanisms for programme impact include use of the skills taught by the curriculum; developing relationships with teachers and pupils; or a direct impact on emotional wellbeing and school attachment due to enjoyment of the workshops. My results do not show a strong relationship between pupils' reports of using the skills and intervention impact, although pupils who claim to use the skills may see a larger postintervention reduction in depression symptoms scores and absence rate. There does not appear to be any relationship between this variable and intervention impact at any other time or for other outcomes. If pupils' reports are a good

[66] This is based on adjacent cohorts in the same schools; schools with information for only one cohort are excluded.

indication of whether they used the skills or not, this would seem to suggest that any observed intervention impact is not primarily due to learning useful life skills. However, it is not clear that pupils' reports are accurate: some pupils may give the socially desirable answer and say they use the skills when in fact they do not; while others may use the skills without being aware that they are doing so.[67] If this variable is noisy it may lead to attenuation bias, meaning that we would be less able to see differences in outcomes between the two groups. In addition, not all intervention pupils answered the question on skills, and there are clear differences on the baseline values of the outcome variable between pupils who report using the skills; pupils who report not using them; pupils who do not answer the question; and control group pupils. For example, pupils who report not using the skills had significantly higher baseline depression scores, a significantly lower absence rate, and significantly higher academic attainment than the control group, while pupils who did not answer the question have significantly lower academic attainment than the control group. Moreover, students in high-intensity workshops are significantly more likely to say that they use the skills than students in low-intensity workshops. Thus any differences could be due to mean reversion. Nevertheless, I do not have any other measures of whether students actually use the skills or not.

The intervention impact on academic attainment does not appear to be related to (reported) use of the skills. Moreover, the changes in depression symptoms, academic attainment and the absence rate are not related to one another. This could be because different pupils are benefitting along different dimensions, but the mechanisms of impact on each outcome may also be different. The magnitudes of all three of these outcomes were associated with my crude measures of workshop quality, which might suggest that the second two mechanisms (relationships and enjoyment of school) could be important. Given the apparently strong association of workshop quality on academic attainment, but no effect of using skills or of improvements in the depression score on attainment, it is consistent with my data that improvements in attainment come from changes in the school environment and greater school attachment rather than directly through using UKRP skills. However, my measures of workshop quality and of use of the skills are noisy.

[67] For more detail see Chapter 8 in Challen et al. (2009), which presents the qualitative work on UKRP by Philip Noden and Anne West, and includes discussions of whether pupils used skills based on interviews with pupils and teachers.

The intervention could also have had an impact on pupils through an impact on teachers. Most facilitators who responded to the satisfaction survey or who were interviewed said that they used the UKRP/CBT skills themselves in their everyday lives (Challen et al., 2009). An evaluation of the impact of UKRP training on teachers' wellbeing found large and statistically significant impacts on self-reported happiness and flourishing, though no impact on symptoms of depression or job satisfaction (Murphy, 2011).[68]

The subject matter of the workshops could also uncover child protection issues, through inadvertently seeming to encourage pupils to disclose problems (Challen et al., 2009; personal communications). If this allowed schools to seek specialised help for pupils at an early stage, then part of the impact we see could be due to this: the impact of detection plus specialised help for troubled pupils.[69]

There was no standardised alternative treatment, and provision for the control group varied between schools. This is not a threat to the validity of the experiment, indeed, part of what makes this a pragmatic trial is the lack of restriction on what the alternative treatment could be (Thorpe et al., 2009). However, most control group students were taught in groups of around 30, so I am not able to disentangle the impact of the UKRP curriculum from the impact of smaller class sizes for one lesson a week whilst teaching PSHE (the usual alternative treatment). The literature on the impact of smaller class sizes tends to find small or no effects. For example, Hattie (2009, pp. 85-88) reports an average effect size on attainment of 0.13 SD for moving from 25 to 15 pupils in a class. Note that these results apply to the class in which the academic lesson is taught; no academic lessons are taught during the small UKRP classes. Thus it does not seem likely that the intervention impact I observe on academic attainment is entirely due to a smaller class size. Hattie notes that studies finding in favour of reduced class sizes tend to find that the greatest improvements were seen in teacher and pupil working conditions, and it may be that the small groups and the subject matter of UKRP promoted positive experiences for pupils and teachers. One previous study of the PRP curriculum that used a placebo control group (small classes in which students could discuss problems, but which did not teach resilience skills), did not find significant

[68] The study used a non-randomised wait-list control design, identifying the intervention impact using difference-in-differences.
[69] Note that early detection of problems as a result of feedback from the research questionnaires should not be responsible for the intervention impact, as feedback was also provided for control group pupils.

differences in depression symptoms between the placebo and the PRP groups (Gillham et al., 2007). In general, studies which compare interventions against placebo controls may produce smaller effect sizes (Calear & Christensen, 2010; Stallard et al., 2012). Moreover, UKRP facilitators said that the smaller class size was an important part of UKRP, although they also rated the curriculum as important (Challen et al., 2009, Ch. 6). The decision about whether or not to adopt UKRP in schools should be based on whether it is more or less effective than current school provision, so the comparison of the intervention being tested with ordinary practice (larger class sizes) does seem to be the appropriate one to make (Roland & Torgerson, 1998). However, we should also be concerned with cost and cost-effectiveness, and if a large part of the programme impact on pupil wellbeing is due to the smaller class size, this impact could be obtained at lower cost by implementing a programme of *smaller classes once a week*, as opposed to adopting the intervention evaluated here, which is *smaller classes once a week plus PRP curriculum.* Because of the design, this trial can only answer the question of whether this intervention has an impact on pupil outcomes; I cannot judge whether it does so in the most cost-effective manner, although this is understandably of concern to policymakers and may merit further research.

Scaling up issues

Compared to previous PRP research in universal samples, I find a greater impact at postintervention (though not substantially so) and a smaller impact at 1-year follow-up. It is common for pragmatic trials of interventions to find smaller effects than explanatory trials. The latter tend to have smaller sample sizes, more experienced facilitators, and maintain tighter control over intervention quality (Malti et al., 2011). Extensive exclusion criteria can also reduce the heterogeneity of the population, making it easier to find an effect (Patsopoulos, 2011). Comparing interventions to no provision rather than ordinary provision may also bias in favour of finding an effect. Trials conducted by independent researchers may also produce smaller effect sizes than those carried out by the research team that developed the intervention. This could be due to the fact that these trials are often explanatory trials rather than scaled-up implementations, but higher programme fidelity when the development team is involved or unintended bias from researchers who are motivated to achieve positive results could also be involved (Eisner, 2009; Petrosino & Soydan, 2005). This trial was designed with input from the programme developers, but all analyses were planned and carried out independently. Programmes may also be less effective when delivered by school staff rather than by specialists (Calear & Christensen, 2010), yet using school staff is more likely to be sustainable and possibly the only way that interventions could feasibly be used on a wide scale.

Yet it is not obvious that the UKRP workshops were of much lower quality than those in previous studies. There was certainly less oversight of the intervention and a frequent lack of organisation within schools, with many schools cancelling workshop groups or failing to schedule time right from the start. Facilitator selection was often fairly arbitrary, and dependent on teachers' holiday plans rather than their appropriateness and availability to teach the intervention. Yet facilitator quality is likely to be important to the success of the programme (Weisz et al, 2006). On-going support for facilitators after the initial training period was also largely perceived to be inadequate and unhelpful, and different support models are now used (Challen et al., 2009; Challen & Bailey, 2012). Recent research suggests that more intensive on-going support may be important to the maintenance of programme quality in dissemination (Lochman et al., 2009). However, programme dosage was generally high – most UKRP pupils received most lessons – and compared favourably with previous studies primarily because the lessons were scheduled during the school day (see Chapter 1, 'Programme adherence'). Previous PRP trials have assessed workshop quality through rating a recordings of sessions, scoring facilitators on their ability to teach programme content and respond appropriately (Brunwasser et al, 2009), and these may be better indicators of workshop quality than the 'input-based' measures I have available (hours scheduled and group size). Nevertheless, these inputs are likely to be associated with how seriously each school took the programme and therefore be proxies for intervention quality, and these measures are associated with a greater impact on several outcomes, particularly depressive symptom scores and academic attainment.[70] Thus although I cannot compare the quality of UKRP workshops with the quality of the intervention in other studies, within my sample higher quality workshops were associated with better results.

Internal validity

One major shortcoming of this study is the lack of randomisation of condition assignment. Random assignment may well produce the same effect sizes as non-random

[70] Although note my concerns above about different workshop groups being associated with differential trends which I cannot control for.

assignment (Oliver et al., 2010), but it is usually more credible: we have a better basis for believing that the results are due to the intervention rather than other factors (Angrist & Pischke, 2009). In this study it is noticeable that classes assigned to the intervention had higher levels of prior academic attainment than the control group, suggesting that assignment was not truly arbitrary. In some cases I know this was chance, but in others teachers felt that the programme would be easier to teach to higher-attaining classes, or that low-attaining pupils could not spare the time for the intervention. I do not think this would have a large impact on the estimated intervention impact on depression symptoms scores: pupils with SEN (who have lower average attainment) experience larger declines in their depression scores as a result of the intervention than other intervention pupils, so my results should *underestimate* the intervention impact seen in a randomly selected sample. However, this is a problem in interpreting the results on academic attainment: pupils with higher baseline values of academic attainment progress faster, and the difference-in-differences strategy I adopt assumes equivalent trends in the intervention and control groups. I attempt to correct for this by conducting robustness checks on samples with particular baseline scores, and by including a trend term for each value of the baseline score. But there may still be other unobserved differences between the two groups, and it is likely that random assignment would have given more convincing results.

I have mentioned the importance of the common trends assumption in estimating the intervention impact, but primarily in relation to the composition of the intervention and control groups: if pupils with different characteristics experience different trajectories on average, and the intervention and control groups are not well matched on these characteristics, they will not be good counterfactuals of one another. However, there is another way in which the common trends assumption may be violated: through shocks associated with one group but not the other. For example, if treatment assignment is non-random, and is related to membership of another group (e.g. class), then there may be many things associated with class membership which affect wellbeing or academic attainment other than the intervention. This confounding of the intervention impact with other policies or pre-existing conditions is clearly a problem when trying to estimate difference-in-difference estimates of a policy change in neighbouring states (as in Card & Krueger, 1994) – there are many things going on before and after the policy in each state which are unrelated to the intervention of interest and which could affect the outcome. However, it is not so clear this is would be a threat in the current case, except when pupil characteristics differ substantially between the intervention and control groups. This is because school classes were put together in July or September 2007, and did not exist before this because pupils were at different schools and it was not even known which secondary schools they would be attending – any pre-existing trends in pupils' outcomes therefore cannot be attributed to class membership, except to the extent that this is related to pupil characteristics (for example, higher-attaining pupils have been doing better at school for some time, and this ensures that they are placed in a high-attaining class if classes are setted by ability). More plausible is that classes assigned to UKRP workshops also experienced other shocks which the control classes did not, and that these drive any results. However, in all schools pupils were taught in classes other than the UKRP class group, reducing the risk that any class-specific shocks will be perfectly correlated with intervention assignment because this class was only one of many. Moreover, many classes will have had their membership changed or have been disbanded altogether at the end of the UKRP intervention and each year subsequently, leaving less chance of class-specific shocks perfectly correlated with the intervention assignment. Finally, random positive or negative shocks to intervention classes would not be a problem unless they are the same across schools, and there is no reason for thinking that this should be the case unless the classes are distinct in terms of pupil composition (e.g. we might expect high-attaining pupils to have a similar trajectory across schools; classes with a high proportion of high-attaining pupils will therefore have different trends to those with fewer high-attaining pupils). For these reason I assume that differences in trends due to pupil composition are more likely to cause bias than shocks restricted to either the intervention or control groups.

One robustness check used with difference-in-differences is to use a placebo test: assume an intervention happened earlier or later than it in fact did, and see if a similar estimate of intervention impact is obtained (Angrist & Pischke, 2009). Doing this for an earlier time period might make some sense in this case: I could test whether pupils' trajectories were different even before the experiment started. However, I do not have data available for any outcome prior to the baseline, so this is not a test I can perform.[71] Moreover, assuming the intervention happened later than it did would not tell me much: some programmes demonstrate a 'sleeper effect' where the main impact is not seen until

[71] Data on Key Stage 1 results is available, giving pupils' academic performance at age 7. However, there are only five possible outcome categories for these tests, meaning that they are not very sensitive.

six months or more after the end of an intervention (e.g. Brunwasser et al., 2009), so this would not mean that the intervention effect was not plausible. For these reasons I do not carry out placebo tests of this kind.

Since all schools contain both intervention and control students, there is a possibility of externalities or spillovers from the intervention to the control group (or the reverse). Positive spillovers would bias downwards the estimate of the effect of the programme, as control group students would receive a partial treatment. Two possible ways this could operate are through staff and students: school staff trained as workshop facilitators could teach the programme skills to students outside the workshops, or behave differently towards them; and intervention students could influence their control group peers through social interactions. For example, if intervention group students behaved better in school as a result of the intervention, any control group students in their classes would also benefit. If this were a substantial problem, then I would expect to see improvements in the intervention group's outcomes mirrored by improvements in the control group. In fact, the large majority of students in both groups do not show large changes in their depression scores: between the baseline and post intervention points, 54% of intervention group students and 56% of control group students have a change in depression score between -3 and +3 (measured in points on the CDI scale of 0-52), with the modal change for both groups being zero. Of course, if the overall trend without the programme would otherwise be a steady worsening of scores, with students becoming more depressed as they got older, then this lack of change could itself be seen as evidence of an 'improvement' in scores as it would reflect maintaining a steady level of depressive symptoms rather than worsening over time. However, based on the population scores reported in Kovacs (2003), this does not appear to be the case. I see a similar pattern for attendance. There is a steady upward trend in academic attainment throughout the three years of data collection, but this is seen in both the 2007 cohort and in cohorts several years older than them. It is not plausible that positive spillovers occurred from the treatment group to the control group one or two years before the programme was implemented. For these reasons I do not think that spillovers are likely to have had a substantial impact on the results I obtain. Even if they did, they would be likely to bias downwards my estimate of programme impact rather than inflate it. If negative spillovers from the control group to the intervention group were observed, this could mean that a scaled-up programme would have a larger impact (and so external validity is also threatened). However, based on evidence presented above I do not think that spillovers are a serious threat.

Pupils and teachers were not blinded to intervention condition, but many pupils probably did not know that they were receiving UKRP lessons as part of an experiment. The teachers who administered the questionnaires were not associated with UKRP, and the questionnaires did not explicitly link the research to UKRP lessons. Absence data is obtained from national databases, and it is very unlikely that schools could have manipulated these data. Academic attainment was obtained from schools' databases, and most grades would have been assigned and recorded by teachers not involved in the intervention. There is therefore little reason to believe that participants were motivated to respond positively in order to make the intervention appear more effective.

External validity

The study was intended to be a pragmatic trial because its main aim was to inform policy in this area. Following the criteria set out by Thorpe et al. (2009), the evaluation is highly pragmatic along some domains but not others. For example, there are no exclusion criteria – all pupils within the relevant cohorts at these schools were included in both the intervention and evaluation if they did not opt out; facilitators and schools had considerable flexibility in how they taught the intervention; facilitators were drawn from regular school staff, few of whom had prior experience in this area; there were no requirements as to what the control group's alternative treatment should be, and this varied between schools; the alternative treatment was taught by regular school staff; measurement of facilitator adherence to the intervention and student participation in the workshops was only crudely and unobtrusively monitored, and no incentives were provided to increase fidelity; and results are reported for all pupils with available data in an intention-to-treat fashion, with no exclusions based on intervention compliance or other factors. However, there are two domains on which this trial was less pragmatic: first, pupils and teachers were surveyed three times postintervention to gauge programme impact on these outcomes, and these data would not be collected in routine practice. Second, the primary outcome (depression symptoms score) is probably not meaningful to study participants. These two factors are mitigated by the use of other routinely collected outcome data obtained unobtrusively from databases, which is meaningful to participants: absence from school and academic attainment are easily understood, and are both of immense interest to teachers and parents, with school-level data on these are reported in school performance tables published by the Department for Education (see, for example, DfE, 2013f).

The trial therefore does appear to be largely pragmatic in its design, which is important for external validity: how well these results would translate to an impact in routine practice. Moreover, because UKRP was integrated into the ordinary school curriculum, many pupils were not aware that the intervention impact was what the questionnaires aimed to measure. This may reduce the risk of a Hawthorne effect biasing the questionnaire responses.

In order for the trial to have good external validity, the sample should also be typical of the population of interest. Although there were no exclusion criteria applied within schools, only certain schools signed up for the programme and these were more deprived than both the population of the LAs from which they were drawn and that of England as a whole. The results might therefore demonstrate the impact of the programme in an above-averagely deprived sample. Nevertheless, the recruitment strategy meant that there was considerable diversity in the regions, schools and pupils involved, and since there is little significant impact heterogeneity by FSM eligibility the results may generalise to a less deprived population.[72] More importantly, the schools and staff involved in the evaluation were probably more interested in PSHE than the schools and staff who were invited but chose not to take up places. We would expect that schools with an interest in this area would already be providing a good PSHE curriculum, so the alternative treatment offered should set a higher bar for finding an impact for UKRP.

One further issue is whether schools could afford to run the intervention under normal conditions. As described in Chapter 1, UKRP is now used in more than 85 schools and more than 800 facilitators have been trained. Schools and LAs have funded the training places and schools have provided the costs of staff cover for training days and smaller classes. School organisation and the selection of facilitators have probably improved since the pilot, partly because school have to invest in the intervention rather than receiving it for free, and partly because the information provided to facilitators has improved (Challen & Bailey, 2012; Challen, Machin & Gillham, forthcoming). Facilitators may also become more skilled with practice: in the first year of the

---

[72] If split by follow-up period, FSM pupils in the intervention group experience a significantly greater reduction in anxiety scores at postintervention. However, there are no significant differences for any other variable or at any other time point.

programme they were new to the intervention and due to the small number of workshops taught per facilitator had relatively limited opportunities to practise.

Thus the intervention is clearly cheap enough to use in schools, and the implementation quality of the first year of the workshops is unlikely to have been any better than in subsequent years (in fact, it may have been worse). This should set a lower bound on the estimated programme impact. Overall, the trial was largely pragmatic and should have good external validity due to its close imitation of ordinary practice.

Cost effectiveness

I can use the cost estimates presented in Chapter 1 to assess the cost effectiveness of the intervention on the outcome evaluated in this chapter. Table 2.9 sets out estimates of cost effectiveness in terms of standard deviations of each outcome variable. The table uses the incremental cost-effectiveness ratio (ICER), defined as:

ICER = (cost of intervention – cost of control)/(effect in intervention – effect in control)

Since the control group will not incur any costs which the intervention group does not also incur, I can use the estimates of the per pupil intervention costs set out in Chapter 1 in the numerator. Effect sizes were also defined relative to the control group, so these are included in the denominator. In effect, this gives the cost per standard deviation of improvement in each of the three outcome variables on which I see an impact of the intervention. A one standard deviation reduction in the depression score would be equivalent to a shift from the 50th percentile of the control group at postintervention to the 1st percentile, but would represent a much smaller shift at the top of the distribution: either from the 80th percentile to the 56th; or the 99th percentile to the 96th. It is not clear whether any of these shifts would be clinically meaningful. For absence, a one standard deviation reduction in absence rates at postintervention would represent a shift from the median to the 1st percentile, or from the 80th percentile to the 56th percentile. For academic attainment a one standard deviation improvement would represent a shift from the median to the 84th percentile at postintervention (or 20th to 47th percentile shift); median to 79th percentile at one-year follow-up (or 20th to 47th percentile shift, again); and a median to 77th percentile shift at two-year follow-up (20th

to 42nd percentile). These are the shifts which would correspond to the stated costs, which are clearly quite different from the intervention impacts actually observed.

Note that although I have calculated separate costs for each outcome at each point, these would only be used separately if a policymaker were interested only in a reduction in depression symptoms scores, for example, with no interest at all in impacts on absence or attainment. Expenditure on the intervention will result in all of these outcomes, so if they are all valued the decision to use the intervention should focus on the per pupil costs of the intervention listed in Chapter 1. Moreover, the costs of the improvement in academic attainment are calculated separately at each follow-up point. Since I found no impact on anxiety symptoms scores or behaviour scores I have omitted these from the table. The figures also do not take into account other benefits that the programme might bring such as earlier identification of pupils in need of specialised help (apart from the extent to which this results in lower depression scores or higher attendance or attainment), or any impacts on teacher wellbeing.

The 2007 costs represent the actual cost of the workshops I am evaluating in this study; the 2009 costs represent the cost of implementing UKRP in schools now or in the future – this latter cost estimate should be the one of interest to a policymaker. Note that per-pupil costs partly depend on the number of workshop groups each trained facilitator teaches (because of the high fixed cost of the training), so here I have given two scenarios for each of the 2007 and 2009 costs, representing the scenarios shown in Table 1.16 (maximum - facilitator teaches 225 pupils; minimum – facilitator teaches 45 pupils). Further scenarios could be estimated using the cost information presented in Table 1.16. Table 2.9 also includes 95% confidence intervals around the point estimates (again in terms of standard deviations) and estimates of the cost-effectiveness of the intervention based on the upper and lower bounds of the intervention impact represented by these.

There is little information on the cost-effectiveness of universal mental health promotion programmes which reduce mental distress (reduce symptoms) but which do not necessarily prevent clinical disorders. There is information on the cost-effectiveness of targeted prevention programmes (e.g. Mihalopoulos & Vos, 2013), and on the cost-effectiveness of preventing mental health *disorders* (e.g. Eisenberg & Neighbors, 2009). Much of this evidence defines health outcomes in terms of quality-adjusted life years (QALYs) or disability-adjusted life years (DALYs). It is outside the scope of this study to convert my outcome measure (effect size on depressive symptoms) into one of these measures, but it would be possible to do so (Mortimer & Segal, 2008). An alternative measure is to use cost-benefit analysis and examine returns to investment. Effective prevention of conduct disorder through school-based programmes has high returns to investment, primarily because of the costs of crime, but also because of longer-run impacts on educational attainment, employment and healthcare expenditure (Knapp, McDaid & Parsonage, 2011; Scott, Knapp, Henderson & Maughan, 2001). However, I found no impact of UKRP on behaviour. School-based interventions to reduce bullying may also produce a positive return, with the longer-run gains a result of higher earnings and employment due to better educational attainment, school attendance and mental health (Knapp et al., 2011). However, as Knapp et al. note, evaluations of anti-bullying programmes have used relatively short follow-up periods and there is little information on whether the impacts are sustained over time. This is a serious consideration in assessing the value of the postintervention improvement in depressive symptoms scores: since it does not last, we cannot assume that the intervention will have a permanent impact on pupils' mental health or in other areas of their lives. Moreover, much of the estimated return to anti-bullying programmes comes through higher earnings due to higher educational attainment, and since academic attainment is the only outcome for which the impact is sustained until two-year follow-up it may be best to base estimates of cost-effectiveness on this.

Given the cost of UKRP, it might not provide good value for money to offer it universally if the only aim is reductions in depressive symptoms. Universal mental health prevention programmes are often more popular with schools, partly because of a lack of stigma attaching to participation (Merry et al., 2011; Offord, Kraemer, Kazdin, Jensen, & Harrington, 1998). But targeted programmes have larger effect sizes on average, and will therefore offer better value for money conditional on the same number of pupils being taught by a trained facilitator. However, note that because of the large fixed cost of the initial training course, the per pupil cost of UKRP declines with the number of workshop groups each facilitator teaches. If offering UKRP universally allows a facilitator to teach many more pupils, the per pupil cost of the programme tends towards the long-run variable cost of staff time and materials, making this better value for money. The greatest cost effectiveness will therefore depend on the number of pupils a facilitator would teach if the programme were targeted or universal and on the expected impact in these groups.

The impact of UKRP on educational attainment is small, and equivalent to around

1-2 months' progress according to the EEF's guidelines, while the cost estimate per pupil would be moderate to very large (Higgins, Kokotsaki & Coe, 2012). These figures make it very similar to many other interventions summarised in the EEF report just cited, but it is certainly not the best value for money for a policymaker interested in educational attainment alone. Moreover, since the UKRP impact on academic attainment is concentrated in pupils who already have above-median attainment at baseline, the intervention may offer particularly poor value for pupils who are low attaining (although conversely, these pupils may receive more benefit in terms of mental health symptoms).

I do not have information on the effect sizes and cost effectiveness of other strategies to promote pupils' attendance at school. However, absence is held to be associated with poor attainment, disruptive behaviour and poor safety, and is sufficiently important to be reported in school performance tables (Malcolm, Wilson, Davidson & Kirk, 2003; DfE, 2013f).

Thus for any individual outcome it is not clear that UKRP would offer returns on investment or good cost-effectiveness relative to other available interventions. However, the cost-effectiveness of UKRP partly depends on efficient use of trained staff (and type/grade of staff): because of the large fixed cost of the training course, training staff who are then unable to teach many pupils is a major contributor to the average per pupil cost of the intervention. With good school organisation and considering the impact on all three outcomes together, the intervention may be more worthwhile.[73]Moreover, the approach to mental illness prevention which compartmentalises mental, social, behavioural and educational problems may be inadequate: programmes such as UKRP which aim to have a broader impact on pupils' functioning may provide a better first step towards prevention (WHO, 2004).

Conclusions and policy implications

I evaluate the impact of a large-scale implementation of the Penn Resiliency Program in a universal sample of pupils at English secondary schools. The experiment was designed as a pragmatic trial, with the intention of implementing the programme and trial in as realistic a setting as possible. I find statistically significant, but small and

[73] Note that this is independent of my finding larger intervention impacts to be associated with higher quality workshops: since I have been defining quality in terms of inputs (time available and class size), high quality workshops may have larger impacts but will also be more expensive to run. Ensuring that trained teachers teach a large number of pupils would not cost more money but may have a positive impact on intervention quality through facilitator experience.

short-lived reductions in depressive symptoms and absence from school, and a small impact on academic attainment which persists until two-year follow-up. I find no impact on anxiety symptoms or self-reported or teacher-reported behaviour. Programme dosage was high, and higher dosage was associated with larger intervention effects. Estimates of cost-effectiveness suggest that the intervention has similar results to many educational interventions in terms of its impact on academic attainment, while the costs would range from moderate to very large. The impacts on depression symptoms and absence are harder to evaluate because they do not last beyond postintervention. However, a policymaker should consider the impact on all three outcomes when deciding whether to fund the programme, and the potential for other benefits such as teacher wellbeing and earlier identification of pupils with more severe mental health problems. Given the small impacts on the outcomes measured, and the heterogeneity in impact by pupil characteristics, offering the programme universally may not offer the best value for money. However, advantages to a universal approach include the avoidance of stigma and its use as a screening mechanism to identify pupils who need more help.

Strengths of the study include the large sample size; the long follow-up period; the inclusion of a large number of schools from three different regions; excellent response rates and low attrition; and the use of a range of outcome measures from different sources. The trial was also a realistic scaled-up implementation of PRP, and was highly pragmatic along most dimensions making my results useful to policymakers. The most important weakness was the lack of randomisation to condition assignment. In addition, it is not clear that all of the outcome measures were sufficiently sensitive to change in a universal sample, or that they were able to tap into the core concept of resilience. The (temporary) lack of attainment data from national exams is also a limitation at the time of writing (August 2013).

Chapter 2: Figures and Tables

Table 2.1: Descriptive Statistics: Depression, Anxiety and Behaviour Scores as Outcomes

	Treatment group	Control group	p-value of test of equality of means	p-value when include school FE
Outcome variables at baseline				
Depression score at baseline	**8.66**	**8.17**	0.135	0.784
Standard deviation	6.98	6.36		
number of observations	1,805	1,888		
Anxiety score at baseline	**9.38**	**9.04**	0.252	0.590
Standard deviation	6.83	6.32		
number of observations	1,781	1,858		
Self-reported behaviour at baseline	**11.14**	**10.98**	0.609	0.367
Standard deviation	6.37	6.19		
number of observations	1,782	1,857		
Teacher-reported behaviour at baseline	**6.07**	**6.64**	0.225	0.154
Standard deviation	6.04	6.25		
number of observations	1,762	1,808		
Demographic variables at baseline				
Gender (% female)	**0.49**	**0.46**	0.218	0.077
Standard deviation	0.50	0.50		
number of observations	1,805	1,882		
Age in September 2007	**11.55**	**11.54**	0.341	0.608
Standard deviation	0.30	0.29		
number of observations	1,805	1,882		
Special Educational Needs	**0.25**	**0.29**	0.238	0.030
Standard deviation	0.44	0.45		
number of observations	1,805	1,882		
Free School Meals	**0.23**	**0.32**	0.001	0.483
Standard deviation	0.42	0.47		
number of observations	1,805	1,882		
KS2 English level	**4.07**	**3.90**	0.020	0.003
Standard deviation	0.78	0.82		
number of observations	1,749	1,802		
KS2 maths level	**4.05**	**3.90**	0.035	0.007
Standard deviation	0.80	0.81		
number of observations	1,760	1,803		

Notes: This table presents baseline means for the first four outcome variables and the demographic variables for this sample to compare treatment and control groups. The third column gives the p-value of the coefficient on ‘treated’ in a regression of the variable on treatment assignment, and the last column gives the p-value when the regressions are run with the addition of school fixed effects. Standard errors are clustered at the level of treatment assignment (class).

Table 2.2: Descriptive Statistics: Absence from School as Outcome

	Treatment group	Control group	p-value of test of equality of means	p-value when include school FE
Outcome variables at baseline				
Fraction of sessions absent (2006-7)	**0.055**	**0.058**	0.483	0.724
Standard deviation	0.06	0.07		
number of observations	1,227	1,221		
Demographic variables at baseline				
Gender (% female)	**0.49**	**0.45**	0.119	0.016
Standard deviation	0.50	0.50		
number of observations	1,227	1,221		
Age in September 2007	**11.54**	**11.55**	0.307	0.314
Standard deviation	0.29	0.29		
number of observations	1,227	1,219		
Special Educational Needs	**0.27**	**0.31**	0.262	0.400
Standard deviation	0.44	0.46		
number of observations	1,227	1,221		
Free School Meals	**0.23**	**0.38**	0.000	0.312
Standard deviation	0.42	0.48		
number of observations	1,227	1,221		
KS2 English level	**4.02**	**3.81**	0.032	0.037
Standard deviation	0.82	0.87		
number of observations	1,220	1,208		
KS2 maths level	**3.98**	**3.83**	0.098	0.064
Standard deviation	0.83	0.86		
number of observations	1,219	1,208		

Notes: This table presents baseline means for absence and demographic variables in order to compare treatment and control groups for this sample. The sample is different to that presented in Table 2.1 because of different availability of the outcome data. The third column gives the p-value of the coefficient on ‘treated’ in a regression of the variable on treatment assignment, and the last column gives the p-value when the regressions are run with the addition of school fixed effects. Standard errors are clustered at the level of treatment assignment (class).

Table 2.3: Descriptive Statistics: Academic Attainment as Outcome

	Treatment group	Control group	p-value of test of equality of means	p-value when include school FE
Outcome variables at baseline				
KS2 English level	**4.42**	**4.32**	0.135	0.094
Standard deviation	0.77	0.82		
number of observations	1,819	4,769		
KS2 maths level	**4.46**	**4.38**	0.217	0.086
Standard deviation	0.83	0.85		
number of observations	1,818	4,767		
KS2 science level	**4.72**	**4.65**	0.243	0.189
Standard deviation	0.69	0.74		
number of observations	1,817	4,763		
KS2 standardised combined score	**0.07**	**-0.03**	0.176	0.105
Standard deviation	0.88	0.94		
number of observations	1,819	4,773		
Demographic variables at baseline				
Gender (% female)	**0.49**	**0.45**	0.072	0.042
Standard deviation	0.50	0.50		
number of observations	1,819	4,768		
Age in September 2007	**11.55**	**12.19**	0.000	0.000
Standard deviation	0.29	0.62		
number of observations	1,819	4,756		
Age at baseline	**11.55**	**11.55**	0.741	0.497
Standard deviation	0.29	0.30		
number of observations	1,819	4,756		
Special Educational Needs	**0.26**	**0.26**	0.913	0.962
Standard deviation	0.44	0.44		
number of observations	1,819	4,756		
Free School Meals	**0.33**	**0.37**	0.187	0.544
Standard deviation	0.47	0.48		
number of observations	1,819	4,756		

Notes: This table presents baseline means for academic attainment and the demographic variables for this sample to compare treatment and control groups. The sample is different to that presented in Table 2.2 and Table 2.3 because of different availability of the outcome data. The third column gives the p-value of the coefficient on 'treated' in a regression of the variable on treatment assignment, and the last column gives the p-value when the regressions are run with the addition of school fixed effects. Standard errors are clustered at the level of treatment assignment (class).

Table 2.4: Programme impact on depression and anxiety symptom scores and absence from school

	Outcome: Depression symptoms score				Outcome: Anxiety symptoms score				Outcome: Absence from school			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treated*Postintervention	-0.099**	-0.113**	-0.176***	-0.150***	-0.002	-0.025	-0.063	-0.052	-0.100**	-0.101**	-0.137**	-0.136*
Standard error	(0.045)	(0.046)	(0.057)	(0.056)	(0.043)	(0.043)	(0.046)	(0.047)	(0.049)	(0.049)	(0.063)	(0.073)
Treated*1-year follow-up	-0.024	-0.028	-0.024	-0.006	0.013	-0.000	-0.011	0.018	-0.068	-0.069	-0.016	-0.018
Standard error	(0.045)	(0.047)	(0.058)	(0.061)	(0.040)	(0.042)	(0.041)	(0.050)	(0.063)	(0.063)	(0.078)	(0.090)
Treated*2-year follow-up	0.029	-0.005	0.041	0.042	0.045	0.003	-0.020	-0.008	-0.026	-0.028	-0.016	-0.026
Standard error	(0.055)	(0.057)	(0.068)	(0.073)	(0.046)	(0.047)	(0.053)	(0.060)	(0.077)	(0.077)	(0.085)	(0.097)
Treated	0.123**	0.157***	0.089*		0.079*	0.113***	0.057		-0.052	-0.010	0.031	
Standard error	(0.050)	(0.042)	(0.047)		(0.045)	(0.041)	(0.044)		(0.053)	(0.045)	(0.042)	
Number of Pupils	3,455	3,455	3,455	3,455	3,458	3,458	3,458	3,458	2,425	2,425	2,425	2,425
Sample size	12,898	12,898	12,898	12,898	12,833	12,833	12,833	12,833	9,633	9,633	9,633	9,633
Number of clusters	147	147	147	147	147	147	147	147	98	98	98	98
Month & day of the week	x	x	x	x	x	x	x	x				
Pupil characteristics		x	x			x	x			x	x	
School, School*Time			x	x			x	x			x	x
Pupil Fixed Effects				x				x				x

Notes: Each column represents a separate regression. Standard errors (in parentheses) are clustered at the level of treatment assignment (class). All specifications include controls for the time period (postintervention, 1-year or 2-year follow-up). Control variables for student characteristics are dummies for gender; SEN; FSM; Key Stage 2 maths and English attainment levels; and ethnicity (5 categories). The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1 based on the pooled (intervention and control group) distribution at baseline.

Table 2.5: Programme impact on behaviour scores and academic attainment

	Outcome: Self-reported behaviour score				Outcome: Teacher-reported behaviour score				Outcome: Academic attainment			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treated*Postintervention	0.042	0.025	0.025	0.027	0.082	0.058	-0.077	-0.007	0.104	0.102	0.088**	0.097**
Standard error	(0.044)	(0.044)	(0.048)	(0.052)	(0.071)	(0.061)	(0.072)	(0.068)	(0.072)	(0.072)	(0.036)	(0.041)
Treated*1-year follow-up	0.019	0.005	0.014	0.035	0.087	0.061	0.051	0.140*	0.091	0.084	0.075**	0.079*
Standard error	(0.040)	(0.041)	(0.050)	(0.057)	(0.067)	(0.061)	(0.087)	(0.081)	(0.056)	(0.055)	(0.035)	(0.040)
Treated*2-year follow-up	0.088*	0.043	0.059	0.043	0.132*	0.103	0.060	0.128	0.127*	0.124*	0.131***	0.131***
Standard error	(0.049)	(0.049)	(0.048)	(0.054)	(0.079)	(0.076)	(0.106)	(0.123)	(0.065)	(0.065)	(0.042)	(0.050)
Treated	0.033	0.079*	-0.000		-0.089	0.014	0.016		0.085	0.061	0.109**	
Standard error	(0.050)	(0.042)	(0.047)		(0.079)	(0.058)	(0.073)		(0.083)	(0.049)	(0.049)	
Number of Pupils	3,460	3,460	3,460	3,460	3,378	3,378	3,378	3,378	5,192	5,192	5,192	5,192
Sample size	12,843	12,843	12,843	12,843	12,393	12,393	12,393	12,393	20,270	20,270	20,270	20,270
Number of clusters	147	147	147	147	144	144	144	144	506	506	506	506
Month & day of the week	x	x	x	x	x	x	x	x				
Pupil characteristics		x	x			x	x			x	x	
School, School*Time			x	x			x	x			x	x
Pupil Fixed Effects				x				x				x

Notes: Each column represents a separate regression. Standard errors (in parentheses) are clustered at the level of treatment assignment (class). All specifications include controls for the time period (postintervention, 1-year or 2-year follow-up). Control variables for student characteristics are dummies for gender; SEN; FSM; Key Stage 2 maths and English attainment levels; and ethnicity (5 categories), except when academic attainment is the outcome when Key Stage 2 attainment is omitted. The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1.

Table 2.6: Programme impact by workshop quality	Depression	Anxiety	Absence	Self-reported behaviour	Teacher-reported behaviour	Academic attainment
Treated*Postintervention*High quality	-0.240***	-0.086	-0.172**	-0.020	-0.016	0.123***
Standard error	(0.072)	(0.061)	(0.078)	(0.063)	(0.070)	(0.041)
Treated*Postintervention*Low quality	-0.064	-0.027	-0.011	0.062	0.034	-0.028
Standard error	(0.054)	(0.050)	(0.094)	(0.054)	(0.115)	(0.086)
Treated*1-year follow-up*High quality	-0.109	-0.080	-0.033	-0.074	0.114	0.107**
Standard error	(0.076)	(0.065)	(0.101)	(0.072)	(0.080)	(0.043)
Treated*1-year follow-up*Low quality	0.090	0.109**	0.033	0.133**	0.253*	-0.055
Standard error	(0.068)	(0.052)	(0.136)	(0.054)	(0.133)	(0.056)
Treated*2-year follow-up*High quality	-0.044	-0.062	-0.035	-0.043	0.101	0.147***
Standard error	(0.084)	(0.074)	(0.107)	(0.064)	(0.120)	(0.055)
Treated*2-year follow-up*Low quality	0.124	0.042	0.005	0.119**	0.257	0.058
Standard error	(0.078)	(0.070)	(0.157)	(0.056)	(0.241)	(0.079)
p-value of test of equality high/low quality: Postintervention	0.007	0.324	0.102	0.174	0.650	0.079
p-value of test of equality high/low quality: 1-year follow-up	0.018	0.008	0.667	0.003	0.263	0.016
p-value of test of equality high/low quality: 2-year follow-up	0.039	0.205	0.812	0.009	0.496	0.315
Number of Pupils	3455	3458	2425	3460	3378	5192
Sample size	12898	12833	9633	12843	12393	20270
Number of clusters	147	147	98	147	144	506

Notes: Each column represents a separate regression. Specification used is the same as in columns 4, 8 and 12 of Table 2.4 and Table 2.5, but with the intervention dummy split into ‘high quality’ and ‘low quality’ interventions (see Table A2.8 and text for definition). The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1. P-values reported are from chi-square tests of the equality of the coefficients in each time period.

Table 2.7: Programme impact by use of skills

	Depression	Anxiety	Absence	Self-reported behaviour	Teacher-reported behaviour	Academic attainment
Treated*Postintervention*Uses skills	-0.204***	-0.026	-0.279***	0.010	-0.006	0.145***
Standard error	(0.072)	(0.065)	(0.086)	(0.065)	(0.083)	(0.049)
Treated*Postintervention*Not use	-0.121*	-0.013	-0.194**	0.019	0.008	0.140***
Standard error	(0.071)	(0.056)	(0.084)	(0.064)	(0.080)	(0.043)
Treated*1-year follow-up*Uses skills	-0.018	0.073	-0.200**	0.078	0.121	0.149***
Standard error	(0.075)	(0.068)	(0.100)	(0.067)	(0.101)	(0.051)
Treated*1-year follow-up*Not use	0.015	0.038	-0.079	0.019	0.153	0.103**
Standard error	(0.083)	(0.067)	(0.106)	(0.072)	(0.097)	(0.044)
Treated*2-year follow-up*Use skills	0.025	0.020	-0.179	0.053	0.056	0.183***
Standard error	(0.089)	(0.077)	(0.123)	(0.068)	(0.146)	(0.066)
Treated*2-year follow-up*Not use	0.056	0.024	-0.051	0.043	0.144	0.179***
Standard error	(0.104)	(0.078)	(0.106)	(0.076)	(0.134)	(0.059)
p-value of test of equality use skills/not: Postintervention	0.171	0.802	0.229	0.854	0.798	0.906
p-value of test of equality use skills/not: 1-year follow-up	0.587	0.563	0.184	0.311	0.623	0.260
p-value of test of equality use skills/not: 2-year follow-up	0.708	0.953	0.287	0.866	0.207	0.944
Number of Pupils	3,084	3,087	2,142	3,085	3,035	4,853
Sample size	11,562	11,509	8,516	11,509	11,144	18,962
Number of clusters	145	145	97	145	142	459

Notes: Each column represents a separate regression. Specification used is the same as in columns 4, 8 and 12 of Table 2.4 and Table 2.5, but with the intervention dummy split by pupils into those who reported using the skills taught by the intervention and those who reported not using them (see text for data sources). A substantial fraction of intervention pupils did not respond to this question, so the sample size is smaller than in previous tables. The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1.

Table 2.8: Heterogeneity in programme impact

	Depression	Anxiety	Absence	Self-reported behaviour	Teacher-reported behaviour	Academic attainment
Treated*Post* Baseline score > median	-0.119* (0.063)	-0.051 (0.054)	-0.083 (0.102)	0.005 (0.052)	0.060 (0.081)	0.125*** (0.046)
Treated*Post* Baseline score <= median	-0.012 (0.048)	-0.020 (0.040)	-0.051 (0.074)	0.007 (0.051)	0.031 (0.069)	0.049 (0.037)
p-value of test of equality	0.089	0.592	0.748	0.975	0.707	0.153
Treated*Post*Boy	-0.080 (0.059)	-0.043 (0.057)	-0.113 (0.088)	0.047 (0.059)	0.049 (0.082)	0.115** (0.046)
Treated*Post*Girl	-0.025 (0.069)	0.001 (0.056)	-0.009 (0.094)	0.012 (0.056)	0.101 (0.076)	0.087** (0.039)
p-value of test of equality	0.466	0.524	0.337	0.623	0.401	0.454
Treated*Post*FSM	-0.074 (0.094)	-0.104 (0.079)	-0.007 (0.131)	-0.015 (0.076)	0.060 (0.096)	0.052 (0.041)
Treated*Post*Not FSM	-0.038 (0.054)	0.018 (0.045)	-0.075 (0.068)	0.056 (0.052)	0.081 (0.076)	0.122*** (0.043)
p-value of test of equality	0.692	0.121	0.589	0.406	0.800	0.117
Treated*Post*SEN	-0.214** (0.089)	-0.062 (0.088)	-0.041 (0.132)	0.020 (0.082)	0.050 (0.090)	0.042 (0.050)
Treated*Post*Not SEN	-0.000 (0.053)	-0.012 (0.043)	-0.055 (0.073)	0.039 (0.050)	0.083 (0.077)	0.122*** (0.042)
p-value of test of equality	0.020	0.553	0.914	0.823	0.681	0.198
Number of Pupils	3,455	3,458	2,425	3,460	3,378	5,192
Sample size	12,898	12,833	9,633	12,843	12,393	20,270
Number of clusters	147	147	98	147	144	506
Month & day of the week	x	x		x	x	
Pupil characteristic*Post	x	x	x	x	x	x
School, School*Time	x	x	x	x	x	x
Pupil Fixed Effects	x	x	x	x	x	x

Notes: Each box within each column represents a separate regression. Specification used is the same as in columns 4, 8 and 12 of Table 2.4 and Table 2.5, but with the intervention dummy split by pupil characteristics at baseline. The first rows split the intervention group by the baseline value of the outcome. All regressions include an interaction of the variable along which the intervention group is split (e.g. FSM) with post. The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1.

Table 2.9: Cost effectiveness of UKRP

		Postintervention			1-year follow-up			2-year follow-up		
		Point estimate	Lower bound	Upper bound	Point estimate	Lower bound	Upper bound	Point estimate	Lower bound	Upper bound
Depression symptoms score	Effect size	-0.15	-0.26	-0.04	-			-		
	Equivalent to	-1 point on CDI			-			-		
	Percentile points	-5.4			-			-		
2007 costs (£)	Max cost/impact	1,230	710	4,586						
	Min cost/impact	519	300	1,935						
2009 costs (£)	Max cost/impact	698	403	2,603						
	Min cost/impact	287	166	1,071						
Absence from school	Effect size	-0.14	-0.28	0.01	-			-		
	Equivalent to	-1.3 days absence			-			-		
	Percentile points	-6.0			-			-		
2007 costs (£)	Max cost/impact	1,318	661	No						
	Min cost/impact	556	279	impact						
2009 costs (£)	Max cost/impact	748	375	-						
	Min cost/impact	308	154	-						
Combined academic attainment (English, maths and science)	Effect size	0.097	0.017	0.177	0.079	0.001	0.157	0.131	0.033	0.229
	Equivalent to	7% of an NC level			6% of an NC level			10% of an NC level		
	Percentile points	4.7			3.3			3.60		
2007 costs (£)	Max cost/impact	1,902	11,089	1,040	2,336	307,547	1,172	1,409	5,592	806
	Min cost/impact	803	4,679	439	986	129,769	495	594	2,359	340
2009 costs (£)	Max cost/impact	1,080	6,295	591	1,326	174,578	665	800	3,174	457
	Min cost/impact	444	2,589	243	545	71,815	274	329	1,306	188

Notes: Table shows estimates of per pupil cost effectiveness under four cost scenarios, partly dependent on teacher utilisation; see Table 1.16 for details. Upper and lower bounds represent 95% confidence intervals. The long-run costs would be represented by the 2009 costs. There were no intervention impacts on behaviour or anxiety symptom scores so these are excluded. In effect, the estimated cost is for a 1 SD improvement in that outcome at that point. NC level = National Curriculum level. I have not attempted to discount the upfront investment in teacher training and workshop delivery by the cost of capital: doing this would mean that facilitators who taught 10 workshop groups in the first year after training would be more cost-effective than teachers who taught the same number of groups over a longer time period (all else being equal) because of the time dimension of money.

Chapter 3: Impact of the UK Resilience Programme on pupil popularity

Introduction

Social relationships form an important part of a child's experience of school. Schools may aim to influence children's personalities and behaviour as well as develop their intellectual abilities, preparing them for social roles in the adult world (Parsons, 1959; Bowles, Gintis & Osborne, 2001). The social learning that takes place at school may be just as important as the development of cognitive and vocational skills to eventual academic attainment and success in the labour market (Heckman, Stixrud, & Urzua, 2006; Segal, forthcoming; Lindqvist & Vestman, 2011; Roberts, Harms, Caspi & Moffitt, 2007; Wolfe & Johnson, 1995). In particular, students' peers, and their relationships with their peers, may be critical to the functioning of schools.

Most of the economic literature on students' associations with others focuses on the influence of *peer groups*, which we could define as all other students in narrow or broad proximity to a student, whatever the student's feelings towards them. Examples of this include the impact of peers' characteristics on educational attainment (Lavy, & Schlosser, 2011; Lavy, Passerman, & Schlosser, 2012; Ammermueller & Pischke, 2009; Burgess & Umaña-Aponte, 2011; Zimmerman, 2003; De Giorgi, Pellizzari, & Redaelli, 2010); participation in recreational activities (Bramoullé, Djebbari & Fortin, 2009); charitable giving (Smith, Windmeijer & Wright, 2013); welfare participation (Bertrand, Luttmer & Mullainathan, 2000); and eating disorders (Costa-Font & Jofre-Bonet, 2013). Alternatively, we could choose to examine the influence of students' *friends* – other people (usually also peers) whom the student likes and chooses to associate with.[74] These two groups are likely to overlap substantially: being part of a peer group is usually a necessary but not sufficient condition for being a friend, as most students draw their friends from the immediate population of peers. However, there are important differences between the two concepts. Given preferences for social interaction with people of similar characteristics to oneself (homophily), we would expect greater similarity on own and friends' characteristics and therefore greater segregation with

[74] I will use these definitions of friends and friendships versus peers throughout: a friend is someone you like and would choose to associate with (or the reverse – someone who likes you and would choose to associate with you). A peer is anyone who occupies similar geographical or functional space and with whom you have to associate because of this proximity.

friendship groups than with peer groups, particularly when peer groups are large (Weinberg, 2007; Currarini, Jackson & Pin, 2009). There is therefore no guarantee that a pupil's friends will have similar aggregate characteristics to their peers. Moreover, any student in a school with a roll of more than one will necessarily have a peer group, but they will not necessarily have any friends. Thus the literature on the influence of peer groups can assess the impact of having peers with particular characteristics (such as having high attaining versus low attaining peers), but does not examine the impact of having any peers at all.[75] By contrast, we can ask what the effect of having few friends is. For both groups (friends and peers) we can ask about the quality or intensity of the relationship. Given that friends will also tend to have closer relationships and more contact than will peers, we might also expect the impact of friends to be greater. Although a pupil's entire peer group may be influential, their friends may be particularly so.

There is good reason to think that having friends might be important in itself, quite aside from the characteristics of those friends. Friendship has a utility value which most people would find self-evident, and the desire for relationships with others appears to be a fundamental human motivation (Baumeister & Leary, 1995). Investigating the characteristics of very happy people, Diener and Seligman (2002) find that while no characteristic is sufficient for happiness, good social relations are necessary, and very happy people are highly social.

Friendships form an important part of the context of adolescent development (Rubin, Bukowski, & Parker, 2006). Friends also offer companionship, affection, and practical and emotional support; promote self-esteem and confidence; and are important to both psychological and biological aspects of well-being (Berndt, 2004; Reis & Collins, 2004). Stable friendships during adolescence can provide protection against stressors and promote good adjustment and adaptation (Berndt, Hawkins & Jiao, 1999; Juvonen, 2006; Hartup, 1996; Buhrmester, 1990). Having friends within school promotes higher educational attainment and reduced risk of dropping out Calvó-Armengol, Patacchini & Zenou, 2009; Coleman, 1988; Roseth, Johnson & Johnson, 2008; Babcock, 2008; Ladd, 1990; Vaquera & Kao, 2008; Finn, 1989; Lavy & Sand, 2012; Woodward & Fergusson, 2000). The absence of friends can be particularly

[75] An exception might be the literature on class size: this examines the impact of having more or fewer peers. Yet even here the observed impact is often thought to operate through the characteristics or behaviour of peers rather than the mere *quantity* of peers e.g. Cortes, Moussa and Weinstein (2012).

damaging: low peer acceptance or peer rejection leads to behavioural problems, emotional distress, and other disorders (Witvliet, van Lier, Cuijpers & Koot, 2009; Ladd, 2006; Klima & Repetti, 2008; Coie, 2004; Buhs & Ladd, 2001; Parker & Asher, 1987; Coie, Lochman, Terry & Hyman, 1992; Bierman & Wargo, 1995; Dodge, 1993; Wentzel, Barry & Caldwell, 2004; Hymel, Rubin, Rowden & LeMare, 1990). Children with few or no friends are also at greater risk of being bullied (Salmivalli, Huttunen & Lagerspetz, 1997). Social isolation may also be a risk factor for poor health behaviours (Ennett & Bauman, 2000; Botvin, Baker, Dusenbury, Botin & Diaz, 1995).

Friendships may also produce economic gains: Conti, Galeotti, Mueller and Pudney (2012) find that moving from the 20th to 80th percentile of high-school popularity distribution yields a 10% wage premium nearly 40 years later. Bandiera, Barankay and Rasul (2009) find in a field experiment that social connections improve the performance of connected workers. Even relatively weak friendship ties may promote social cohesion and bring benefits such as improved access to information (Granovetter, 1973). Having friends might be merely *indicative* of social capital in some of these instances, but having friends is also essential to the development of social skills (Youniss & Haynie, 1992).[76] Youth friendships will therefore predict adult outcomes because they are both a measure of the stock of social skills as a teenager, and because they provide a training ground for social competence of use in later life.

A number of interventions and strategies have been developed to develop positive peer relations amongst children and adolescents at school. The more diffuse strategies focus on the aspects of school organisation which are thought to promote connectedness and belonging, such as the use of form groups or house systems. These reduce the number of different students and teachers a student will have contact with (the effective number of peers) whilst increasing the time they spend with the same groups of pupils, with the aim of encouraging connectedness and belonging, particularly for students at risk of dropping out (Wetz, 2009). Sports, arts and other extracurricular clubs are also used to promote good peer relationships, and there is evidence that these can be successful (Little, Wimer & Weiss, 2008).

There are also curriculum-based interventions which aim to teach social skills explicitly. The most widely used of these have been developed for targeted populations

---

[76] Another example of this is the positive impact that high school leadership positions and sports participation have on adult wages: these are both indicators of social skills but participation provides the essential environment for development of these skills (Kuhn & Weinberger, 2005; Postlewaite & Silverman, 2005).

with serious social deficits or which suffer from peer rejection.[77] For example, the UCLA PEERS intervention (Program for the Education and Enrichment of Relational Skills) is designed to promote the social skills of adolescents with autistic spectrum disorder (Laugeson, Frankel, Mogil & Dillon, 2009; Laugeson, Frankel, Gantman, Dillon & Mogil, 2012). Similarly, the PATHS intervention (Promoting Alternative Thinking Strategies) was developed for use with at-risk populations, though has now also been used with universal samples (Greenberg, Kusche & Cook, 1995). Yet there are benefits of popularity and social skills across the distribution, and there have also been a number of trials of universal interventions in different countries. These target a range of aspects of peer relations including peer rejection and isolation, bullying, self-esteem and prosocial behaviour (Waasdorp, Bradshaw & Leaf, 2012; Barret, Webster, & Wallis, 1999; Witvliet, van Lier, Cuijpers & Koot, 2009; Dobeš, Fedáková, Lehotská and Koscelníková, 2010). However, very few of these programmes have measured friendship or peer acceptance through sociometric measures, with most relying on assessments from children or teachers of classroom climate or others' behaviour.[78] This is important, because measures of peer interaction based on self- or teacher-reports may not accurately reflect pupils' friendship networks, while peer-based nominations of friendship (the in-degree of friendship) may be considered the gold standard in measuring friendship connections (Henry & Metropolitan Area Child Study Research Group, 2006; Parkhurst & Hopmeyer, 1998; Gest, 2006; Hartup, 1996; Bondonio, 1998).

I evaluate the impact of the UK Resilience Programme (UKRP) in a universal population of 6,500 11-12 year old pupils at 22 non-selective English secondary schools. The programme is based on cognitive behavioural therapy, and includes substantial components on social problem solving and prosocial skills, as well as emotional awareness and coping skills. The evaluation was designed as a pragmatic controlled trial, with intervention and control group pupils in every school. Pupils were asked to list all the people they were good friends with at three times postintervention, and using this I can measure pupils' popularity using the number of friendship nominations they receive from their peers. I test whether UKRP had an impact on pupils' popularity measured in

[77] Bierman (2003) summarises the research on chronic peer rejection and interventions and strategies to overcome it.
[78] The exceptions are Dobeš et al. (2010), who measure impact through friend nominations; and Witvliet et al. (2009) and Conduct Problems Prevention Research Group (1999; 2002), in both of which children were asked to nominate an unlimited number of friends who they liked most and least. All three studies found positive impacts at postintervention on these measures. However, all three trials used only a moderate-sized sample of children (fewer than 1,000), so cannot be viewed as cases of fully scaled-up programme implementation.

this way, finding a small increase in popularity in the short run (immediately postintervention) for intervention pupils, which is zero at one-year and two-year postintervention. This short-run effect is equivalent to one half of an extra friend nomination per intervention group pupil relative to the control group, or a shift of 3 percentiles in popularity, and it appears that this is driven by being listed more often as a more distant friend (friend number 11 or below), rather than any change in nominations as a close friend. Due to the absence of a placebo control group I cannot distinguish between mechanisms such as greater contact with other pupils during the programme, and improvements in prosocial skills, although the two are not mutually exclusive. The result does not appear to be generated by a greater tendency to list more friends in the intervention group, suggesting that it is genuinely an improvement in popularity and not merely in reporting. This evaluation provides evidence for the impact of a scaled-up wellbeing and prosocial skills programme in a real world setting, taught by regular school staff. My results suggest that social capital such as social relationships and popularity with peers can be lost as well as gained, and that continued efforts may be needed to maintain it.

The paper addresses the following questions:

1. What impact did the UK Resilience Programme have on pupils' popularity with their peers?
2. Is there heterogeneity in intervention impact by workshop organisation or pupil characteristics?
3. Is the intervention impact on popularity mediated through an impact on depressive symptoms?

I described the UKRP intervention and evaluation in Chapter 1, including the recruitment of LAs and schools; population characteristics; attrition; and the data available. Here I outline the identification strategy, and describe the statistical approach and outcome data, before presenting the results. I include some information on cost effectiveness. This is followed by the discussion section and conclusions.

Method

Intervention and control groups

As described in Chapter 1, classes of pupils in the 2007 cohort in participating schools were arbitrarily assigned to the treatment group based on availability of rooms and trained staff. It was hypothesised that this assignment would be arbitrary and would therefore mimic random assignment, allowing an accurate identification of the treatment effect through comparing control and treatment group outcomes. The 2006 (year above) cohort was also available in 13 schools as an additional control group. This was because these schools had intended to include the whole of their 2007 cohort in workshops, so needed to use another yeargroup as a control group. In the event only 7 schools included the entire 2007 cohort in workshops, but the 2006 cohorts continued to be surveyed in all 13 schools. For the 2006 cohort to be a suitable control group, we must assume that adjacent cohorts in the same school are formed quasi-exogenously, and so could be viewed as the counterfactuals of one another when one cohort receives the intervention and another does not. As discussed in Chapter 1 ('Intervention assignment'), since cohort membership is almost always determined by date of birth and rarely modified subsequently this seems like a reasonable assumption to make. However, it is important to compare pupils when they are the same age and at the same stage in their school careers, as these are factors which could have a sizeable impact on their psychological health and their friendship groups. Year-above (2006 cohort) students surveyed at the end of Year 8 in 2008 are therefore compared with 2007 cohort pupils surveyed at the end of Year 8 in 2009, for example. The 2007 cohort is split roughly equally into control and intervention students, while all students in the 2006 cohort are in the control group. The intervention and control groups do appear to be well-matched on their demographic characteristics at baseline (see below). However, I also include controls for prior academic attainment and demographic characteristics in regressions to account for any remaining differences.

Data

Data from students was collected through paper questionnaire booklets containing psychological inventories measuring anxiety, depression and behaviour, plus some demographic questions, as described in Chapter 1. From June 2008 the questionnaires included a question on pupils' friends, and this is the source of the primary outcome data for this paper.

Surveying points

The 2006 cohort was surveyed at the end of Year 7 (June 2007); end of Year 8 (June 2008); and end of Year 9 (June 2009).[79] The 2007 cohort were surveyed at the start of Year 7 (September 2007); end of Year 7 (June 2008); end of Year 8 (June 2009); and end of Year 9 (June 2010).[80] All students in the relevant cohort at participating schools were surveyed at each point. However, the question about pupils' friends (the source of the data for this paper) was only included in questionnaires from June 2008. At this point the 2007 cohort was at the end of Year 7, while the 2006 cohort was at the end of Year 8. This means that I have two years of friendship data for the 2006 cohort, and three years for the 2007 cohort. Table 3.1 presents the number of observations on friendship nominations available at each point, by treatment status, year group and cohort. Here we can see that we have a smaller control group for the data for Year 7. This is because Year 7 data is only available for the control pupils who were in the 2007 cohort, while in years 8 and 9 we also have data for control group pupils from the 2006 cohort. This is because I have to compare pupils when they are at the same stage in school and the same age, and the 2006 cohort were not asked about their friends until they were at the end of Year 8. About half of the 2007 cohort is treated, with half in the control group, while the entire 2006 cohort is in the control group.[81]

Outcomes

Data on pupils' friendships was obtained from one page of the evaluation questionnaire, which asked pupils to list who their good friends were, along with each friend's form group and school to aid us in matching their names to codes. An example of this page is given in Table A3.1. There were 24 blank lines to fill in names, but these were not numbered. Some pupils opted to add extra names along the margins or on an additional piece of paper, and these were recorded by the research team after the names within the 24 lines. A team of research assistants matched these names to codes using class lists from participating schools, and school lists for all other schools in each LA (obtained from the NPD). These codes can be used to match pupils into the NPD, and

[79] This cohort started school before the evaluation had begun and so it was not possible to obtain a start of Year 7 baseline for them. The first data collection in July 2007 was ordered by DCSF only two weeks before it had to happen, and so the organisation of it (including the decision on which questions to include) was rushed.

[80] There was also some additional surveying for some members of the 2007 cohort in 2007-08, depending on workshop timing, but none of this included questions on pupils' friends.

[81] Information on the number of pupils who actually answered the friends question at each point is given in
Table **A3.2**.

thereby obtain access to demographic and attainment data for all matched pupils listed as friends, not just those in the evaluation sample at UKRP schools. The codes are also used to match listed friends to evaluation data and generate the outcome measures.

Number of times a pupil is listed as a friend: in-degree friends

The primary outcome is the number of times a pupil is nominated as a friend by others in the evaluation sample, which I will call the in-degree of friendship (following Conti et al., 2012, and others). This does not necessarily gauge the quality of the friendships a pupil has, but it can be taken as a measure of popularity, and the number of friends a pupil has appears to be positively related to the emotional support they receive (Cauce, 1986). I use this as the main outcome measure because peer-nominated friendship (the in-degree) is likely to be a more reliable measure of popularity than self-reports (the out-degree; Bondonio, 1998; Henry & Metropolitan Area Child Study Research Group, 2006; Hartup, 1996). The in-degree is also made up of information from multiple individuals, rather than relying on a single source. It is also similar to measures used in other papers (e.g. Conti et al., 2012; Burgess & Umaña-Aponte, 2011; Conduct Problems Prevention Research Group, 1999), although my questionnaire allowed for more friendship nominations than most.

One advantage of this measure is that a pupil does not have to complete a questionnaire in order to have a score: the dataset includes all pupils who were on their schools' registers at the time of the survey, regardless of whether they themselves returned a questionnaire. A pupil who is on the register but who receives no friendship nominations from anyone else in the evaluation sample will receive an in-degree score of zero. Thus a pupil's absence or refusal to complete questionnaires will bias downwards their friends' in-degree scores, but not their own. This also means that I have very little missing data on pupils, making attrition even less of a problem than for the other questionnaire outcomes for which attrition is reported in Chapter 1 ('Attrition').[82] I also run robustness checks excluding observations where this variable is equal to zero to check that the results are not solely due to excess zeroes in the data.

There are certain potential biases to this measure. In the example given in Table A3.1, we can see that the handwriting is very poor, which, combined with the pupil only listing their friends' forenames, means that most of them will not be positively

[82] Of course, school rolls may not be entirely up-to-date, so some pupils who have left may still appear, and some pupils who have recently joined the school may not yet be on roll. It is not clear how this would bias my estimates of programme impact for these pupils, beyond introducing classical measurement error.

identified. This problem – misspelling, ambiguity due to use of forenames with no further information given, and poor handwriting – means that some listed friends will not be matched even if we know that they are in the same year and at the same school. If the six friends listed in Table A3.1 are in the evaluation sample then this will affect their in-degree scores: we cannot credit them with this respondent's nominations as we cannot identify who they are. The in-degree score may therefore be biased downwards slightly, particularly for pupils whose friends have poor literacy.[83] We would expect the same for pupils whose friends have high absence rates (and who are therefore less likely to be present to fill in a questionnaire), or who are less inclined to answer questionnaires for whatever reason. Pupils whose forenames which are very common within their school are also more likely to have their in-degree score biased downwards, as any nomination which does not include their surname or form group will be ambiguous. However, the majority of friend lists were clearly written and largely unambiguous, and the matching rate for friend names is very high: only 6% of all the friends listed were not properly matched to an NPD code (i.e. not positively identified as a particular pupil), and most of these were out of school. Only 1.2% of all friends listed were within the same school but not matched due to ambiguity, and these are the individuals whose in-degree scores will be biased downwards.

Pupils with the majority of their close friends outside of the evaluation sample (in different cohorts or schools, for instance) will have low scores, even if they do in fact have many friends. We might expect that pupils who are popular outside of school will also be popular within their school and cohort, so this would only understate the popularity of pupils who *only* have friends outside of school. For example, pupils who have only recently joined the school may not get any friend listings.[84] However, 84% of friends listed were at the same school as the respondent, so these pupils will be a small minority. Moreover, as long as these possible reporting biases affect the intervention and control groups equally then this should still be a valid measure of programme impact. I also present robustness checks which exclude these pupils.

---

[83] Poor literacy could affect responses in a number of ways: pupils may be less likely to get to the friends section of the questionnaire (page 12) in the time available; less likely to list many friends because of the effort of writing; and less likely to be able to give enough information to positively identify the friend listed, e.g. surname or form group as well as forename.

[84] The geographical proximity of many of the UKRP schools means that many pupils leaving one UKRP join another and pupils are therefore still likely to have friendship listings from their old schools. Thus only pupils joining a school from outside the area or from other non-UKRP schools are likely to receive no listings at all.

Number of friends a pupil lists: out-degree friends

I also look at the number of friends each respondent lists themselves – the out-degree of friendship. This could itself be an outcome, but could also mediate any impact on the number of friend listings a pupil receives. This variable could be seen as indicating a more pro-social attitude or more social confidence. If the UKRP has an impact on these, then we might expect to see a difference in this variable between the intervention and control groups. If this is the case, and intervention pupils are more likely to be in regular contact with other intervention pupils (because they are more likely to be in the same classes), then the intervention could have an impact on the number of listings intervention pupils receive simply because other intervention pupils tend to list more friends. It is also important to describe this variable to show reporting patterns through time. As for the main outcome variable, all pupils who were on their schools' registers at the time of the survey will have a value for this variable. If a pupil did not fill in a questionnaire, or did not fill in the friends section, this will be equal to zero. Because this may generate excess zeroes, I also run robustness checks excluding zeroes and using zero-inflated negative binomial regressions. Note that this measure will not be affected by matching failures: we can see that the respondent in Table A3.1 listed 6 friends, even though we do not know who they are.

Number of times a pupil is listed as a friend by rank

I further examine the main outcome variable – in-degree friends – by splitting it into two: the number of times a pupil is listed as a top ten friend (in the first ten spaces in the questionnaire); and the number of times a pupil is listed as friend number 11 or below. The spaces for friends' names in the questionnaire were not numbered, and pupils were not asked to list friends in order of closeness or preference, but in practice the top-listed friends are more likely to reciprocate listings; are more likely to list each other as part of a close group; and are more likely to still be on the list a year later. This suggests that pupils do list friends roughly in order of preference, so that a friend in the top ten is a close and regular friend, while lower-listed friends are more distant but pleasant acquaintances. Given this, it is interesting to see where any programme effect appears: it may be more difficult to influence a pupil's top ten in-degree score, than to make pupils slightly more attractive people and therefore gain them more low-ranking listings. The same caveats apply to this measure as do to the total number of friend listings a pupil receives. In addition, I am treating rank as meaning the same thing for all pupils: being listed as friend number 8 by a pupil who lists only 8 friends is weighted the same as being listed number 8 by someone who lists 50 friends. As for the other outcome variables I run robustness checks excluding zeroes.

Statistical analysis

I compare intervention and control group outcomes at postintervention (end of Year 7), one-year follow-up (end of Year 8) and two-year follow-up (end of Year 9). I do not have information on pupils' friends at baseline (pre-workshops), so I cannot use difference-in-differences to estimate programme impact as I did in Chapter 2 (when evaluating the intervention impact on psychological symptom scores, absence, and attainment). Since I do not require that pupils have baseline data, all students are included in the analyses if they have any data in any period. As described above, it is not necessarily for pupils to have completed a questionnaire to have data available: they can still be assigned both an in-degree and an out-degree score. This means that any pupil who was on the school roll and in the relevant cohorts at a survey point will be included in the analysis at that point in time.[85]

I used negative binomial regressions in Stata (StataCorp, 2011) to estimate the impact of the intervention on each outcome variable separately. The outcome data are all count data, but Poisson regression is inappropriate because the variance is substantially greater than the mean for all four outcomes. Thus negative binomial regressions are the most appropriate method for count data with less restrictive assumptions (Greene, 2003). I have also run the same specifications with Ordinary Least Squares as a robustness check. I include three specifications for each outcome: the basic specification including all observations with any outcome data; the same specification but with the sample restricted to observations for which I have all demographic data; and finally a specification which includes demographic variables. This allows me to compare results with the full sample and the sample with demographic characteristics available, and to directly compare the results in the same sample with and without demographic control variables. This acts as a robustness check of the sensitivity of the results to using different specifications and sample, which is particularly important given the non-random condition assignment and lack of a baseline for this measure.

[85] I also run robustness checks using only pupils with a non-zero outcome score; when I do this the sample attrition will look almost identical to the attrition statistics shown for the depression score in Chapter 1 (Table 1.10).

Thus for each outcome measure (described above) I present the following analyses:

$$y_{it} \sim Negbin(\mu_{it}, \kappa)$$

(1) All observations:

$$\log \mu_{it} = \beta_1 TREAT_i * \gamma_t + \gamma_t + \varepsilon_{it}$$

(2) Same specification as (1), but sample restricted to observations with demographic variables available:

$$\log \mu_{it} = \beta_1 TREAT_i * \gamma_t + \gamma_t + \pi_{it}$$

(3) Adding in demographic control variables, with same sample of observations as in (2):

$$\log \mu_{it} = \beta_1 TREAT_i * \gamma_t + \gamma_t + \beta_2 \sum X_i + s_i + \nu_{it}$$

Where for each student i in school s at time (yeargroup) t, y_{it} is the outcome of interest; $TREAT_i$ is a treatment assignment dummy; γ_t is a set of three time dummies representing the yeargroup a pupil is in at the time (or, equivalently, their age); X_i is a vector of student characteristics; s_i is a school fixed effect; and ε_{it} , π_{it} and ν_{it} are unobserved individual random effects. In each case β_1 , the coefficient on treated*yeargroup, estimates the impact of the intervention. Note that all the yeargroup dummies correspond to time periods which are postintervention or follow-up periods; there is no baseline available. When controlling for the school a student attends I use the school they attended at baseline, as subsequent school moves could be endogenously determined.

Thus the basic specification uses equation (1) with all data, and includes three dummies for yeargroup (representing the end of Year 7; Year 8; and Year 9), and three interactions of treatment*yeargroup on the right hand side. The second regression (2) uses the same specification as (1), but restricts the sample to those observations for which I have full demographic data, for comparison with the fully controlled specification in (3). The final specification (3) includes dummies for gender; SEN; FSM eligibility; broad ethnic group (5 dummies); month of birth (giving relative age within the cohort – 12 dummies); and school attended at the start of Year 7 (22 dummies). In addition, the mean Key Stage 2 test score (mean of English, maths and science scores in national tests at age 11) is included as a continuous variable. These characteristics were included because of their availability and association with the outcome variables (see Chapter 1, 'Pupil characteristics'). Since assignment into the intervention or control group was by class, I cluster the standard errors of all regressions by class grouping at baseline, which should take into account correlation within the class at a point in time and across different time periods (serial correlation).

Equation (3) represents the preferred specification, and I also split the sample by student and workshop characteristics to examine heterogeneity in impact using this specification. For example, to test whether there is a differential impact of the intervention on girls and boys I will run two separate regressions with the specification outlined in (3), one for girls only and one for boys only.

I also present a simple test of mediation, following the procedure in Baron and Kenny (1986). This is to explore whether any programme impact on one variable is related to a programme impact on another, or whether the effects operate independently. I test whether the intervention impact on friendship was mediated through the depressive symptoms score at each point. Using the fully-controlled specification in equation (3) above, I present the following steps:

Step 1: Regressing the dependent variable (friendship nominations) on the independent variable (intervention assignment). This is to check whether intervention assign is a significant predictor of the dependent variable.

Step 2: Regressing the mediating variable (depressive symptom scores) on intervention assignment (the independent variable). If there is no significant association, then the mediator cannot be mediating the impact on the dependent variable.

Step 3: Regressing the dependent variable on both the independent variable and the mediator (this will involve adding the depression symptoms score to equation (3) as an additional predictor). If there is still a significant relationship between the mediator and the dependent variable, and if the strength of the relationship between the independent variable and the dependent variable is reduced relative to the relationship seen in Step 1, this suggests that the mediator mediates the impact of intervention assignment on the dependent variable. This cannot establish the direction of causation, but it can provide evidence as to whether a change in one variable is associated with a change in another.

Results

Intervention and control groups

Table 3.2 displays descriptive statistics of the demographic characteristics of the sample at baseline. I cannot show that the intervention and control groups had similar scores on the outcomes variables at baseline because I do not have a baseline measure for the outcome variables. However, because most of the demographic and attainment data is obtained from the National Pupil Database, it is possible to obtain information for almost all pupils at the experimental baseline i.e. when they joined Year 7 at participating schools.[86] The control group is significantly older than the treatment group. This is because roughly half the pupils who started secondary school in September 2007 were in the control group, while all of the pupils who started in September 2006 (so were one year older) were surveyed as an additional control group However, I compare pupils' results when they are at the same age, so that the outcomes of pupils in the 2006 cohort when at the end of Year 8 in 2008 are compared with the outcomes of the 2007 cohort when they are at the end of Year 8 in 2009 (see Table 3.1). Moreover, the treatment and control groups are no more likely to be young or old within the year – a comparison of the month part of their age in September each year reveals no significant difference between the two groups. The only other difference between the groups is that the control group contains a slightly higher fraction of black pupils (p=0.09). Although I cannot show that the intervention and control groups looked similar on the outcome variable at baseline, these results suggest that they are at least very similar in terms of demographic characteristics. This suggests that the control group does provide an appropriate comparison for the intervention group.

Intervention impact on friendship measures

Figure 3.1 presents the distribution of the number of friends respondents listed (out-degree friends). The distribution has a trimodal shape, with peaks at zero, 10 and

---

[86] The exceptions are the variables 'lives with mother' and 'lives with father', which are measured 1-2 years after baseline. Two time periods are provided for absence data: the first term of Year 7, and when pupils were in Year 6 at primary school. The latter data was not available for pupils in the 2006 cohort, for the reasons outlined in Chapter 1.

24. The mode response was zero, with 15.2% of 16,317 observations listing no friends. This could be for several reasons: because the pupil did not complete a questionnaire due to absence, having left the school, or through choice, or because they completed most of the questionnaire but left the friends section blank (or did not get to it in time). There were 24 lines on the questionnaire to list friends (although these were not numbered), so the next most frequent number is 24, with 14.6% of observations listing this number of friends. However, pupils could list more names on additional pieces of paper or fit extra names in the margins, and 4% of observations have more than 24 friends listed, with the maximum number being 71. The panels in Figure 3.1 show that treatment and control group pupils had similar patterns of responses, but with the control group listing slightly fewer friends at each point. Only in Year 7 was the control group distribution substantially different from the treatment group distribution.

Figure 3.2 presents the distributions of the number of friend listings pupils receive (the in-degree). This looks like a lognormal distribution, with a long right tail, although note that it is important to use negative binomial regression rather than a Poisson model because the variance is much greater than the mean, and using Poisson regressions would underestimate the standard errors (Greene, 2003). The overall shape closely resembles that reported for in-degree nominations by Strauss and Pollack (2003), even though the Add Health dataset they use does not allow respondents to list more than 10 friends. The median in-degree score is 10, and the maximum 47, with the mode at 8. Here we can see that the treatment and control group distributions are very similar, but that in Year 7 the treatment group distribution appears to have been shifted to the right, suggesting that pupils in the treated group are listed as a friend by more people.

These differences are shown graphically in Figure 3.3, which gives the means and 95% confidence intervals for the in-degree score. Here we can see that there is a significant difference in the number of nominations between the treatment and control groups taking all years together, and that this is driven by a large difference in Year 7, with the gap disappearing by Year 9. Figure 3.4 shows the same thing but for the out-degree score, showing a smaller gap between treatment and control which fades through time and is at most points not statistically significant.

Figure 3.5 and Figure 3.6 show that the overall difference in the in-degree score is driven by differences in both the top ten in-degree score and the 11 or lower in-degree score. Table 3.3 presents the raw means for the outcome variables by time period (year group). Here we can see that treated pupils are listed as a friend by more people than control pupils in each time period, and that this difference is statistically significant in Year 7 and in Year 8. This difference is due to a higher rate of listings in both top ten friends and in friend number 11 or below (except in Year 9). It may also be partly driven by an increased propensity to list more friends themselves – although the difference here only approaches statistical significance in Year 7.[87]

Table 3.4 puts this data into negative binomial regressions. Column (1) presents the raw specification with the number of listings a pupil receives on the right hand side and only the treatment dummy on the left hand side. Here we can see that in Year 7 treated pupils receive 13% more listings than control group pupils (IRR=1.132), and in Year 8 they receive 6% more listings (IRR=1.057), with the difference becoming insignificant in Year 9. Demographic and attainment data is not available for the full sample; including this reduces the sample by 780 observations and 431 pupils. Column (2) reports the same raw specification (no controls) as column (1), but for this reduced sample, for the purposes of comparison. This reduces the impact slightly, particularly in Year 8 where the point estimate on treatment is now insignificant. Column (3) uses this reduced sample and adds in controls for gender; special educational needs (SEN); free school meals eligibility (FSM); mean score in Key Stage 2 exams; broad ethnic background; month of birth; and school attended at baseline. This leaves only the coefficient on treated in Year 7 marginally significant – pupils in the treatment group receive 5% more friend listings in Year 7, with the impact fading in subsequent years. Including the control variables and reducing the size of the sample accordingly halves the size of the coefficient on the treated variable in years 7 and 8, and reduces their significance. This suggests that there are significant differences between the treatment and control groups and that demographic characteristics are important predictors of the number of friend listings a pupil receives.[88]

The remaining columns in Table 3.4 present the same specifications for different outcome variables. The increase in the number of listings a treatment group pupil receives does not appear to be due to treated pupils listing more friends – the coefficients on 'treatment' when the outcome is the number of friends listed are positive but small and insignificant. This is important because it suggests that the risk of spillovers from the treatment group to the control group is lower: if treated pupils listed

[87] Treatment assignment was by class, so if treated pupils are more likely to list more friends, and friends usually share at least some classes, this would result in an increase in listings for treated pupils.

[88] In particular, academic attainment is a strong predictor of the number of friend listings a pupil receives.

more control group pupils in the same school as friends, this would bias downwards the impact of the programme on the number of times a pupil was listed as a friend. Moreover, to the extent that treated pupils are more likely to share a class with other treated pupils it could indicate a reporting effect rather than an increase in the number of pupils considered to be friends. The number of times a person is listed as a friend can be split into two parts: listings as a close friend (friends 1-10), or as a more distant friend (friend 11 or below). Here it appears that treated pupils receive more nominations in both of these categories, but that once we control for school attended at baseline and demographic characteristics only the likelihood of being listed as a more distant friend is increased, and this appears to be driving the effect on the total count of friend listings. This suggests that treated pupils do not have any more close friend listings than control group pupils, but are more likely to be listed as an additional friend. The magnitude of these effects on the total number of times listed as a friend and as friend 11 or below is equivalent to half an extra listing per pupil in the treatment control relative to the control group (0.54 of a friend in the total listing; 0.46 of a friend 11 or below). Thus it would appear that treated pupils received 0.5 more friend listings in Year 7 than control group pupils, and that this was primarily but not entirely due to an increase in the number of listings they received as friend number 11 or lower.

Heterogeneity

Table 3.5 uses the same specification as column (3) of Table 3.4, but splits the sample by demographic characteristics, with each column representing a separate regression. Here we see a much stronger impact on girls' friendship listings, with boys showing no impact of treatment. SEN pupils have no differential impact to pupils without SEN, while pupils entitled to FSM appear to benefit much less from the programme than those without FSM – the coefficient on 'treated' in Year 9 even becomes negative and marginally significant for FSM pupils. Pupils who are older within their cohort (born September-February) appear to be benefit more in the short run than younger pupils (born March-August). However, pupils with a mean Key Stage 2 score below the national target of level 4 (about 21% of pupils in this sample with available Key Stage 2 data), have a coefficient on treated*Year 7 which is substantially higher than pupils who scored at or above the national targets. Prior academic attainment is a strong predictor of the number of friend listings a pupil receives, so it may be that there is more room for an impact with very low attaining pupils.

As discussed in Chapter 1 ('Programme adherence'), there was some variation in the time allocated to UKRP workshops and the number of pupils in UKRP groups, despite guidance from the development team that at least 18 hours should be available and groups should contain no more than 15 pupils. In practice, most treated pupils (60%) were not taught in workshops meeting these criteria. I take hours and class sizes as measures of workshop quality, both because of the direct effect that these may have had on intervention quality, and because few hours and large classes may indicate disorganised and relatively uncommitted schools which made less effort to teach the programme well. Table A3.3 shows the number of treated observations and pupils and a summary of class sizes and hours available for this sample.[89] Designating all workshop groups as 'poor quality' which had either over 16 pupils and/or fewer than 17 hours available I find that 28% of pupils were in poor quality workshops, with 28% of treated observations coming from these.

Table 3.6 presents the main results by workshop quality.[90] Each column represents a separate regression, using the specification with all control variables as in columns 3, 6, 9, and 12 of Table 3.4, but with interactions of treatment*year group with workshop quality. Here we see that the coefficient for treated in Year 7 is larger for the 'high quality' (high dosage) workshop group than for the low quality (low dosage) group for all outcomes, although the p-value of a test of equality of the coefficients is only significant when the outcome is the number of times a pupil is listed as friend 11 or below. The picture is different in Year 8, with the coefficients about the same size and none significantly different. In Year 9 the coefficients again look similar, but treated pupils in low quality workshops appear to be less likely to be listed as friend 11 or below than pupils in the control group as well as compared to pupils in the high quality treatment group. This could be because low quality workshops also indicate poor school-level planning, and pupils may have had less time to complete questionnaires and could therefore list fewer friends. However, there are always control group pupils within the same school, so this would not fully explain the result. Overall, there may be some heterogeneity in treatment impact by intervention quality, with the Year 7

[89] The numbers here will differ slightly from the numbers presented in the Appendix for Chapter 2 because of the different number of observations available for these outcome measures.

[90] It is also possible to enter the workshop time available into the regression in place of the treatment variable, setting the control group's value to 0. However, the resultant variable is highly collinear with the treatment variable (linear correlation coefficient = 0.99) and gives identical results to simply using the treatment dummy.

treatment effect on the number of times a pupil is listed as a friend stronger in high quality workshop groups than in low quality ones.

Robustness checks

One mechanism through which the intervention effect could operate is if treatment group pupils are less likely to receive zero listings, or themselves list no friends at all. This matters, because reporting zero friends could be an indication of absence from school rather than lack of popularity, and the same could be true of receiving no listings (either that the pupil is frequently absent, or that their friends are frequently absent). Since the UKRP workshops appear to have had some small effect on absence rates in Year 7 (see Chapter 2, 'Programme impact'), this could be generating the difference. Table A3.4 reports the results of logistic regressions which follow the same patterns of controls as Table 3.4, but on the left hand side is a dummy to indicate whether the corresponding outcome variable is zero. Here we see that the treated group tends to have fewer zeroes than the control group for most outcomes at most time periods, but this is not usually statistically significant. However, treatment pupils are significantly less likely to have zero listings as friend 11 or below, with the fully controlled specification suggesting that they have only 0.77 as many zeroes as the control group. An alternative way of checking the reporting patterns is by rerunning the regressions reported in Table 3.4 but excluding cases where each outcome variable is 0. These regressions are reported in Table A3.5, showing almost identical results to those reported in Table 3.4 but with slightly smaller coefficients, most notably when the outcome is the number of times a pupil is listed as friend number 11 or lower in Year 7.

It is also possible that there are excess zeroes in the outcome variables, particularly in the number of friends a respondent listed: a pupil could have a zero on this measure because they did not consider themselves to have any friends; because they chose not to respond to this section or to complete the questionnaire at all; or because they were absent when the questionnaires were administered. Running the same regressions using zero-inflated negative binomial regressions does not change the results for any of the outcome variables (Table A3.7). Likewise, running the same specifications using ordinary least squares regressions generates almost identical results (Table A3.6).

As mentioned above, pupils whose friends are not in the same school may are likely to have an in-degree friendship score which is biased downwards, as their friends are less likely to be in my evaluation sample. In most cases I know the proportion of friends a pupil lists who are in different schools.[91] I can therefore run a robustness check excluding pupils who have more than 50% of their out-degree friends in a different school. This accounts for 608 pupils, and 695 observations, and excluding them from regressions does not change the results I obtain (Table A3.8). Thus my results are robust to using different estimation techniques and to excluding pupils whose friendship scores may be particularly biased.

Mediation through depression score

The primary outcome of the UKRP evaluation was the depression symptoms score on the Children's Depression Inventory, as described in Chapter 2, and data for this was obtained from the same questionnaire as the friendship data. Internalising problems are known to be related to poor peer relationships, although the causality is complex (Wentzel, Barry & Caldwell, 2004; Klima & Repetti, 2008; Rees, Goswami, Pople, Bradshaw, Keung & Main, 2013). Barrett et al. (1999) suggest that a major failure of many social skills interventions is to teach skills without addressing self-esteem or confidence, the absence of which may inhibit the expression of prosocial behaviour. It is therefore possible that the impact of UKRP on peer nominations partly operated through improvements in mental health.[92] In Table 3.7 I present estimates of mediation following the procedure outlined by Baron and Kenny (1986). Since only the total number of in-degree friends and the in-degree friends score for listings of 11 or lower are significantly affected by the intervention in my preferred (fully controlled) specification, I only report these two outcomes here (for the other two outcomes there is no significant impact of intervention to be mediated here, see Table 3.4).[93]

In order for this procedure to give evidence of mediation, there should be a significant impact on intervention assignment on the outcome variable in Step 1; a significant impact of intervention on the depression symptoms score in Step 2; and when the depression score is included in the regression of outcome variable on intervention assignment in Step 3, the coefficient on the depression score should be significant and the coefficient on intervention should be attenuated relative to that seen in Step 1. In this sample (restricted by the availability of the depression symptoms score)

[91] The only cases in which I would not know this is when the listed friends are ambiguous. However, as discussed above, most names are matched.
[92] Depression symptoms score is not an ideal measure of confidence or self-esteem, but many items ask about this.
[93] I also present the specification with the three outcome periods pooled together for simplicity, but separating out the outcome periods does not change my results.

we see no impact of intervention on the total number of in-degree friends in Step 1, and an insignificant coefficient when the depression score is regressed on intervention assignment in Step 2. There is therefore no statistically significant effect to mediate, and the depression score cannot be a mediator. In addition, although the depression score is a significant (negative) predictor of the in-degree in Step 3, the coefficient on ‘treated’ is almost unchanged relative to Step 1. When looking at in-degree friends listed 11 or lower, this outcome is significantly related to intervention in Step 1, although again the depression score is not related to the intervention assignment variable in Step 2. In Step 3 the depression symptoms score is again a highly significant predictor of the in-degree of friendship, but the coefficient on ‘treated’ is almost unchanged relative to Step 1. In addition, the correlation between the *change* in depression symptoms score and the change in in-degree friend scores is not significant (table not shown). Thus there is no evidence that the intervention impact on friends was mediated by pupil’s depression symptom scores.

The two measures are certainly correlated: a one standard deviation increase in the depression symptoms score is associated with a 7.3% reduction of in-degree friend nominations, and the depression symptoms score is weakly negatively correlated with the in-degree friends measure across all periods (see Table 1.15). However, one complication of the analysis presented here is that the intervention impact for both the friends score and the depression symptoms score is only seen at postintervention, and since I do not have a baseline friends score I cannot calculate the correlation between the change in friendships and the change in depression score from baseline to postintervention. Thus I may not be able to detect mediation at the point at which it would be most likely to be occurring. It is also notable that there is no significant impact of the intervention on the depression score when using the same statistical approach as used for the friendship scores (differences), while there is an impact when using difference-in-differences as reported in Chapter 2. This may indicate a limitation of the method used to estimate the intervention impact on friends. [94]

Cost-effectiveness

I can use the cost estimates presented in Chapter 1 to assess the cost effectiveness of the intervention in promoting pupils’ popularity, as I did for the intervention impact

[94] Another complication here is that due to missing depression symptom scores, including this measure reduces the sample size, which also tends to reduce the size and significance of the coefficients on ‘treated’ for the friendship measures.

on depressive symptoms scores, absence from school and academic attainment in Chapter 2. Table 3.8 sets out estimates of cost effectiveness in terms of standard deviations of the main outcome variable (in-degree friendship nominations). The table uses the incremental cost-effectiveness ratio (ICER), defined as:

ICER = (cost of intervention – cost of control)/(effect in intervention – effect in control)

Since the control group will not incur any costs which the intervention group does not also incur, I can use the estimates of the per pupil intervention costs set out in Chapter 1 in the numerator. The effect size is in the denominator. In effect, this gives the cost per standard deviation of an increase in popularity.

I calculated two effect sizes: one using the standardised in-degree friends score as the outcome in OLS regressions (Table A3.6), which gives an effect size of about 0.08 SD. An alternative is to use the IRRs from the negative binomial regressions in Table 3.4 to estimate the increase in friendship nominations at the median of the control group in Year 7, and to divide this by the standard deviation of the in-degree score in the control group at this point in order to generate an impact in terms of standard deviations. This gives a figure of 0.068. Taking the mean of the two gives an effect size of 0.074 SD, which is the figure I use to estimate cost effectiveness, and is equivalent to 3 percentile points improvement or half an extra friendship nomination at the (control group) median. Note that per-pupil costs partly depend on the number of workshop groups each trained facilitator teaches (because of the high fixed cost of the training), so here I have given two scenarios for each of the 2007 and 2009 costs, representing the scenarios shown in Table 1.16 (maximum - facilitator teaches 225 pupils; minimum – facilitator teaches 45 pupils). Further scenarios could be estimated using the cost information presented in Table 1.16. Table 3.8 also includes 95% confidence intervals around the point estimates (again in terms of standard deviations) and estimates of the cost-effectiveness of the intervention based on the upper and lower bounds of the intervention impact represented by these (the lower bound cost estimate has been omitted, since the estimated effect size is slightly negative and it would not make sense to pay for a negative impact).

Because the effect size on friendship is slightly smaller than that for depression symptoms or absence reported in Chapter 2, the cost associated with an increase of 1

SD in this measure will consequently be higher. However, in the absence of any other literature on interventions of this sort which report cost information the magnitudes are difficult to interpret.

Discussion and conclusions

The programme impact I estimate suggests that in the short run UKRP pupils gain half a friendship nomination on average, and that this will usually be driven by being listed more often as a more distant friend – friend number 11 or below. The intervention does not appear to have a significant impact on the number of in-degree nominations a pupil receives as a close friend (friends 1-10). Thus the programme increases pupils' popularity in the short run, without necessarily changing their main friendship group.[95] Thus the popularity effect is not large: it is equivalent to moving up 3 percentiles in terms of friend nominations taken at the control group median. However, there are two reasons for thinking that even a small effect such as this could be practically significant. First, much of the clinical literature on social relationships focuses on the bottom end of the distribution, particularly on pupils who are rejected by their peers and the impact this has on their development (e.g. Bierman, 2003). Table A3.4 shows that intervention group pupils are less likely to receive no nominations at all than control group pupils, suggesting that peer rejection may be lower as a result of the programme, with important consequences for the pupils concerned. Second, having a greater number of 'weak ties' such as connections of liking or acquaintance may prevent overall social fragmentation, serving as linkages between more strongly-connected cliques (Granovetter, 1973). This could be important both for overall classroom climate, and for maximising the number of contacts individuals can call upon for practical support and information, at school and in the future (Yakubovich, 2005; Granovetter, 1995).

The economic impact of having more in-degree friend nominations could also be important. Conti et al. (2012) estimate that moving from the 20th to the 80th percentile in the distribution of popularity at high school results in a 10% wage premium 40 years later. Applying this estimate to my data would suggest a 0.5% wage premium on

[95] The lack of programme impact observed in the nominations as a close friend (friends 1-10) could also be due to a ceiling effect in this measure: in the control group, the median number of friends listed (out-degree friends) is 12. This suggests that the out-degree of friends numbered 1-10 is already full for most pupils, and so intervention group pupils could only improve their in-degree score in one of three ways: by being nominated as an additional friend by the 25% of pupils who list 1-9 friends; as a result of an increase in the response rate (fewer zeroes); or by displacing other pupils (control group pupils) in respondents' top ten. By contrast, 40% of control group pupils nominate between 10 and 23 out-degree friends (inclusive), so there may be more room here to see an effect.

average for a shift of 3 percentile points, but since the intervention impact on friend nominations does not last beyond the first year it is unclear whether we should expect any longer term impact on other outcomes. Alternatively, the estimates of cost-effectiveness I report essentially measure the cost of an improvement of 1 SD in popularity, which in my data would be equivalent to moving from the control group median to the 84th percentile of popularity at postintervention. If the wage premium reported by Conti et al. (2012) could be applied linearly to this population, and if the observed impact could be sustained beyond postintervention, the intervention could very well be cost effective: a 5% wage premium on a median full-time salary of £26,500 (ONS, 2012) would be £1,325, which is within the range of the costs presented in Table 3.8. However, it is also not clear whether an intervention that is capable of shifting most pupils up 3 percentile points is capable of shifting any one individual by 30 percentile points, so this may be a misleading metric to use. It is also not clear that a 10% wage premium for a 60 percentile shift necessarily means a 0.5% wage premium for a 3 percentile point shift, as the impact across the distribution is not necessarily linear (e.g. Lindqvist & Vestman, 2011). Interestingly, estimates of heterogeneity in programme impact suggest that pupils with low prior attainment see a much greater increase in popularity as a result of the programme. Lindqvist and Vestman (2011) find that noncognitive skills are more important to labour market outcomes for individuals with low cognitive ability than for those with higher levels of cognitive ability, so any programme which could have an impact on the noncognitive skill levels of these pupils could be particularly valuable. However, the decay of the programme impact is also important. Interpreting the popularity effect as indicative of social capital accumulation, this would suggest that social capital can be lost as well as gained. Interestingly, the evaluation of the Fast Track intervention with at-risk children found a similar result – an impact on popularity as measured by in-degree friend nominations at postintervention, but no effect two years later (Conduct Problems Prevention Research Group, 1999, 2002). Maintaining levels of social capital may require continued teaching of social skills, or maintenance of an environment conducive to pro-social behaviour.

UKRP teaches social problem solving skills. Popularity is associated with social behaviour (Youniss & Haynie, 1992), so it is plausible that by using these skills intervention pupils became more likeable and popular. Pupils also reported using the pro-social UKRP skills more than any other skills, particularly negotiation, assertiveness and compromising, and teachers' reports of pupils' use of skills agreed with this (see Chapters 5 and 8 of Challen et al., 2009). However, the experimental design did not include an attention placebo control, and in the absence of this I cannot disentangle the impact of pupils using the taught skills to improve peer relations (an effect of teaching the curriculum) from the impact of spending time in a small class with a supportive teacher, talking about matters of concern to them. Both of these features of the programme could contribute, and indeed, facilitators rated the small group size as being very important to UKRP lessons (Chapter 6, Challen et al., 2009). Barrett, Webster and Wallis (1999) suggest that the majority of peer relationship interventions have poor impacts because they focus on developing social skills alone, without also addressing the feelings of insecurity which can accompany poor relationships, and they develop an intervention aimed at promoting self-esteem as well. Thus the opportunity to get to know other pupils intimately in a small class, just after pupils have moved from small primary schools to large secondary schools, could have a substantial impact on peer relations. Teachers reported that pupils were generally supportive of each in UKRP lessons, and 75% thought that the workshops would improve pupils' relationships with their peers, and that shy pupils and those lacking in confidence would particularly benefit from the programme (Chapter 6, Challen et al., 2009). Thus there is anecdotal evidence that some of the non-skill components of the programme might benefit pupils as much as the social skills elements. This may also help to explain the fade-out of the effect into years 8 and 9: after Year 7 pupils will no longer have regular small group sessions with their UKRP peers, and the positive impact on peer groups may therefore decline. Of course, pupils may also forget the skills and stop using them.

Another factor which could have an impact on the measured intervention effect is the allocation to intervention by class, meaning that intervention pupils were likely to have many academic lessons with other intervention pupils, rather than just meeting them in the UKRP lessons. If regular contact with other intervention pupils positively reinforced the impact of the programme, we may be observing a larger impact than if pupils had been individually allocated to intervention or control (this is a version of the Manski reflection problem, Manski, 1993). However, in almost all cases classes rather than pupils were allocated to the intervention, so I cannot test whether this made a difference to the results. Related to this is the concern that there may be substantial spillovers from the intervention group to the control group, since all these pupils are in the same school. However, I would expect this to bias downwards my estimates of intervention impact. Moreover, I would expect spillovers to be more acute for the within-year control group than for the year-above control group: 85% of friends listed are at the same school as the respondent, and 80% are in the same year at the same school. However, a significant intervention impact is only seen in the year (Year 7) for which there is no year-above control group and for which spillovers should therefore be more pronounced. Thus although there is still a risk of spillovers within the school, this would only bias downwards my estimates of programme impact, and does not appear to do so in the expected way.

If pupils were more willing to list others as friends because of the intervention, this could also have an impact, especially since intervention pupils were grouped in classes with other intervention pupils and so are more likely to be nominated by them. However, the intervention did not appear to have a significant impact on the number of friends pupils listed, once we control for pupil characteristics, so it is unlikely that it is this reporting effect which is driving the impact on nominations received. The intervention appears to have had a positive short run impact on other factors, such as absence and depression symptom scores (see Chapter 2, 'Programme impact'). Lower absence in the intervention group would mean that they were more likely to list at least some out-degree friends, and because of concentration within groups this could also mean that intervention pupils received more in-degree nominations. If pupils are happier, they may be more prosocial in addition to any social skills they have learned (Barrett et al, 1999). However, we have already seen that the impact on in-degree friendships cannot be attributed to an increased propensity to list more out-degree friends, and the intervention impact does not seem to be mediated through the depression symptoms scores.[96] Thus it may be possible that the social skills components of the intervention are primarily responsible for the impact on friends, independently of any impact on pupils' subjective wellbeing.

One major strength of the study is the use of a large sample with reliably measured outcome data. The peer nomination measures I use as outcomes are similar to measures found elsewhere, such as those in the Add Health dataset, which asked students to name up to five male and five female friends at their school (UNC Carolina Population Center, 2013), or the sociometry measures described in Conduct Problems Prevention Research Group (1999), which asked children to list the classmates they

[96] Lower absence could also raise a pupil's in-degree nominations because they would be at school more often and therefore have more social contact with their peers. However, given the small average magnitude of the intervention impact on absence – 1.3 more school days attended per year – this seems unlikely to be the main channel.

‘most liked’ and ‘least liked’. However, my dataset is new, and drawn from a population broadly representative of students at nonselective secondary schools in England. Peer-nomination data is more reliable than teacher- or self-reports of friendship networks: children themselves are the only true judges of who they consider to be their friends, and even teachers are constrained in their ability to observe peer relationships in all relevant contexts (Henry & Metropolitan Area Child Study Research Group, 2006; Parkhurst & Hopmeyer, 1998; Gest, 2006; Hartup, 1996; Bondonio, 1998). The in-degree (peer-reported) number of friends is also made up of multiple reports of friendship, from all other respondents in the cohort, and this may also reduce measurement error.

Pupils were also able to list a large number of friends. There were 24 response lines on the questionnaire, and 5% of respondents listed more friends than this by using margins and additional sheets of paper, so it is unlikely that the data is severely censored. The amount of space available for listing, along with the low rate of nonresponse, and the inclusion of the large majority of nominated friends in the evaluation sample, means that I can fully reconstruct the majority of friendship networks. This allows me to detect even small changes in a pupil’s popularity, and may be important more generally for understanding behaviour (Granovetter, 1973). By comparison, the Add Health data allowed for a maximum of 10 friends; the Wisconsin Longitudinal Study asked for up to three same-sex friends (Conti et al., 2012); while the Avon Longitudinal Study of Parents and Children (ALSPAC) survey asked for up to five friend nominations (Burgess & Umaña-Aponte, 2011). In all three of these datasets a minority of students within each school year is sampled, resulting in either incomplete networks or selection on those pupils with completed networks (which may not be typical of the population as a whole).

However, there may be limitations to the measures I use. Pupils with many friends outside of their school and yeargroup will have an in-degree score which is biased downwards, because these friends are not in the evaluation sample and cannot respond. However, a robustness check which excludes the 695 responses where more than half of a pupil’s out-degree friends are not at the same school does not change the size or significance of the results on any measure. In addition, the method of response – asking students to write down names themselves – could bias downwards the out-degree friendship listings of pupils who have literacy problems or who are more likely to be absent from school, as these pupils are less likely to be able to list all their friends.

Given homophily in friendships on these characteristics (Weinberg, 2007; Currarini, Jackson & Pin, 2009), this could mean that pupils with lower literacy or higher absence also receive fewer in-degree nominations than they have friends. This is aggravated by problems of deciphering barely-legible questionnaires as in Table A3.1. The large number of friends a pupil could potentially list will also increase the difference in out-degree nominations between pupils with good literacy and those without; if the question required only three friend nominations this gap would be likely to be smaller. However, this would not compromise the identification of programme impact unless the intervention and control groups differed in terms of their literacy and absence rates. The baseline data presented in Table 3.2 suggests that there were no differences at all in terms of the absence rate of treated and control pupils. Intervention pupils had slightly higher test scores, although this difference is not statistically significant (p=0.19).

The major limitation of this study is the lack of randomisation. Although the intervention and control groups appear to be well-matched on their baseline demographic characteristics, we cannot be sure that the outcome differences were not due to pre-existing differences between groups generating differential trends between the intervention and control groups.[97] When pupil characteristics are added to the outcome regressions (e.g. in Table 3.4), the size of the coefficient on 'treated' is reduced substantially for most outcome measures. This suggests that pupil characteristics are important, and since these may be correlated with class assignment this will complicate the estimate of the programme effect. In addition, the absence of an attention control group (classes of 15 students covering similar topics) means that I cannot disentangle the impact of the taught curriculum itself from greater teacher attention and a more intimate workshop setting for pupils. If the latter aspect of the programme is the active ingredient in promoting good peer relations, then this could be implemented more cheaply and easily than by teaching the UKRP curriculum.

One further issue is the lack of a baseline measure of popularity, which means that I cannot show that intervention and control group pupils were matched at baseline on the outcome variables. I also cannot use difference-in-differences methodology to assess programme impact as I did for the other outcomes in Chapter 2. As shown above in the section on mediation, this difference in methods makes a difference to the estimated intervention impact on depression symptoms, and could also make a difference when popularity is the outcome.

[97] See Chapter 2 for a discussion of the common trends assumption.

This paper contributes to the literature on pragmatic real-world implementations of wellbeing interventions for universal adolescent populations. The UK Resilience Programme was trialled with the aim of developing pupils' emotional resilience, but the programme also contains substantial components on social problem solving and developing social skills. The format of the lessons was conducive to developing closer relationships between pupils and teachers, with half the normal class size and relatively informal and fun activities. I estimate the impact of the intervention at three points in time, finding a small impact on the number of friendship nominations from peers, equivalent to half an extra friend for each intervention pupil, or a popularity shift of 3 percentiles. I present cost-effectiveness estimates, but in the absence of any other interventions of this sort which report cost information these are difficult to interpret. Using estimates of the impact of popularity on wages from a longitudinal study suggests that under certain assumptions the intervention could be cost effective, quite aside from any impacts on depressive symptom scores, absence from school and academic attainment. However, the lack of a lasting impact suggests that social capital can be lost as well as gained, and that continued efforts are required to maintain it.

Chapter 3: Figures and tables

Figure 3.1: Distribution of number of friends listed (out-degree) by year group and treatment status

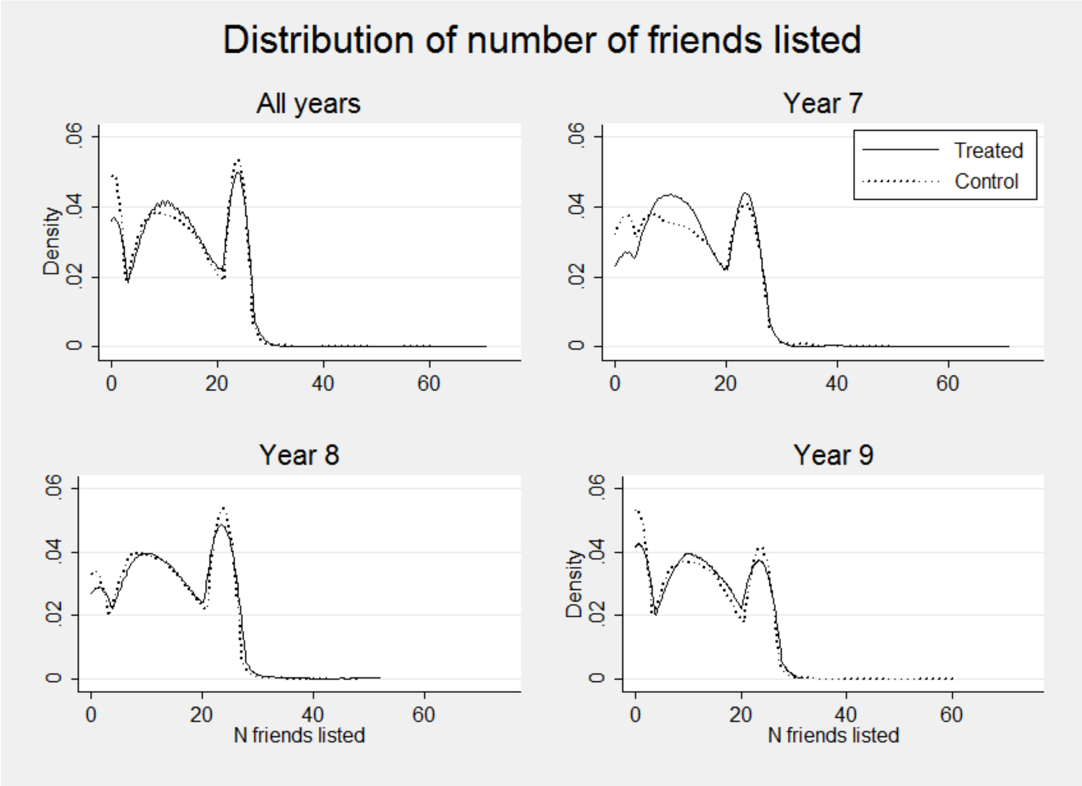


Figure 3.2: Distribution of times listed as a friend (in-degree), by year group and treatment status

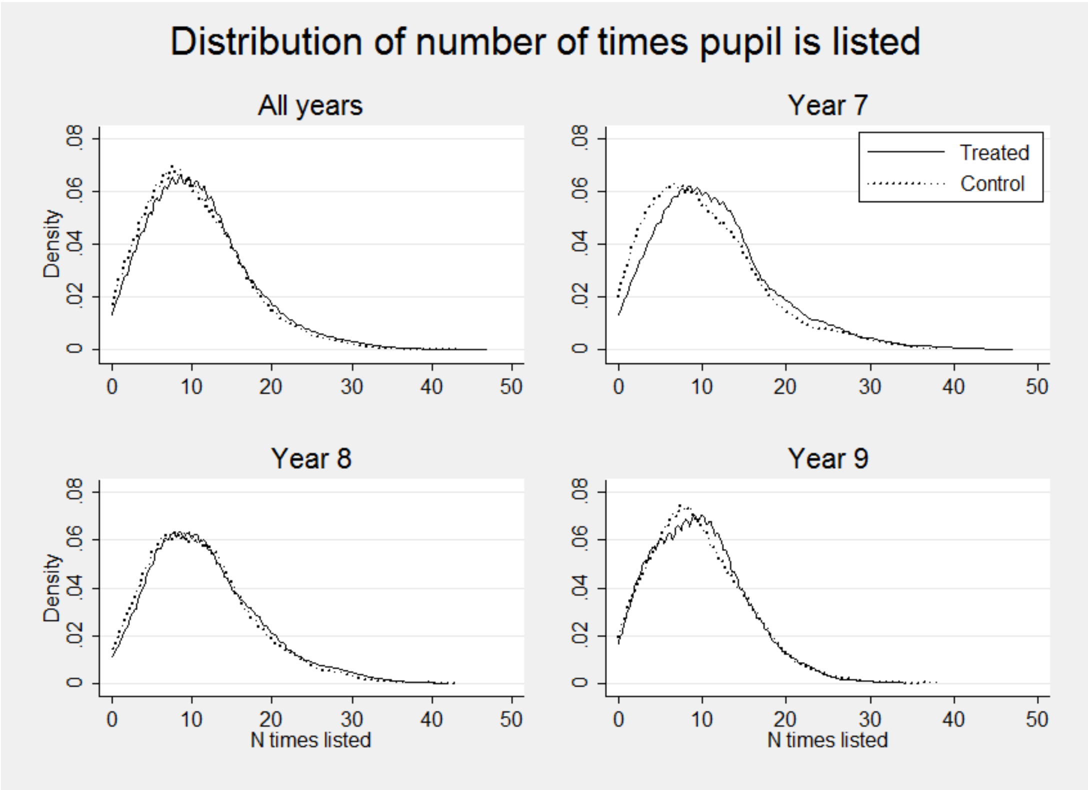


Figure 3.3: Mean in-degree friend score, by year group and treatment status

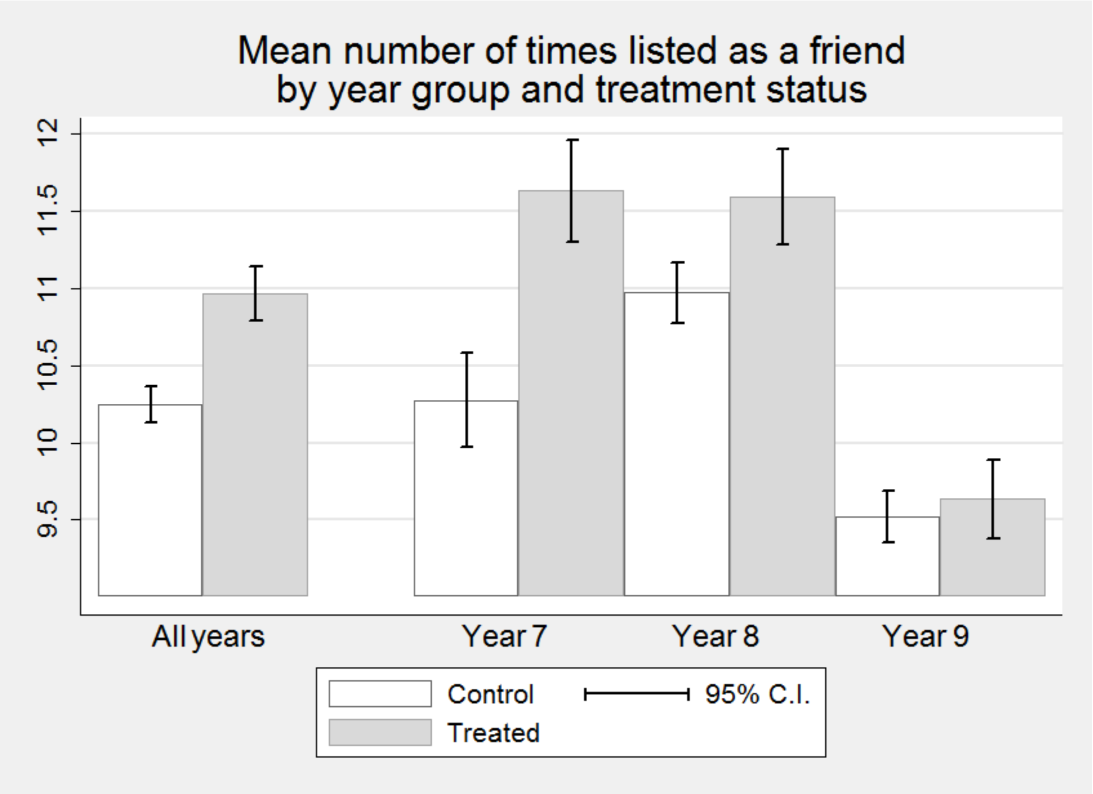


Figure 3.4: Mean out-degree friend score, by year group and treatment status

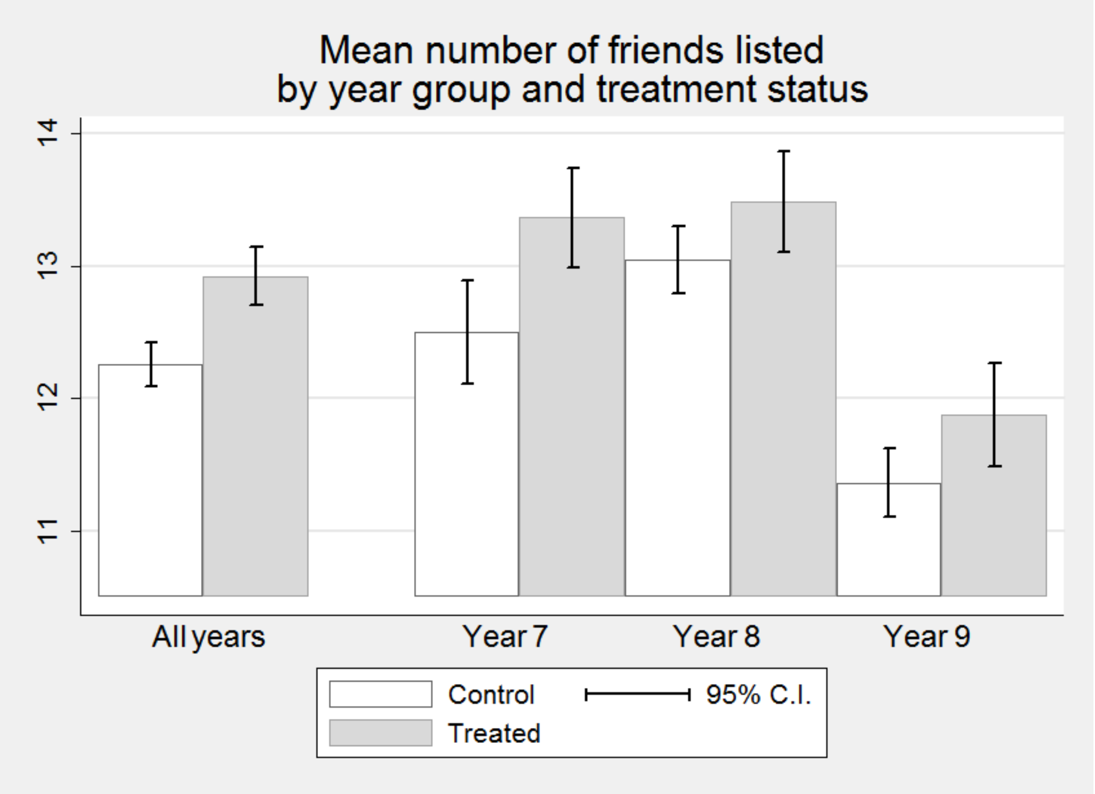


Figure 3.5: Mean in-degree top ten friend score, by year group and treatment status

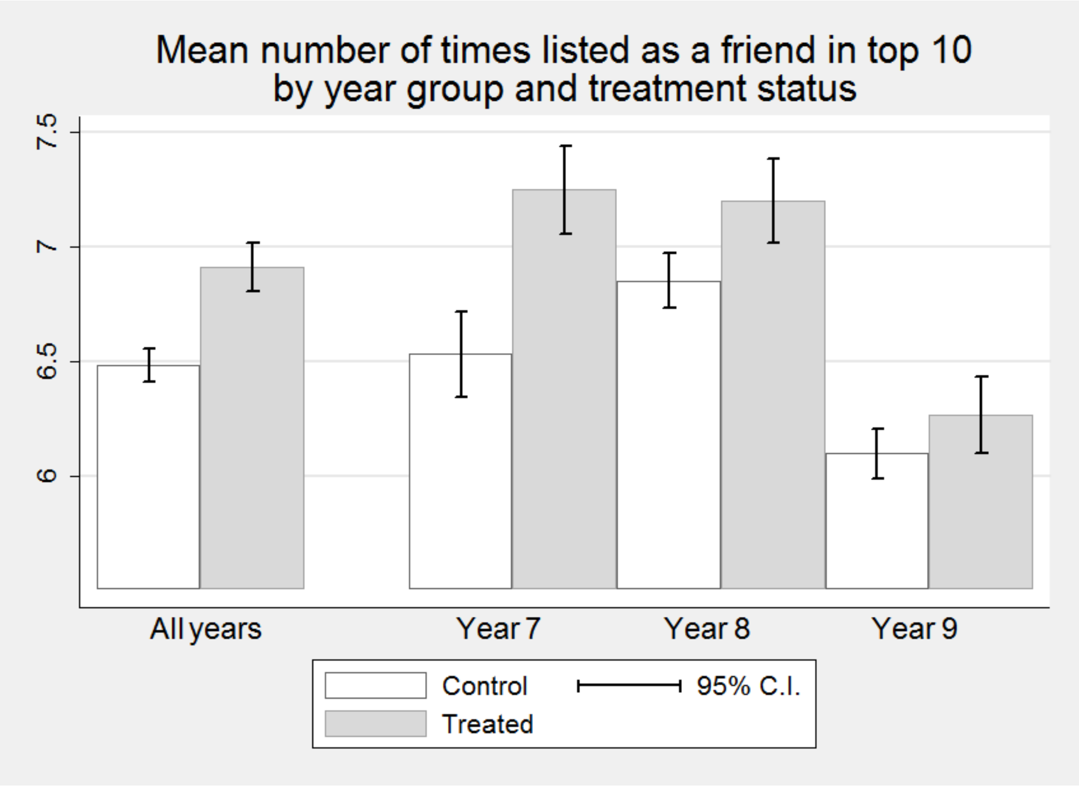


Figure 3.6: Mean in-degree score as friend 11 or below, by year group and treatment status

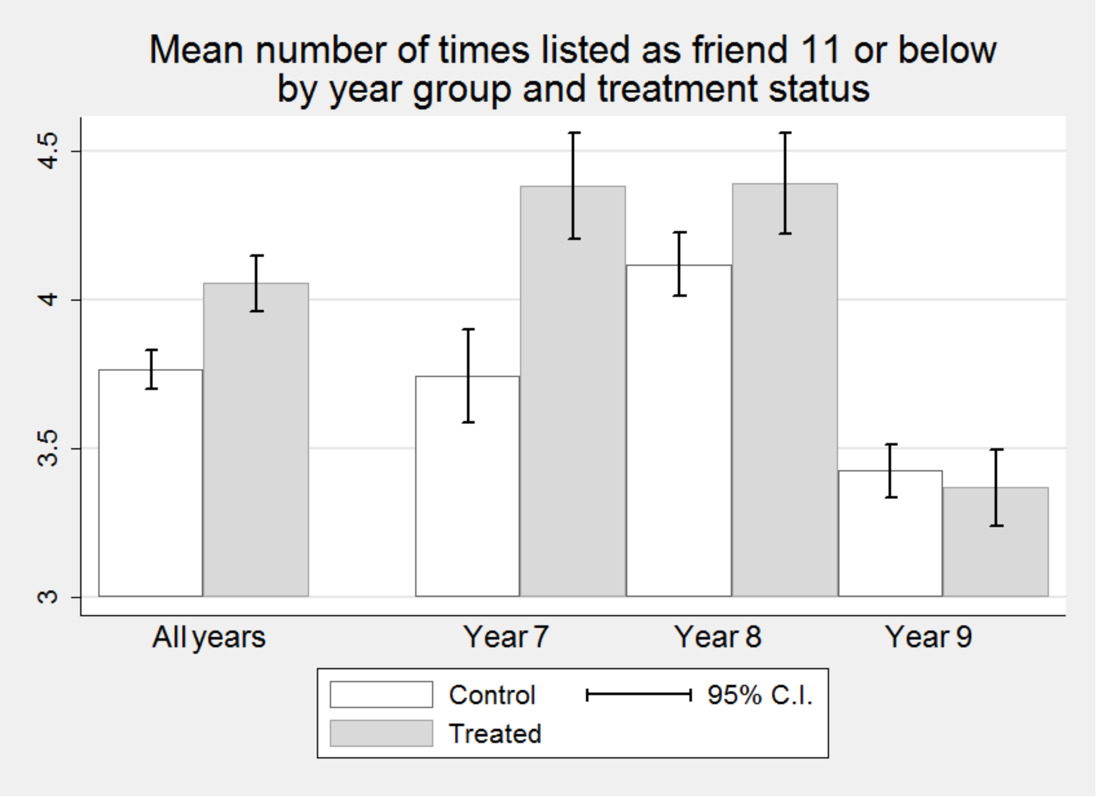


Table 3.1: Treatment and control group observations by timing and cohorts

		Year 7	Year 8	Year 9	Total
	Treated	1,940	0	0	1,940
June 2008	Control	1,967	2,252	0	4,219
	All	**3,907**	**2,252**	**0**	**6,159**
	Treated	0	1,898	0	1,898
June 2009	Control	0	2,046	2,269	4,315
	All	**0**	**3,944**	**2,269**	**6,213**
	Treated	0	0	1,848	1,848
June 2010	Control	0	0	2,097	2,097
	All	**0**	**0**	**3,945**	**3,945**
	Treated	1,940	1,898	1,848	5,686
Total	Control	1,967	4,298	4,366	10,631
	All	**3,907**	**6,196**	**6,214**	**16,317**

Notes: Table shows the observations available at each point in time, by year group and treatment status. There were 6,510 pupils in total: 4,189 pupils in the Year 7 cohort and 2,321 pupils in the Year 8 cohort. An observation does not necessarily mean that a pupil responded to the friendship question, or completed a questionnaire; rather, all pupils are counted who were on their schools' registers at the time. Pupils who did not complete a questionnaire (or who did not complete the friends question) will have the number of friends they listed equal to zero. However, they can still be listed as a friend by others in the sample. Cohorts are determined by date of birth, with over 99% of pupils in the cohort predicted by their date of birth. 4 of 6,510 pupils moved cohort during this time (0.06%), and are treated as being a member of their original cohort throughout.

Table 3.2: Demographic characteristics of sample

		All	Min	Max	Control	Treated	p-value of T-C difference
Gender	Mean	0.48	0	1	0.47	0.49	0.279
	SD	0.50			0.50	0.50	
	N	6,510			4,574	1,936	
Special Educational Needs	Mean	0.32	0	1	0.32	0.31	0.707
	SD	0.46			0.47	0.46	
	N	6,497			4,561	1,936	
Free School Meal eligibility	Mean	0.25	0	1	0.26	0.24	0.516
	SD	0.44			0.44	0.43	
	N	6,497			4,561	1,936	
Age in September 2007	Mean	11.82	10.83	13.92	11.97	11.47	0.000
	SD	0.56			0.58	0.30	
	N	6,510			4,574	1,936	
Month part of age in September	Mean	5.50	0	11	5.47	5.58	0.225
	SD	3.46			3.45	3.51	
	N	6,497			4,561	1,936	
Key Stage 2 mean score	Mean	4.48	1	5.80	4.46	4.53	0.193
	SD	0.72			0.73	0.70	
	N	6,079			4,187	1,892	
White ethnicity	Mean	0.79	0	1	0.78	0.82	0.283
	SD	0.40			0.41	0.39	
	N	6,510			4,574	1,936	
Black ethnicity	Mean	0.04	0	1	0.05	0.03	0.089
	SD	0.21			0.22	0.18	
	N	6,510			4,574	1,936	
Bangladeshi or Pakistani ethnicity	Mean	0.09	0	1	0.09	0.08	0.477
	SD	0.28			0.29	0.27	
	N	6,510			4,574	1,936	
Other Asian ethnicity	Mean	0.02	0	1	0.02	0.02	0.747
	SD	0.15			0.15	0.15	
	N	6,510			4,574	1,936	
Other or mixed ethnicity	Mean	0.05	0	1	0.05	0.05	0.680
	SD	0.23			0.23	0.22	
	N	6,510			4,574	1,936	
Lives with mother	Mean	0.90	0	1	0.90	0.89	0.438
	SD	0.30			0.30	0.31	
	N	6,328			4,411	1,917	
Lives with father	Mean	0.56	0	1	0.55	0.56	0.853
	SD	0.50			0.50	0.50	
	N	6,328			4,411	1,917	
Fraction of sessions absent	Mean	0.061	0	0.98	0.061	0.061	0.948
in Autumn term of Year 7	SD	0.08			0.08	0.08	
	N	6,153			4,225	1,928	
Fraction of sessions absent	Mean	0.037	0	0.57	0.038	0.037	0.902
in Year 6 (primary school)	SD	0.06			0.06	0.06	
	N	3,770			1,880	1,890	

Notes: Demographic data represents pupil characteristics at the start of secondary school in September of Year 7 (experimental baseline – September 2007 or 2008 depending on the cohort), unless stated otherwise. Each characteristic (e.g. gender) was regressed on a dummy for treatment assignment, with OLS regressions clustered by class. The p-value on the coefficient on 'treated' in these regressions is reported in the final column.

Table 3.3: Outcome variables by time and treatment status

Time	Variable		All	Control	Treated	p-value of T-C difference
End of Year 7	In-degree friend nominations	Mean	10.95	10.28	11.63	0.001
		SD	7.19	6.94	7.36	
		N	3,907	1,967	1,940	
	Out-degree friend nominations	Mean	12.93	12.50	13.37	0.103
		SD	8.58	8.75	8.38	
		N	3,907	1,967	1,940	
	In-degree nominations in top ten	Mean	6.89	6.53	7.25	0.001
		SD	4.30	4.24	4.34	
		N	3,907	1,967	1,940	
	In-degree as friend 11 or below	Mean	4.06	3.74	4.38	0.006
		SD	3.79	3.53	4.01	
		N	3,907	1,967	1,940	
End of Year 8	In-degree friend nominations	Mean	11.16	10.97	11.59	0.042
		SD	6.63	6.54	6.81	
		N	6,196	4,298	1,898	
	Out-degree friend nominations	Mean	13.18	13.05	13.49	0.225
		SD	8.41	8.38	8.48	
		N	6,196	4,298	1,898	
	In-degree nominations in top ten	Mean	6.96	6.85	7.20	0.046
		SD	4.02	3.98	4.08	
		N	6,196	4,298	1,898	
	In-degree as friend 11 or below	Mean	4.20	4.12	4.39	0.113
		SD	3.62	3.57	3.75	
		N	6,196	4,298	1,898	
End of Year 9	In-degree friend nominations	Mean	9.56	9.52	9.63	0.773
		SD	5.64	5.68	5.55	
		N	6,214	4,366	1,848	
	Out-degree friend nominations	Mean	11.52	11.36	11.88	0.273
		SD	8.68	8.74	8.54	
		N	6,214	4,366	1,848	
	In-degree nominations in top ten	Mean	6.15	6.10	6.27	0.493
		SD	3.67	3.66	3.70	
		N	6,214	4,366	1,848	
	In-degree as friend 11 or below	Mean	3.41	3.43	3.37	0.767
		SD	2.97	3.03	2.81	
		N	6,214	4,366	1,848	

Notes: Each outcome variable was regressed on a dummy for treatment assignment, with OLS regressions clustered by class membership (the unit of treatment assignment). The p-value on the coefficient on 'treated' in these regressions is reported in the final column.

Table 3.4: Programme impact on friends by year group

Outcome	In-degree friends			Out-degree friends			In-degree friends 1-10			In-degree friends 11+		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Year 7*Treated	0.124***	0.110***	0.052*	0.067	0.055	0.018	0.104***	0.090***	0.015	0.157***	0.144**	0.115**
SE	(0.039)	(0.038)	(0.030)	(0.042)	(0.042)	(0.041)	(0.032)	(0.032)	(0.024)	(0.058)	(0.057)	(0.048)
IRR	1.132	1.116	1.053	1.069	1.057	1.018	1.110	1.094	1.016	1.170	1.155	1.122
Year 8*Treated	0.055**	0.032	0.027	0.033	0.018	0.013	0.050**	0.027	0.012	0.064	0.041	0.055
SE	(0.027)	(0.026)	(0.024)	(0.027)	(0.027)	(0.029)	(0.024)	(0.024)	(0.022)	(0.040)	(0.040)	(0.037)
IRR	1.057	1.033	1.027	1.034	1.018	1.013	1.051	1.028	1.012	1.066	1.042	1.056
Year 9*Treated	0.012	-0.016	-0.022	0.044	0.021	0.019	0.027	0.001	-0.014	-0.017	-0.045	-0.035
SE	(0.041)	(0.039)	(0.028)	(0.040)	(0.040)	(0.037)	(0.040)	(0.038)	(0.025)	(0.056)	(0.055)	(0.045)
IRR	1.012	0.985	0.978	1.045	1.022	1.019	1.028	1.001	0.987	0.984	0.956	0.966
N pupils	6,510	6,079	6,079	6,510	6,079	6,079	6,510	6,079	6,079	6,510	6,079	6,079
N	16,317	15,537	15,537	16,317	15,537	15,537	16,317	15,537	15,537	16,317	15,537	15,537
Clusters	178	169	169	178	169	169	178	169	169	178	169	169
Demographic controls	no	no	yes	no	no	yes	no	no	yes	no	no	yes

Notes: Table shows results of negative binomial regressions of the four outcome variables (given by column headings) on dummies for year group (timing) and treatment status*year group. The first column for each outcome includes all observations; the second uses the same raw specification as the first but includes only observations which have full demographic data; and the third includes demographic and school controls. Standard errors are heteroskedasticity robust and are clustered by class group, which is the unit of treatment assignment. Controls included as dummies are: gender; SEN; FSM eligibility; broad ethnic group (5 dummies); month of birth (giving relative age within the cohort – 12 dummies); and school attended at baseline (22 dummies). In addition, the mean Key Stage 2 test score (mean of English, maths and science scores in national tests at age 11) is included as a continuous variable.

Table 3.5: Heterogeneity in programme impact by pupil characteristics
Outcome: In-degree friends (number of times listed as a friend)

	Girls	Boys	SEN pupils	Non-SEN pupils	FSM pupils	non-FSM pupils	Younger pupils	Older pupils	Below KS2 target	At/above KS2 target
Year 7*Treated	0.096***	0.000	0.085	0.041	0.004	0.068**	0.016	0.085**	0.154**	0.030
SE	(0.033)	(0.041)	(0.055)	(0.028)	(0.052)	(0.031)	(0.037)	(0.035)	(0.073)	(0.029)
IRR	1.101	1.000	1.089	1.042	1.004	1.070	1.016	1.088	1.167	1.030
Year 8*Treated	0.049*	0.002	0.038	0.025	-0.023	0.040	0.017	0.035	0.038	0.026
SE	(0.027)	(0.032)	(0.043)	(0.023)	(0.039)	(0.025)	(0.028)	(0.029)	(0.056)	(0.024)
IRR	1.050	1.002	1.039	1.025	0.978	1.041	1.017	1.035	1.039	1.026
Year 9*Treated	-0.036	-0.012	0.017	-0.034	-0.094*	-0.005	-0.048	0.003	0.011	-0.026
SE	(0.032)	(0.035)	(0.045)	(0.028)	(0.049)	(0.028)	(0.032)	(0.032)	(0.059)	(0.029)
IRR	0.965	0.988	1.017	0.967	0.910	0.995	0.953	1.003	1.011	0.974
N pupils	2,924	3,155	1,928	4,151	1,512	4,567	3,058	3,021	1,273	4,806
N	7,481	8,056	4,884	10,653	3,938	11,599	7,775	7,762	3,208	12,329
Clusters	156	159	163	163	166	166	167	166	143	163
Demographic controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

Notes: Table shows results of negative binomial regressions with the number of times a pupil is listed as a friend as the outcome, regressed on dummies for year group (timing) and treatment status*year group. Each column represents a separate regression which includes demographic and school controls, as in column (3) of Table 3.4. However, here the sample is split by demographic characteristics. Standard errors are heteroskedasticity robust and are clustered by class group, which is the unit of treatment assignment.

Table 3.6: Programme impact by intervention quality

Outcome	In-degree friends	Out-degree friends	In-degree friends 1-10	In-degree friends 11+
Year 7*High dosage treated	0.070**	0.025	0.023	0.150***
SE	(0.031)	(0.041)	(0.025)	(0.052)
IRR	1.072	1.025	1.023	1.162
Year 8* High dosage treated	0.024	-0.001	0.006	0.057
SE	(0.027)	(0.033)	(0.026)	(0.041)
IRR	1.025	0.999	1.006	1.058
Year 9* High dosage treated	-0.011	0.016	-0.017	0.003
SE	(0.032)	(0.042)	(0.029)	(0.048)
IRR	0.990	1.017	0.984	1.003
Year 7* Low dosage treated	0.003	-0.001	-0.004	0.019
SE	(0.045)	(0.064)	(0.038)	(0.073)
IRR	1.003	0.999	0.996	1.019
Year 8* Low dosage treated	0.032	0.048	0.027	0.047
SE	(0.034)	(0.039)	(0.027)	(0.056)
IRR	1.033	1.050	1.027	1.048
Year 9* Low dosage treated	-0.054	0.025	-0.005	-0.141**
SE	(0.037)	(0.056)	(0.038)	(0.059)
IRR	0.947	1.026	0.995	0.868
p-values of tests of equality between high-quality and low-quality treatment groups				
Year 7	0.136	0.634	0.480	0.075
Year 8	0.828	0.268	0.520	0.876
Year 8	0.292	0.885	0.795	0.017
N pupils	6,079	6,079	6,079	6,079
N	15,537	15,537	15,537	15,537
Clusters	169	169	169	169
Demographic controls	yes	yes	yes	yes

Notes: Table shows results of negative binomial regressions of the four outcome variables (given by column headings) on dummies for year group (timing) and treatment status*year group*workshop quality. Workshop quality is as defined in Table A3.3. Each column represents a separate regression which includes demographic and school controls, as in column (3) of Table 3.4. P-values of tests of equality are from pairwise comparisons of treatment effect for high- and low-quality workshops at the same time point, e.g. Year 7*High dosage treated is compared with Year 7*Low dosage treated. Standard errors are heteroskedasticity robust and are clustered by class group, which is the unit of treatment assignment.

Table 3.7: **Mediation of programme impact by depression symptoms scores**

Outcome	**In-degree friends**			**In-degree friends 11+**		
	Step 1	Step 2	Step 3	Step 1	Step 2	Step 3
Treated	0.026	-0.037	0.024	0.056**	-0.037	0.055**
SE	(0.017)	(0.023)	(0.016)	(0.026)	(0.023)	(0.025)
IRR	1.026	0.963	1.024	1.057	0.963	1.056
Depression score			-0.008***			-0.008***
SE			(0.001)			(0.001)
IRR			0.992			0.992
N pupils	5,967	5,967	5,967	5,967	5,967	5,967
N	14,825	14,825	14,825	14,825	14,825	14,825
Clusters	167	167	167	167	167	167
Demographic controls	yes	yes	yes	yes	yes	yes

Notes: Table shows estimates of mediation of the dependent variable (all in-degree friend listings or in-degree friend listings at number 11 or lower) through the depression symptoms score measured at the same time as the friendship data. Following Baron and Kenny (1986), step 1 gives the regression results for the dependent variable on ‘treated’ for this sample; Step 2 uses the same specification but with the mediator (depression score) as the dependent variable; and Step 3 runs the same specification as Step 1 but adds in the mediator (depression score) on the right hand side. All regressions use the preferred specification, including all demographic control variables. The impact of ‘treated’ is pooled across all three periods, but results are the same when entered separately as ‘treated*time’. Results for the out-degree friends and in-degree friends 10 or lower are not shown because the coefficients on ‘treated’ are not statistically significant in this sample and using this specification.

Table 3.8: Cost effectiveness of UKRP in promoting friendship

		Postintervention				
		Point estimate	Lower bound	Upper bound	1-year follow-up	2-year follow-up
Friend nominations	Effect size	0.074	-0.010	0.158	-	-
	Equivalent to	0.5 extra friend			-	-
	Percentile points	3			-	-
2007 costs	Min cost/impact	£ 1,052.18	No impact	£ 493.80		
	Max cost/impact	£ 2,493.62	-	£ 1,170.29		
2009 costs	Min cost/impact	£ 582.28	-	£ 273.27		
	Max cost/impact	£ 1,415.50	-	£ 664.31		

Notes: Table shows estimates of per pupil cost effectiveness under four cost scenarios, partly dependent on the utilisation of trained teachers; see Table 1.16 for detail. Upper and lower bounds represent 95% confidence intervals. The long-run costs of the intervention would be represented by the 2009 costs. In effect, the estimated cost is for a 1 SD improvement in the outcome at that point in time. Note that I have not attempted to discount the upfront investment in teacher training and workshop delivery by the cost of capital: doing this would mean that facilitators who taught 10 workshop groups in the first year after training would be more cost-effective than teachers who taught the same number of groups over a longer time period (all else being equal) because of the time dimension of money.

Chapter 4: Pupil behaviour in secondary schools

Introduction

Recent research suggests that noncognitive skills are critical to academic attainment and labour market success, as well as in many other areas of life (Heckman, Stixrud, & Urzua, 2006; Segal, forthcoming; Lindqvist & Vestman, 2011; Roberts, Harms, Caspi & Moffitt, 2007; Wolfe & Johnson, 1995). Cognitive traits may be defined as general intelligence and the ability to solve abstract problems, leaving 'noncognitive' traits to cover a broad range of skills, attitudes and preferences (Borghans, Duckworth, Heckman & ter Weel, 2008).[98] These include personality traits such as conscientiousness and antagonism (Roberts, Kuncel, Shiner, Caspi & Goldberg, 2007; Chamorro-Premuzic & Furnham, 2003; Mueller & Plug, 2006); discount rates (Borghans & Golsteyn, 2006); capacities such as self-discipline (Duckworth & Seligman, 2005); and emotional health and stability more generally (De Neve & Oswald, 2012); all of which have been shown to be associated with success in domains such as earnings or attainment.

Given their impact, it would seem important to understand the determinants of noncognitive skills. In particular, can they be taught in schools? Bowles, Gintis and Osborne (2001) suggest that schools may influence personality through social interactions and incentive systems similar to those observed in workplaces, and that this might account for the impact of schooling on earnings which remains after controlling for academic attainment. But there is disagreement on the extent to which noncognitive traits are malleable, and on whether they are stable through the life course. We can distinguish between individuals' rank order and their mean level of a trait: as with cognitive skills, it is possible for mean levels to change while relative rank remains the same. Helson, Kwan, John and Jones (2002) present evidence of substantial mean-level change in personality characteristics during adulthood, particularly on traits such as conscientiousness and agreeableness that are relevant to behaviour. Borghans et al. (2008) suggest that even the rank order of personality measures may become stable only between the ages of 50 and 70, much later than rank order IQ (which they estimate as stabilising around middle childhood), and that noncognitive skills are more malleable

---

[98] Nevertheless, noncognitive traits or skills are likely to involve cognitive processes, and their development may be influenced by cognitive skills.

than cognitive skills. Against this, others have argued that the rank order of many personality traits becomes stable in early adulthood and remains so until old age, particularly for conscientiousness (Terracciano, McCrae & Costa, 2010; Lucas & Donnellan, 2011; Specht, Egloff, & Schmukle, 2011). These two views may not be incompatible: the average member of the population might not exhibit much (rank order) personality change during adulthood, but this is not to say that they are unable to change. We should be more interested in whether individuals *could* change, possibly with the help of interventions designed to develop noncognitive skills and environments which promote them.[99] There is a wealth of evidence from the intervention field suggesting that improvements are possible (Durlak, Weissberg, Dymnicki & Schellinger, 2011; Heckman, 2000; Wilson & Lipsey, 2007).

However, one issue with these estimates is that the expression of noncognitive skills depends on the context in which people find themselves (Borghans et al., 2008; Roberts, Harms, Caspi & Moffitt, 2007). Different roles and incentives will induce people to display different traits, so observed behaviour should be interpreted as an interaction of individuals' traits with the situations in which they find themselves (Roberts, 2007). Helson et al. (2002) stress that social roles and situations shape personality development through the life course, so if people sorted into specific environments early in life and stayed there we would see substantial persistence of measured noncognitive traits even if such traits were in fact malleable. The value or usefulness of particular noncognitive traits is also likely to depend on context.[100] We might therefore question whether the sort of behaviour exhibited under artificial incentives in laboratory experiments reflects what happens in the outside world. Since it is impossible to reconstruct in the lab the environmental and incentive factors governing everyday behaviour at work and at home these experiments may be unable to measure the traits most relevant to life outside the lab, a problem further complicated by the use of inadequate methods of measurement (e.g. using self-reports rather than actual data on behaviour). Laboratory measures may identify traits which are *related* to those most relevant to life in the real world, but ultimately these are only proxies for the personality and behavioural factors we should be interested in. Thus to understand the nature and

[99] Borghans et al. (2008) call the life course changes observed in a typical member of the population 'normative', while changes due to deliberate effort or atypical life events 'nonnormative' change.
[100] There may also be individual heterogeneity in returns to noncognitive traits. For instance, Mueller & Plug (2006) and Bowles et al. (2001) report differential returns to specific personality traits by gender.

distribution of noncognitive skills, it is important to observe and measure these skills in appropriate and realistic environments (Bowles et al., 2001).

Even when noncognitive skills are investigated in context, standard measures of these traits may be less reliable than measures of cognitive ability, with most relying on self-reports or responses to hypothetical situations (Borghans et al., 2008). Self-reports are unreliable indicators of behaviour and performance, with the average respondent overly optimistic about their own abilities and prospects (Dunning, Heath, & Suls, 2004). Yet studies of personality and social psychology largely rely upon self-reported measures, despite the clear benefits to studying observed behaviour (Furr, 2009; Baumeister, Vohs, & Funder, 2007). Exceptions to this include Segal (2008), who uses teachers' reports of students' behaviour to assess noncognitive skills along five dimensions, although these rely on teachers' global assessments of behaviour at a point in time, which may be subject to recall and perceptual biases. Lindqvist and Vestman (2011) use scores from a structured psychologist-conducted interview, yet even here the interviewer does not know the interviewee and does not observe his character and behaviour in context. Thus although these measures are both clear improvements over self-reports, they may still be biased and noisy measures of noncognitive skills and attributes.[101] I use data on behavioural incidents from school databases as a measure of noncognitive skill (following Segal, 2008; and Bowles et al., 2001), investigating its determinants and persistence through time and in different contexts.

I believe that this provides an excellent measure of economically-relevant noncognitive skills, for a number of reasons. First, the data I use is probably a more reliable measure of behaviour and noncognitive skills than others in the literature. I use database records of behavioural incidents from four secondary schools (students aged 11-16). At these schools, staff are expected to record every behaviour incident above a certain threshold of severity, and include details such as date and time of day; the nature of the incident; the activity at the time of the incident; and the staff involved. I can therefore construct incidents counts per pupil, in total and by incident type. The data is also the result of multiple observations of students' behaviour by multiple observers over an average of more than two academic years per pupil, reducing the likelihood of bias from single observers and reducing the noisiness of the measure through repeated

[101] Petrides, Frederickson and Furnham (2004) use data on absences and exclusions from school as realised behaviour measures. However, exclusions are the result of behaviour not a direct measure of it, and absences measure only a very specific dimension of behaviour.

observations. Using administrative data also means that I do not have a problem with selection biases into measurement – outcome measures are available for all students within each school in the relevant cohorts. Moreover, all of the behaviour data I use and most of the other accompanying data on students' characteristics was collected routinely for administrative databases, so is unlikely to be subject to reporting biases resulting from the Hawthorne effect or from selective non-reporting. I am also able to compare the behaviour incident scores with pupil- and teacher-reported questionnaire measures of pupil behaviour for a subset of pupils.

Second, as discussed above, context is an important determinant of noncognitive skills (Borghans et al., 2008; Roberts, Harms, Caspi & Moffitt, 2007). Observing behaviour at school provides me with a measure of noncognitive performance in a real-life situation, which may be more reliable than measures constructed from laboratory-based experiments. In particular, the incentives for and consequences of good or bad behaviour are real, well-understood and habitual, as are the routines and activities students engage in. Thus my use of behaviour data fulfils the criterion of being a situation-specific measure of noncognitive skill as recommended by Bowles et al. (2001). Moreover, I present evidence that the schools for which I have data are nationally representative, and that they use standard disciplinary methods. There is also relatively little selection into these situations. Adults select into and out of jobs based on their preferences and traits, so there is likely to be a strong pre-existing association between noncognitive characteristics and the environments in which we observe them (Rutter, 2006). Pupils at these schools have little discretion over their environment: they have little or no choice over the school they attend, cannot choose their teachers or classmates, and have limited choice over the subjects they study.[102] I would therefore expect significantly less selection into this specific context than would be observed in the average adult workplace, allowing a more general understanding of the personality traits displayed.[103] So although I am only able to observe behaviour in one context, and this might not generalise to all other situations, this is at least a real-world context with many similarities to other environments encountered through the life course, increasing the probability of my findings having some external validity.

[102] Specifically, and depending on the school attended, pupils usually have no choice over subjects studied at ages 11-14 and limited choice at 15-16.

[103] There is likely to be strong selection into schools by parental characteristics, but I provide evidence below that the schools for which I have data are fairly typical of state secondary schools in England.

Third, the behaviour required by schools may be similar to that valued by employers (Bowles et al., 2001). Roberts, Harms, Caspi and Moffit (2007) find that conduct disorder during adolescence is a significant predictor of counterproductive work behaviours. Observed school behaviour is also a reliable indicator of an individual's possession of other noncognitive traits such as self-control and aggression (Coolidge, DenBoer & Segal, 2004; Petrides et al., 2004; Resing, Bleichrodt & Dekker, 1999; Aluja-Fabregat, Balleste-Almacellas & Torrubia-Beltri, 1999; Simo & Perez, 1991; Maliphant, Hume & Furnham, 1990). Thus observed behaviour in schools may approximate exactly the sort of noncognitive skills rewarded in the workplace.

Behavioural incidents in schools are also interesting because they can be a serious problem. A survey by a UK teachers' union found that two-thirds of teachers had had to deal with a violent pupil in the current academic year (ATL, 2012); and two-thirds of secondary school teachers responding to the 2012 Teacher Voice Omnibus survey said that they believed negative pupil behaviour was driving teachers out of the profession (NFER, 2012). Poor behaviour is directly associated with reduced academic success for the perpetrators, which could further reduce labour market and life chances (Gutman & Vorhaus, 2012). Misbehaving pupils also impose negative externalities on other pupils, reducing peers' attainment by taking up teacher time and reducing the time that can be spent on learning for the whole class (Lazear, 2001; Lavy, Passerman & Schlosser, 2012). Thus as well as representing noncognitive skills, behavioural incidents may play a role in the formation of cognitive skills. Poorly behaved pupils may also induce other pupils to misbehave, further reducing the effectiveness of lesson times, and causing their peers to develop counterproductive attitudes and behaviour.

I use a dataset of pupils' behaviour in four schools over six academic years to investigate the following questions:

(1) What is the distribution of behaviour incidents - do all pupils misbehave?
(2) Do demographic characteristics predict behaviour incidents?
(3) Are there different dimensions of behaviour incidents?
(4) Is behaviour persistent?
(5) Is behaviour context specific?

My main findings are as follows:

- Two-thirds of pupils have at least some behaviour incidents, but behavioural incidents are highly concentrated amongst a small group of pupils. Less than 10% of pupils are responsible for over half of all recorded behaviour incidents.

- Demographic characteristics are strong predictors of the number of behaviour incidents a pupil is involved in. Boys; pupils eligible for free school meals; pupils from white ethnic backgrounds; pupils who are autumn-born (and therefore older within the cohort); pupils with lower academic attainment and ability; and pupils who do not live with both of their parents all have significantly more behaviour incidents. However, these characteristics explain only a small part of the variation in incidents per pupil.
- Different types of incident represent different dimensions of misbehaviour. Specifically, there appear to be two main dimensions to the incidents reported: general rule-breaking (covering all incidents), and confrontational behaviour (verbally and physically aggressive acts only). Pupils with many confrontational incidents tend to be involved in many minor incidents as well, but the reverse is not true. The majority of incidents are minor ones; violence is infrequent and few pupils are involved. The demographic characteristics predicting incidents do not vary much by incident category.
- A pupil's relative rank in terms of behaviour incident rate is highly persistent through time. However, the overall incident rate varies with age.
- Pupils who misbehave appear to do so consistently with different teachers. There are significant differences in behaviour rates by time of day, day of the week, and in different subjects. This suggests that misbehaviour is partly driven by context, and that it may therefore be malleable. However, a pupil's relative rank is maintained throughout high-incident and low-incident times of the day and week, and may also be maintained in different subjects, suggesting strong trait persistence.

These findings suggest that pupil misbehaviour is related to background characteristics, but that these do not account for much of the variation in incident rates. In particular, background characteristics do not explain the strong concentration of incidents amongst a few pupils. The difference in behaviour rates in different contexts (different subjects and times) suggests that behaviour is malleable: conditions and incentives could be adjusted so as to minimise disruption and promote positive behaviour. However, the maintenance of relative rank across different contexts suggests that the underlying tendency to misbehave may represent a stable personality trait. One consequence of the extreme concentration of behaviour incidents is that targeting this group of pupils for intervention could be cost effective even if the interventions were expensive: only a few pupils would need to improve their behaviour to see a large decrease in problematic behaviour in schools.[104] This is particularly true if both mean level and rank order of noncognitive traits are malleable well into adulthood (Borghans et al., 2008). Thus policies to rearrange the external environment to minimise misbehaviour (the context) as well as policies to address individuals' tendencies to misbehave could both be pursued. If effective, they could improve noncognitive outcomes and increase the efficiency of schooling in developing cognitive skills through reducing disruption.

In the next section I present information on the policy context and the schools for which I have data, and provide an overview of the data. I then use these data to outline the types of incident and their relative frequency, and investigate whether incident counts are associated with pupil characteristics. I then look at these associations by incident type, before going on to examine the persistence of (bad) behaviour through time and in different contexts. The final section concludes.

Context and data

I use data on pupils' behaviour from the databases of four schools in two different regions of England. All four schools had participated in the UK Resilience Programme evaluation, and had provided behaviour data as part of the evaluation. However, not enough schools were able to provide behaviour data for enough cohorts for me to be able to use incident data as an outcome variable for the programme evaluation. Nevertheless, the schools which provided data are broadly typical comprehensive schools, and the data are very detailed, so I use them to investigate aspects of pupil behaviour in this chapter and the next. The recruitment of LAs and schools was described in detailed in Chapter 1 ('Context and recruitment'). Here I outline the behaviour data available from the schools which were able to provide it; the policy context for behaviour in schools; and provide descriptive statistics for the schools relative to the general population of schools in England. These can also be compared to the characteristics of all 22 UKRP schools presented in Chapter 1 (Table 1.9). This is important to understand whether these schools are typical of secondary schools in England and so whether my results are likely to have external validity.

[104] Assuming, of course, that any such interventions were actually effective in reducing incidents.

Data

All four schools use standardly available software to record behaviour incidents, such as IRIS (IRIS, 2013). Schools are not legally obliged to record behaviour incidents in this way (DfE, 2012b), but many schools use software to record behaviour incidents, detentions, or other data relating to pupil behaviour.

The four schools for which I have data use different database systems, and therefore record different information. However, the schools record data on a similar range of incidents. It is clear that one of the four schools records more minor incidents than the other three. For instance, this school will record even minor infringements of school rules such as failing to bring equipment to a lesson once; in the other three schools, only repeated failures to bring equipment are recorded. Even with incidents involving disruption and violence it is difficult to gauge the severity of each incident, and it is possible that schools may also have differing thresholds with respect to reporting these. Thus the number of incidents per pupil is only meaningful in the context of the school the pupil attends, but controlling for this, reporting appears broadly consistent through time within a school, and the characteristics of pupils who are involved in incidents are similar across schools. However, I am unable to judge whether these schools have similar levels of poor behaviour overall, and I do not aim to do so. All the comparisons I will make will be within schools.

The datasets give details of pupils' behaviour incidents for each school day, usually with the type of incident, the outcome, and often with the lesson or location, and time of day. Two schools have included descriptions of incidents, and one school has included details of the staff members involved. The date of the incident is always included. The data covers dates between 7th September 2005 and 7th April 2011, or 1,116 school days across 5.5 academic years.

Not all pupils at the school will appear in these databases: if a pupil is not involved in any incidents, they will not appear. For this reason, it is necessary to use school census data from the National Pupil Database (NPD). The NPD covers all pupils at state schools in England, giving the school attended (with entry and leaving dates) and basic characteristics of pupils, as well as academic attainment of pupils in national exams at ages 11 and 16, attendance at school, and the dates and durations of any exclusions from school. By merging this into the behaviour dataset I can establish the full roll of a school on a given date, including pupils who were at these schools but who were not involved in any behaviour incidents. Thus I am able to include the whole population of these schools and avoid selection bias.

I supplement these datasets with a dataset of psychological characteristics collected as part of the UK Resilience Programme Evaluation This dataset contains 20,000 observations of 6,000 pupils over three years. Here I will use measures of depressive symptoms, anxiety symptoms, self-reported and teacher-reported behaviour, and some of the information on family characteristics (see Chapter 1 for details, 'Measures and data'). Pupils were surveyed up to five times over a period of three years; when a measure is available for more than one period I use the mean score across time periods. I also collected information on academic attainment, and developed ability scores (also called 'IQ scores') at entry to secondary school as described in Chapter 1. Pupils at the four schools for which I have behaviour will only be in the UKRP dataset if they were in the relevant cohort to be part of the UKRP evaluation. This means that 1,308 pupils who appear in the behaviour databases also appear in the psychological datasets. Pupils at these four schools who do not appear in the UKRP dataset do so primarily because they are in different cohorts to the ones surveyed for the UKRP evaluation.

Behaviour management

Here I describe typical school behaviour management systems, based on the descriptions of incidents in the databases, the outcomes listed, and discussions with school staff. Serious incidents such as those involving violence are dealt with quickly, with pupils sent home or separated from their ordinary classes for a time to cool off (they are still expected to complete their schoolwork when away from the class). These incidents are certain to be recorded, but this is not necessarily the case for other types of incident. If pupils are being mildly disruptive in class but stop when the teacher warns them, the incident is unlikely to be recorded. If the disruption persists and is serious enough, members of the senior leadership team (SLT) within the school will be on call and can be called out to discuss the issue with the pupil(s) involved. SLT call outs are always recorded. If a pupil is defiant towards senior staff or is otherwise still causing a problem, they will be taken to another classroom to work in isolation from the rest of the class. This is always recorded. Other penalties for poor behaviour include detentions at break times or after school; informing parents; close monitoring of behaviour; and more rarely, community service, and fixed-term or permanent exclusions from school for the most serious offences.[105]

There are two important consequences of this approach. First, there is likely to be underreporting of minor incidents. Second, the disruption experienced by other pupils in a class with a disruptive pupil is reduced when the perpetrators are removed from the class. This could reduce any potential negative impact on other (largely non-disruptive) pupils' behaviour, academic attainment or emotional wellbeing. It is also not clear whether teachers use the same reporting thresholds in breaktimes and before and after school as they would during lessons: disruption is the most commonly listed issue, and this is less likely to be a problem outside of lessons. Thus it is likely that teachers report behaviour that is disruptive or problematic *in context*, meaning that the same behaviour at a different time of day might not be a problem and might not be reported. This is not necessarily a weakness of these databases; appropriate behaviour is defined by its context, and since talking is not a problem in the playground at breaktime but may well be in lesson time, the former is not recorded as an incident. There may also be differential reporting by teacher: teachers may be better or worse at dealing with incidents, or may fear being considered incompetent if they use the SLT call out too often. There may be underreporting of complex incidents: when an incident escalates it may involve several different behavioural problems over a period of time, but will be reported as a single event.

Are these strategies typical of schools in England? The Teacher Voice Omnibus survey (NFER, 2012) finds that 75% of secondary school teachers report removing misbehaving pupils from the classroom at least sometimes, with a further 18% reporting using this often. 45% reported sending misbehaving pupils to the headteacher or senior staff sometimes, with 3% reporting using this often, and 62% said they used detention after school sometimes, with 25% using this often. 48% said they sometimes gave parents feedback about their children's behaviour (both good and bad), while 51% said they used this often. Likewise, a teacher union's survey found that 82% of teachers said that removal from a lesson was used in their schools; 56% sent pupils to the headteacher or senior staff; 82% informed parents; 80% used warnings; and 68% used detention (ATL, 2012). Thus these behaviour strategies do appear to be typical of many secondary schools in England.

[105] Schools may also inform the police if a criminal offence is committed, e.g. theft.

Schools

Four schools provided detailed information on behaviour incidents as part of the UK Resilience Programme evaluation. For my results to have external validity it is important that the sample is nationally representative. Schools 1-3 are located within the same town in Hertfordshire, while School 4 is in South Tyneside. These schools are broadly typical of the LAs from which they are drawn – see part of Chapter 1 for demographic data on these regions (Table 1.1 and Table 1.9). Table 4.1 presents a comparison of the characteristics of pupils at these schools with the average for state-funded secondary schools in England.[106] Figures reported for the schools are for pupils for whom I have data, so these figures reflect the cohorts included in the data and may not be representative of the whole school (and particularly in small schools there may be variation between cohorts). Moreover, included in these numbers are all pupils who are ever in the behaviour databases, and since disadvantaged pupils move school more often these figures may overestimate the percentage of poorly behaved and disadvantaged pupils in a school at any one time (Machin, Telhaj & Wilson, 2006). However, they are broadly representative of the intake of each school. All four schools have a roughly equal number of boys and girls. Across the four schools, 20% of pupils have special educational needs (SEN), in line with the average for England (21%). However, there is notable variation by school: School 1 has 43% of pupils registered as having SEN, while School 3 has only 16%. The percentage of pupils eligible for free school meals (FSM; an indicator of deprivation) in the previous six years is higher in these schools (30%) than in England as a whole (27%), but again, this varies substantially by school: School 3 has only 14% of pupils eligible, while School 4 has 40%.[107] The Income Deprivation Affecting Children Index (IDACI) score gives the proportion of children in an area (lower layer super output area) who are deprived, so a score of 0.246 means that 24.6% of children under 16 are living in families which are income deprived. The average for England is 0.22, while the average value for the home postcodes of children at these four schools is 0.29, suggesting that pupils at these schools live in areas with a higher than average proportion of deprived children. Key Stage 2 exams are nationally set and marked exams taken at the end of primary school at the age of 11. These exams

[106] Approximately 91% of pupils in England aged 11-15 attended mainstream state schools in 2009 (DCSF, 2009b).

[107] I use this measure of FSM eligibility rather than eligibility at a single point in time because my dataset covers six academic years, and FSM eligibility during a single year at secondary school may not accurately reflect disadvantage (DfE, 2010c).

therefore offer an indication of the academic attainment and ability of pupils entering secondary school. In England as a whole in 2006, 79% of pupils attained level 4 or better in English, 76% achieved this level in maths, and 87% achieved this level in science. Attainment for pupils entering these four schools was similar: 77% achieved this level in English, 75% in maths and 88% in science, so their average attainment was very slightly lower than the national average. Pupils at these four schools were more likely to be white (88%) than the national average (82%), although this reflects the ethnic composition of the two regions they are drawn from, which have very small non-white populations. The median size of a state secondary school in 2009 was 950 pupils, while these schools had between 700 and 1120 pupils, with three of four schools at or below the median number for England.

For my results to be informative it is particularly important that the behaviour seen in these schools is typical, and one piece of evidence for this is exclusion rates, as schools face a strict code of practice governing how and why they can exclude pupils. These four schools have permanent exclusion rates below the national average (although there is variation between schools), suggesting that very serious behaviour problems are at least as rare in these schools as in all schools. The number of fixed-term exclusions per pupil is very similar to the national average, although the number of sessions of fixed-term exclusions per pupil appears higher, suggesting that just as many pupils are excluded but for longer periods. Thus, taken together, pupils at these four schools are broadly typical of state school pupils in England as a whole: they are slightly more likely to come from poor backgrounds or live in deprived areas; have similar levels of prior attainment and SEN; and are more likely to be from white ethnic backgrounds. Further information on these schools can be obtained from school inspection reports by the Office for Standards in Education (Ofsted).[108] All four schools were inspected between 2008 and 2010, with three schools obtaining ‘satisfactory’ grades for the schools overall, and being rated as average in terms of pupil behaviour. The fourth school was graded as ‘good’ overall with pupil behaviour deemed better than average. Thus it seems that these four schools are reasonably representative of comprehensive schools in England as a whole, with similar pupil characteristics, similar exclusion rates, and with pupil behaviour graded as average or better by Ofsted. This is important, because if these schools are typical my results may have some external validity.

[108] See Ofsted (2012a) for information on the school inspections framework.

Behaviour incidents and pupil characteristics

Types of behaviour incident

Table 4.2 presents the classification of all recorded behaviour incidents. Of 45,493 incidents, 8,494 are missing classification (19%). Based on the outcomes of these incidents (e.g. no action taken; detention; sent home etc.) they do not appear to be any more or less serious than the average for all other incidents. There is also no relationship between missing data and the time period. It may therefore be reasonable to assume that categorisation is missing at random in most cases, although when I present data by incident type I include 'missing classification' as a separate category for the purposes of comparison. The most commonly reported incident types are disruption and defiance: 27% of incidents are listed as disruption, with a further 17% classified as defiance. From descriptions, it appears that defiance is often disruption, and disruption usually involves defiance as well (as do many other incidents). Defiance may arise because pupils are doing something wrong and refuse to stop doing it, so it could be seen as a catch-all term. Lateness is common (9% of incidents), as is failing to bring equipment (7%). Verbal abuse, rudeness and insolence form the next largest category (4.5% of incidents), after which there are a number of smaller categories. In line with reports by Ofsted (2005), the Department for Education (DfE, 2012a), and the Association of Teachers and Lecturers (ATL, 2012), incidents involving violence are relatively rare: combined assault, aggression and fighting come to only 3% of the total number of incidents, with threatening behaviour a further 0.1% and all bullying (which could also be verbal) 1% of incidents. Taking these three categories together gives 1,883 incidents over 1,484,111 pupil-days: 0.0013 incidents per pupil per day, or 1.3 incidents per day in a school of 1,000 pupils. This suggests that behaviour in these schools is not generally violent, but that disruption and failure to follow rules are common.

Pupil characteristics

The first columns of Table 4.3 give details of all pupils in the dataset. There are 3,284 pupils in total. There is an average of 13.9 incidents per pupil during the time they are covered in the dataset, over an average period of 451.8 school days. Thus we have information on behaviour for an average of 2.38 academic years per pupil, with an average of 0.036 incidents per pupil per day. The full five years of compulsory secondary schooling covers approximately 950 school days (190 days per year over five years), so it is clear that the dataset does not cover the whole of secondary school for the majority of pupils in the dataset. In September 2009 pupils were aged between 11 and 20, with a mean age of 14.6, although pupils will only be included in the dataset between the ages of 11 and 16 as education beyond this age is not compulsory.

Concentration of incidents

Table 4.4 illustrates the concentration or skewness of behaviour incidents by pupil: a small number of pupils are responsible for the majority of incidents, while nearly one third of pupils are involved in no incidents at all. The first column shows that 10% of all behaviour incidents are accounted for by only 0.94% of pupils, 50% of incidents are due to 7.98% of pupils, and so on. This is also illustrated in Figure 4.1 and Figure 4.2 which represent the cumulative frequency of incidents per pupil and incidents per pupil per day respectively.[109] This concentration is seen for all incidents, but is also true of each broad category of incidents taken separately (Table 4.4, later columns). However, the degree of concentration differs by type of incident, with just under 40% of pupils accounting for all incidents of disruption, defiance and miscellaneous minor incidents, while lateness and truancy, dangerous behaviour and damage, and assault and fighting are each accounted for by around 20% of pupils.[110] Part of this difference may be due to the frequency of incidents in each category, but this does not account for all of the difference: there are more incidents involving lateness and truancy than are in the 'other (minor)' category, but half as many pupils account for all incidents of lateness and truancy. This could mean that behaviour incidents are heterogeneous, and that pupils have different preferences for the types of misbehaviour they are involved in (or alternatively, that they face different costs to behaving well in different ways). Relatively few find it difficult to avoid fighting, assault, and dangerous behaviour, but a larger number have a tendency to be disruptive or to infringe uniform regulations. This in turn may be partly a response to the school environment: if teachers are more likely to tolerate lateness than violence, incidents of lateness are likely to be higher in equilibrium.

The end columns of Table 4.3 present information on the characteristics of the 10% of pupils with the most incidents, who are responsible for more than half of total

[109] Incidents per pupil per day takes into account the amount of time a pupil appears in the datasets, although this does not change the overall shape of the graph.

[110] Incidents have been placed into broad categories based on the more detailed category in Table 4.2. The category of 'Other (minor)', includes inadequate work or homework, no equipment, incorrect uniform, and other minor incidents which I have a categorisation for. 'Missing category' covers the 8,494 incidents for which I do not have any category information, and is included for comparison.

incidents.[111] The final column contains the p-values of a regression of each characteristic on a dummy for membership of the top 10% and school dummies, clustered by class membership. The pupils with the most incidents have significantly higher rates of incidents in all categories, with an overall mean of 0.18 incidents per day, or just under one incident per week. These pupils are significantly more likely to be male, have SEN, be eligible for FSM, come from a white background, and live in a more deprived neighbourhood, than the 90% of pupils with the least incidents. They are also older, and have significantly lower test scores, lower developed ability scores, higher (worse) depression scores, higher (worse) self-reported and teacher-reported behaviour scores, and are significantly less likely to report living with both of their parents. However, there is no significant difference between the two groups in terms of the number of school days for which I have data; the calendar quarter in which they were born[112]; the fraction from black, other Asian or mixed backgrounds, and their anxiety symptom scores. These figures suggest that pupils with poor behaviour have different characteristics on average, but that none of the characteristics is in any way deterministic of behaviour: the 10% with the most incidents includes pupils with a variety of different demographic backgrounds.

Predicting behaviour incidents

An alternative way of looking at the relationship between pupil characteristics and behaviour is presented in Table 4.5. This uses negative binomial regressions to predict the number of incidents per pupil, reporting coefficients, standard errors and incidence rate ratios (IRRs) for each regressor.[113] As before, we cannot interpret these associations as being *causally* predictive of incidents, but they may be informative as to which pupil characteristics are most strongly associated with poor behaviour. The outcome variable here is an instance of count data, but I cannot use Poisson regression because the mean is substantially greater than the variance (Greene, 2003).[114] Column 1 includes only

---

[111] As in Table 4.4, the top 10% is calculated within school to account for reporting threshold differences between schools. However, failing to take school into account does not substantially change the results. The top 10% is also based on the rate of behaviour incidents per day, not the total number of incidents, although the correlation between the two measures is 0.96 and using the total number of incidents does not change the results.

[112] A pupil's date of birth determines which cohort they belong to, with September-born children the oldest in their cohort and August-born children the youngest. Moving cohorts is rare; in this dataset 99% of pupils are in the cohort predicted by their date of birth.

[113] Note that 'prediction' does not mean causation; I use the term to suggest that in my data certain characteristics are strongly associated with outcomes. This is the way the term is often used in psychology when one wishes to assert association without making statements about causality (see e.g. Duckworth & Seligman, 2005).

[114] I have run all of these specifications using OLS and Poisson regressions, and the results are very

basic demographic characteristics, indicating that being male, FSM eligible, having SEN, coming from a white ethnic background, and living in a more deprived area are all significantly associated with having a higher number of behaviour incidents. Having SEN is the most strongly associated with more behaviour incidents, with an IRR of 2.2, but this variable may be tautological: many pupils with problematic behaviour are categorised as having Emotional or Behavioural Disorder (EBD) Special Educational Needs, and in 2012 EBD SEN was the commonest type of SEN in state maintained secondary schools (DfE, 2012f). Thus the dummy for SEN will identify pupils with EBD SEN as well as pupils with low attainment, specific learning disabilities (e.g. dyslexia), or physical impairments. Column 2 adds in the Key Stage 2 score (the age 11 national test score, averaged over English, maths and science): this is a significant predictor of incidents, and also decreases the coefficient on 'SEN', suggesting that both low attainment or ability and the identification of pupils with EBD SEN are associated with worse behaviour. When the average test score is included the coefficients on being born in spring (March-May) or summer (June-August) become significant, with pupils born in the second half of the academic year having 0.76-0.81 as many incidents as those born in autumn (September-November; these will be the oldest pupils within each cohort). The remaining columns supplement the basic demographic data from the national databases with psychological and family data collected through questionnaires for the UK Resilience Programme evaluation. This greatly reduces the sample size. Column 3 reruns the regression from column 2 but on this reduced sample, showing a similar pattern of relationships, the only major difference being that the coefficient on winter born is now larger and statistically significant. Column 4 adds in a number of variables: a dummy for EBD SEN; measures of symptoms of depression and anxiety; self-reported and teacher-reported behaviour measures; pupils' reports of who they live with; and a developed ability score from a test taken at the start of secondary school. EBD SEN is a highly significant predictor of incidents, and its inclusion reverses the sign on the 'SEN' variable, suggesting that only insofar as SEN pupils have EBD SEN, low developed ability or low attainment do they have more incidents; all else being equal they have significantly fewer incidents than non-SEN pupils. A higher (worse) depression score is associated with more incidents, as are higher (worse) self- and teacher-reported behaviour scores, as we might expect. However, higher (worse) anxiety scores are negatively associated with behaviour incidents. This could reflect the

largely unchanged.

difference between internalising and externalising psychological traits, with behaviour associated with the latter: a small amount of anxiety might prevent pupils from misbehaving, perhaps through increased inhibition. Pupils who report living with neither or only one of their parents have significantly more behaviour incidents than those who report living with both parents. However, this variable is pupil-reported and probably contains a lot of measurement error. Including all these additional variables causes the coefficients on FSM, white ethnic background, the deprivation score of the neighbourhood, and the mean age 11 test score to become insignificant, suggesting that individual psychological characteristics are more important in predicting behaviour than general demographic categories. Column 5 runs the same regression as column 4, but excludes the 'tautological' variables which directly measure poor behaviour (such as EBD SEN and teacher- and self-reported behaviour scores). This is to understand which non-behaviour-related characteristics of pupils are associated with poor behaviour. Here we find that all included variables except neighbourhood deprivation are to some extent significant. Boys have 2.5 as many incidents as girls; while increasing the depression symptom score by one point is associated with a 13% increase in the daily incident rate, a sizeable impact given the scale.[115] There is a consistent negative association with being younger within the cohort.[116] Interestingly, of the two variables tracking deprivation – FSM and IDACI score – only FSM is still significant once more detailed characteristics are included, and even this is similar in magnitude to the dummies on being born in winter, spring, and summer; white ethnicity; and of not living with both parents. The coefficient on FSM is also significantly smaller than the coefficient on being a boy (p=0.0002). These results suggest that although demographic variables are importantly associated with behaviour, it is more idiosyncratic variables such as psychological characteristics which are just as good or better predictors of behaviour incidents.

It is worth looking more closely at the predictive power of the teacher-reported and pupil-reported behaviour scores in column 4. The coefficient on the teacher-

[115] It is difficult to compare the magnitudes of the coefficients here because 'boy' is a dummy while the depression score is a scale. In linear regressions including the same variables the depression score has a significantly larger partial correlation coefficient than the dummy on FSM, p-value of a test of equality of the coefficients=0.01.

[116] We might expect younger pupils to have worse behaviour because they may be less emotionally developed. For instance, Crawford, Dearden and Greaves (2011) find that August-born children have worse teacher-reported behaviour scores. However, this is only the case up to the age of 9, and they also find that August-born children aged 16-19 are significantly less likely to say that they use cannabis or drink alcohol than September-born children. Using the full UKRP evaluation dataset I also find that August-born children are significantly less likely to say they bully other children and significantly more likely to say that they have been bullied than September-born children.

reported behaviour score is significantly larger than that on the pupil-reported score ($p=0.0048$), however, the distributions of the two scores differ (the teacher score is more skewed, with many zeros). In linear regressions including the same variables the teacher score has the largest partial correlation coefficient of any of the regressors ($\rho=0.26$), which is also significantly larger than that on the self-reported behaviour score ($\rho=0.07$; p-value of test of equality = 0.035). Thus it would appear that the teacher-reported behaviour score is more strongly related to the number of incidents a pupil is involved in than the pupil-reported score, suggesting that teachers' reports are a better measure of poor behaviour in schools. This is in line with the literature on the reliability of self-report measures (Dunning et al., 2004).[117] It is reassuring that the behaviour scores are strongly related to the number of incidents a pupil is involved in, but even controlling for these scores there is still a large role for other characteristics in predicting incidents. This suggests that even teacher reports are a noisy measure of pupils' actual behaviour.

It may be worth examining the characteristics which predict incidents separately by gender, since gender appears to be such an important predictor itself. Table 4.6 presents the same regressions as Table 4.5, but includes only boys; and Table 4.7 does the same for girls. The coefficients are in fact broadly similar in both cases, always in the same direction and of similar magnitude. However, it does seem that being from a white ethnic background is a stronger predictor of incidents for girls than for boys; that the impact of season of birth is a stronger predictor for girls; and that the association with the depression and anxiety scores is stronger for girls.

Pupils who are frequently absent will have fewer behaviour incidents (all else being equal), so it is possible that associations between certain pupil characteristics and better behaviour are mainly due to higher absence rates rather than better behaviour when actually at school. Table A4.1 presents information on the mean absence rates by pupil characteristics for this sample.[118] Here we can see that the characteristics associated with more behaviour incidents are also associated with higher absence rates, meaning that the estimates discussed above are likely to be an *understatement* of the true propensity to misbehave for these pupils. This is reflected in a correlation

[117] Interestingly, I cannot reject that the coefficient on the self-reported behaviour score is equal to the (negative) anxiety score coefficient in both the linear and the negative binomial regressions, whether or not we control for the teacher-reported score. Thus self-reported behaviour appears to be a relatively poor predictor of actual behaviour.

[118] The absence rates correspond to the academic years for which I have behaviour incident data for each pupil. I do not have absence rates available for all pupils in this sample.

coefficient of 0.13 between the incident rate and the absence rate. The only exceptions to this are for the quarter of birth, for which absence rates do not seem to vary much but behaviour incident rates do, and for gender: boys have more behaviour incidents but slightly lower absence rates (although this is not a statistically significant difference in this sample). Table A4.2 presents the same regressions as Table 4.5, but includes the absence rate as a control variable. This makes very little difference to the estimates of the strength of the relationship between the demographic characteristics and the behaviour incident rate, with only the coefficient on FSM being somewhat attenuated in certain specifications. An alternative way of accounting for absence is to use the absence rate to adjust the exposure variable in the negative binomial regressions to show the number of days attended per academic year, rather than the number of days on roll. Doing this produces almost identical results to those presented in Table 4.5 (results not shown). These robustness checks suggest that absenteeism is not driving the patterns observed above.

There is a sizeable effect of absolute age or yeargroup on behaviour incident rates, but because the negative binomial regressions reported above were constructed using an incident count for all academic years a pupil appears in the dataset I could not take this into account.[119] Figure 4.3 shows the mean number of behaviour incidents per pupil per day by yeargroup: there is a strong association of yeargroup with the rate of behaviour incidents. Interestingly, incident rates do not increase monotonically with age: they increase to Year 10, then decline significantly in Year 11, when pupils are aged 15-16. The change in the rate of behaviour incidents by yeargroup is further evidence for the change in mean-level noncognitive traits through the life course, especially since the context pupils are in largely does not change here.[120] It is not clear why older students should be more likely to misbehave. Steinberg (2007) suggests that adolescents are more risk-seeking than both younger children and older adults because puberty impels them towards thrill seeking, while their ability to control these impulses is still not fully developed. The drop in incidents in Year 11 may be due to the increasing costs of misbehaviour. Students sit national school leaving exams at the end of Year 11, making

[119] I included season of birth dummies to account for relative age within a cohort, but since most pupils appear in this dataset over more than one academic year their yeargroup will change.

[120] Note that if the reduction in incidents in Year 11 were largely due to teachers being less tolerant of poor behaviour, we would expect to see similar behaviour rates at the start of Year 11 as in Year 10, with a decline through the year as pupils update their expectations about the consequences of misbehaving. This is not the pattern we see here.

this year critical for attainment and future prospects.[121] If students have a high discount rate, then the consequences of poor behaviour will be more apparent to them in the last year of compulsory education. The predilection for misbehaviour may also change: from Year 10 students have more choice about the subjects they study, which may make the average lesson more pleasant and misbehaviour less attractive.

Dimensions of behaviour incidents

Predicting behaviour by type of incident

Table 4.8 runs the same specification as column 2 of Table 4.5, but with counts of the behaviour incidents by type as the outcome variable. Overall, the impact of these basic demographic variables is fairly consistent across the different categories of incident, but there are some differences. For instance, the coefficient on ‘boy’ is similar (IRRs 1.4-2.1) for all types of incident except assault and fighting and dangerous behaviour (boys have 3.5 and 2.8 times more of these incidents than girls, respectively) and lateness and truancy (girls are just as likely to be late or truanting as boys). The impact of FSM is very consistent across categories, with IRRs in the range of 1.69-1.96, with the exception of incidents missing categorisation, which has a lower IRR of 1.39. It is not clear why this should be the case – the missing category is included in case missing the description is related to the nature of incident, rather than being arbitrarily missing data, but it is possible that incidents of this type represent a different dimension of behaviour. SEN is again very significant for all incident types, but the IRR is notably smaller for minor incidents. Being of white ethnicity is not significantly associated with the number of assaults or fights, although it is associated with all other categories. The neighbourhood deprivation score (IDACI) has a consistent impact with the exception of dangerous behaviour and incidents missing categorisation, while it is a particularly strong predictor of verbal offences and of minor incidents. The association of season of birth is not consistent across categories, although being born in summer does seem to be associated with a reduction in most types of incidents. Test scores at age 11 also have very consistent associations across incident types, with only ‘other minor’ types of incident having a slightly higher IRR (suggesting that higher academic attainment is not so strongly associated with being involved in minor incidents such as not having equipment or inadequate work). Overall, the characteristics available here predict different categories of incident in a remarkably consistent way, although the strongest

[121] I have anecdotal reports from headteachers to this effect.

predictors vary by category, and the coefficients on the demographic characteristics are generally weaker for predicting minor incidents.[122]

I can compare these results with those of Segal (2008), who examines the predictors of boys' misbehaviour using teacher reports across five categories: absenteeism, disruptiveness, homework noncompletion, inattentiveness, and lateness. Absences are generally not reported in my behaviour data (these are recorded separately), but the other categories Segal uses correspond roughly to incidents covered in my data. She finds a similar relationship between parental income and behaviour, with pupils from richer or more educated households less likely to be late and disruptive and more likely to complete their homework. Similarly, I find that FSM eligibility is a consistently strong predictor of behaviour incidents. Segal finds that students who do not live with both parents have significantly worse behaviour in every category, as do I. We find differing impacts of ethnicity: Segal finds that black students are more likely to have some form of problematic behaviour, while I find that pupils from white backgrounds are significantly more likely to have incidents across most categories. In England, pupils from most ethnic minority backgrounds perform better at school than pupils from white British backgrounds (DfE, 2013e), so we might expect their behaviour to be better too.

Segal uses a much richer set of controls to predict poor behaviour but finds low values of the r-squared (0.35 or lower) when using linear regressions, even after including teacher as well as pupil characteristics. She concludes that individual-specific traits explain a substantial amount of variation in behaviour, although demographic characteristics are also important. Likewise, when I run the same specifications reported in Table 4.5 using OLS rather than negative binomial regressions, the basic specification (as in column 2 of Table 4.5) produces an r-squared of 0.055, rising to 0.059 when all the variables except the ones directly measuring behaviour are included (similar to column 5 of Table 4.5). The r-squareds from these regressions are reported in Table 4.9.[123] Here we see that the characteristics I include in the regressions explain very little of the variance in the outcome measure, with the highest r-squared for assault and

[122] This may be partly because of the different frequency of incidents in different categories: there are many incidents of disruption, but few of assault, so the same IRR on the same demographic characteristic may represent a different shift across the distribution for different outcomes.

[123] Note that I exclude school fixed effects here because they contribute substantially to baseline r-squared due to differential reporting thresholds. Also, the outcomes in these regressions are the number of incidents per pupil per day.

fighting, defiance and verbal offence.[124] This suggests that the low r-squared Segal finds are not solely due to the lack of variation in her data: in the National Educational Longitudinal Survey data she uses, the outcome is a binary variable equal to 1 if a teacher reports that a student misbehaved on that dimension. About half of the boys in the sample misbehave in at least one category, and the proportion misbehaving for each dimension is never higher than 0.34 (as reported above, in my sample about two-thirds of pupils have at least one incident). Since my data includes the number of behaviour incidents per pupil, this should allow for greater discrimination in the outcome variable. Thus my results are in line with hers: background characteristics are strong predictors of behaviour incidents in most categories, but explain little of the overall variance in behaviour.[125]

Relationships between incident categories

Table 4.10 displays pairwise Spearman rank correlation coefficients for the incident rates of different categories of behaviour incidents. For all categories, a pupil's rank is significantly positively correlated with their rank in all other categories, suggesting that pupils who misbehave in one way also misbehave in other ways. However, there are differences in the strength of association between categories. Disruption shows a moderate to strong correlation with all other categories, suggesting that pupils with incidents of any type are also likely to be disruptive. Assault and fighting and dangerous behaviour are both only weakly related to lateness and other minor incidents. Since one third of pupils have zero incidents in all categories, Table 4.11 presents the same rank correlations with only pupils who have been involved in at least one incident in any category. This attenuates the coefficients but leaves their relative values largely unchanged. Table 4.12 displays results of negative binomial regressions of each incident category on all others. The column heading gives the regressand. Looking along each row we can interpret the coefficients as the impact of more incidents in that category on the total incidents in the regressand category, controlling for all other types of incident. For instance, looking along the row for 'disruption', we can see that incidents of disruption explain little or none of the

[124] Note that if a characteristic 'explains (much of) the variation' in the outcome, this is purely observational and means that the characteristic is statistically associated with the variation in the outcome; this is not meant to imply that the association is causal.

[125] The r-squared from OLS regressions may not be the best way of evaluating contributions to variance here, given the extremely skewed data. Moreover, the background characteristics I have available for pupils are limited in scope, so it is also possible that I am unable to predict incidents simply because I have the wrong variables.

variation in incidents of assault, defiance, dangerous behaviour, and verbal offence. However, each disruptive incident is associated with a 0.6% increase in minor incidents, and is associated with a 1.5% decrease in incidents of lateness and truancy, once all other incident types have been controlled for. This suggests that disruptive incidents are not good predictors of pupils' propensity to misbehave in other ways. However, we can also look at regressions where disruption is the outcome (looking down the column with 'disruption' at its head): here we see that every other category is a significantly positive predictor of incidents of disruption, with the exception of lateness and truancy, which is negatively associated with it. This suggests that pupils who are defiant or who have a number of incidents of assault, verbal offence etc. are also very likely to be disruptive; but the reverse is not true – disruptive pupils are not necessarily prone to other types of misbehaviour. Similarly, having more incidents of assault or fighting is very strongly related to more incidents in all categories except lateness and minor incidents, but only defiance, dangerous behaviour and verbal offence have any predictive power for incidents of assault. I would interpret this as evidence of two broad groups of pupils with behaviour incidents: those with the more serious incidents such as assault or fighting, who also have incidents in many other categories too; and those who only have incidents in the minor categories such as lateness or minor incidents such as uniform issues.

Table 4.13 presents the factor loadings and unique variances for a principal factor analysis of incident types. This produces a two-factor model, with the first factor having strong positive loadings for each item (incident category), which we could interpret as the general propensity to misbehave. The second factor has positive loadings on assault, defiance, dangerous behaviour, and verbal offence, and negative loadings on disruption, lateness and minor incidents. We could interpret this second factor as an indicator of aggressive or oppositional behaviour. We could speculatively link these factors into the Big Five personality traits: the first factor seems indicative of low conscientiousness, associated with rule-following and self-discipline. Segal (2008) suggests that this may be the factor most strongly associated with classroom behaviour. The second factor suggests low agreeableness (uncooperative and antagonistic behaviour) but also high emotional instability (emotionally reactivity).[126] Again I would interpret this as evidence for two groups of misbehaving pupils: all pupils who misbehave have low

[126] For a summary and comparison of 11 inventories measuring the Big Five see Grucza and Goldberg (2007).

conscientiousness and at least some non-confrontational incidents, but only those who are also low in agreeableness and emotional stability are also antagonistic. Most incident categories have relatively high communality, but assault or fighting has little of the variation explained by the model. This suggests that violent incidents are poorly predicted by poor behaviour in general, but are in some way exceptional (or the pupils who commit them are exceptional), even within the group of pupils with confrontational incidents.[127]

Persistence of behaviour

Persistence of behaviour through time

Is behaviour (good and bad) persistent through time, or does it appear to be generated by temporary shocks? Figure 4.4 presents information on the persistence of behaviour through time. I divide the sample into deciles of behaviour incidents per day in the first term a pupil appears in the dataset, then track their mean rate of incidents per day across five subsequent terms (just under two academic years).[128] Because of the skewness of behavioural incidents, more than 60% of pupils fall into the first decile during the first term they appear in the dataset, with no incidents or very few, thus there are only five categories available: the lowest decile, then deciles 7-10.[129] There is substantial mean reversion, with the incident rates of the top two deciles falling sharply over a few terms, and the incident rate of the first decile rising. This is despite the tendency for behaviour to worsen with age. However, the *ordering* of the groups is maintained throughout the five terms, so on average pupils whose behaviour put them in the 10th decile in term 1 still have many more incidents per day five terms later than those in lower deciles at the start. This suggests substantial persistence in behaviour

[127] Running the specifications reported in Table 4.12 using OLS – predicting incident counts in one category using incident counts in all other categories – I find that the r-squared when incidents of assault and fighting is the outcome is 0.33 - lower than for any other incident category (e.g. 0.61 when disruption is the outcome), again indicating exceptionalism. This is robust to including pupil characteristics. We also see demographic and character differences between pupils with different types of incidents. The depression scores of pupils with at least one violent incident are significantly greater (worse) on average than those of pupils with other oppositional incidents, which in turn are significantly greater than those with only minor incidents. The same pattern applies to the likelihood of not living with both parents.

[128] An academic year consists of three terms. A term generally lasts between 48 and 75 school days, with a one week holiday in the middle separating it into two half terms.

[129] The decile is calculated within cohort and school to allow for different reporting thresholds by school and for the differing prevalence of incidents by age and term. Because of the concentration of behaviour incidents deciles 7 and 8 have very similar incident rates.

over five terms, both for pupils with high levels of behaviour incidents and pupils with very few incidents.[130]

I can also use the number of incidents during the first half term a pupil appears in the dataset to predict incidents in later half terms. Table 4.14 presents coefficients and IRRs for negative binomial regressions, showing that 18 half terms later (9 terms, or nearly 3 years later) the number of incidents in the first half term is still a significant predictor of behaviour, even when controlling for demographic characteristics. Its predictive value declines through time: this is clearly shown in Figure 4.5, which plots the IRRs for each half term's regression. Again this suggests that behaviour is persistent through time. We should also look at the correlation in behaviour rank through time. Table 4.15 displays Spearman rank correlation coefficients by half term: these show a moderate positive relationship up to about two years, after which the association is positive and significant but weaker. This is partly due to short panels in several schools which result in a substantial reduction in sample size after two years, but even when we condition on pupils who are in the data for the full 18 half terms we see that rank is persistent but declining in importance through time.

We might be particularly interested in the change in rank over time for pupils with the most behaviour incidents. Table 4.16 shows the transition between deciles of pupils who are ranked in the 10th decile for behaviour during their first half term in the dataset. During the second half term, of the pupils who are still in the dataset, 54.7% are still in the 10th decile; 18.8% have dropped to the 9th decile; and 26.5% are in the 8th decile or lower. Only 0.9% of the original 239 are missing data at this point. Missing data is important: pupils with very poor behaviour may be excluded from school, and may be more likely to leave for other reasons, so failing to take this into account could understate the persistence of behaviour. However, pupils may also disappear from the dataset because of censoring (the panel which the school provided stops), or because they reach the end of compulsory schooling. I therefore provide three explanations for missing data: censored data, leaving school and finishing school - only the second kind of missing data is likely to bias results. We see a steady decline in the percentage of pupils originally in the 10th decile for behaviour who are still there, with only 40.6% of pupils still there 7 half terms (one academic year) later.[131] This is important because the

[130] The graph includes only pupils for whom I have at least 5 terms of data, and this results in excluding a higher proportion of pupils with poor behaviour. This is a particular problem when examining periods of more than two years.

[131] Note that this does not take into account pupils who move down a decile then move up again – all the

top 10% of pupils is responsible for more than half of incidents, so moving out of the top 10% suggests a large absolute improvement in behaviour, while the difference in incident rates is smaller between subsequent deciles. However, note that 20% are now missing data, and this percentage increases with each half term, making it difficult to draw firm conclusions. Having about 40% of the same pupils in the top behaviour category after one year does again suggest persistence in behaviour in the short run, though this is down to 21% of pupils remaining there three years later.

Persistence of behaviour with different teachers

Another question is whether pupils who misbehave do so in only a few lessons or with only some teachers, or whether their behaviour is consistent across different contexts.[132] If poor behaviour is highly context specific, we would expect few teachers to be involved even when pupils have a high incident count. I have information on the teachers involved in behaviour incidents for one school only. More than one teacher can be listed with each incident, but the first teacher to be listed is usually the one whose lesson the incident took place in, or, if outside of lessons, the first teacher to notice and confront the misbehaviour. Using this information on the first listed teacher I can look at the total number of incidents a pupil is involved in, and the number of different teachers who were involved in these incidents - this is shown in Figure 4.6. We see that there is an approximately linear relationship between the number of incidents and the number of teachers involved, with on average two incidents per teacher throughout the distribution. Consistent with this, the linear correlation coefficient between the number of incidents per pupil and the number of different teachers involved is 0.96. This does not tell us much about the concentration of incidents per teacher: some teacher-pupil pairs may produce a disproportionately large number of incidents. But it does suggest that pupils who misbehave do so with a number of different teachers.[133]

Behaviour incidents by time of day and day of the week

Figure 4.7 presents the mean rate of behaviour incidents by day of the week, showing that there are significantly more incidents on Mondays than on later days, with the rate dropping sharply through the week. This difference is sizeable: in a school of

---

table records is the number originally in the top 10% who are also in the top 10% in that half term.

[132] At this level teachers are specialists in a particular subject. A pupil aged 11-14 will typically study up to 15 subjects per year, and will usually have a different teacher for each. Pupils aged 15-16 typically study at least 8 subjects with a different teacher for each.

[133] I only have data on teachers when a pupil is involved in an incident; this means that I cannot reliably compare behaviour rates by teacher as I do not know most pupils' timetables and teacher assignments.

1,000 pupils, there will be an average of 36 incidents in total on a typical Monday, but only 27 on Wednesdays, Thursdays and Fridays.

Figure 4.8 shows a similar impact of time of day: these schools have five lessons per day - the first lesson has few incidents per pupil, and the incident rate rises significantly through the day. In a school of 1,000 pupils we would expect 4.2 incidents to take place during the first lesson, with 7.4 taking place during lesson 5.[134] Putting this into a regression and controlling for pupil characteristics, school, the week of half term, and days before and after holidays, I find that these patterns persist. Thus it is clear that pupils do not behave uniformly well or badly, but that the context in which they find themselves is important and has an impact on the probability of incidents.[135] Table 4.18 and Table 4.19 show Spearman rank correlation coefficients between incident rates by day of the week and by time of day respectively. These are moderate to large and highly significant, suggesting that pupils' relative behaviour rank is maintained at different times and on different days, despite large changes in the probability of an incident.

Incident rates by subject

I have data on the subject being taught when an incident occurs for 23,460 incidents. In many cases I know the timetable frequency of each subject, so I can work out the relative incident rates by subject. These are reported in Table 4.20. Here we can see that some subjects have much higher incident rates than others. For instance, there are approximately 2.23 times as many incidents recorded during maths lessons than during English lessons. The lessons with the fewest incidents are performing arts, physical education (PE) and technology lessons, which may be more fun and less demanding than standard academic lessons. Maths and science have the highest incident rates.[136] This again suggests that behaviour is partly context specific: pupils may be more likely to misbehave when the activities they are engaged in are less appealing. Table 4.21 presents Spearman rank correlation coefficients between the behaviour rates by subject. The coefficients here are positive and moderate in size, but highly

---

[134] Incidents also take place outside of lessons, such as before and after school, registration, lunch and break times.

[135] In Chapter 5 I explore the robustness of these results; provide evidence that these patterns reflect differences in behaviour rather than being solely due to misreporting or selective absence; explore possible causes; and show that these patterns holds for all groups of pupils, both those with high incident rates and those with few incidents. For this reason I do not discuss other potential causes of these differences in incident rates here.

[136] This could also be the result of differential reporting by certain teachers, e.g. if teachers of classroom-based lessons find the same behaviour more disruptive than teachers of lessons which are not classroom-based, or if maths and science teachers are more inclined to record incidents than English teachers.

significant, suggesting substantial rank order consistency across subjects. The correlation coefficients are certainly smaller than those by time of day and day of the week, or on behaviour in adjacent terms, but they are about the same size as those between different incident types. This suggests lower rank order consistency in behaviour by subject than by scheduling, but about as much consistency across incident types. It is likely that the coefficients on subjects are biased downwards because of the large amount of missing information on subjects (which results in excess zeros in the data). However, the smaller coefficients might be partly driven by pupils having different tendencies to misbehave in different lessons, e.g. because pupils enjoy some subjects more than others, and preferences over subjects vary by pupil. This would generate different behaviour rankings by subject.

Discussion and conclusions

Behaviour at secondary school is an expression of noncognitive traits which are highly relevant to success in later life. Using data on behaviour incidents in schools, I examine the determinants, distribution and persistence of incidents across time and in different contexts. The use of data on specific incidents over several years represents a marked improvement in data quality over both teacher-reported and self-reported measures of behaviour at school.

The majority of incidents recorded are relatively minor, reflecting disruptive behaviour or failure to follow the rules, but not serious delinquency. However, 15% of incidents for which some categorisation is available involved verbal or physical confrontation. The majority of pupils misbehave at least once over a period of about two years: only 26% of boys and 40% of girls have no incidents at all. Segal (2008) finds that almost half of her sample of 8th grade boys are rated by their teachers as having some form of misbehaviour, but teachers might only be inclined to rate pupils as having tendencies to misbehave after repeated incidents, while my data can identify pupils with single incidents. The concentration of incidents is striking: less than 10% of pupils are responsible for more than half of incidents, suggesting that occasional rule-breaking is common but that problematic behaviour is confined to a small minority of pupils. There also appear to be two main dimensions of behaviour: factor analysis produces a two factor model tracking general rule breaking (all incidents), and aggressive or confrontational incidents. I suggest that these dimensions are associated with low conscientiousness (rule breaking) and low agreeableness/low emotional stability respectively. Most pupils who misbehave have only incidents of the former type, but pupils with confrontational incidents also have many minor incidents and account for the majority of incidents overall. This suggests that infrequent low-level poor behaviour may not identify serious deficits in noncognitive skills, particularly not in the domains that change naturally with age and which reflect a lack of conscientiousness (such as a tendency to lateness), but that assault or other confrontational incidents may suggest more deep-rooted psychological problems.

Demographic characteristics are strongly predictive of the number of incidents in all categories, and the coefficients on the characteristics do not vary much by incident type. Like Segal (2008), I find that pupils from richer households, and pupils living with both parents, have fewer incidents in all categories. But like her I find that demographic characteristics account for relatively little of the overall variation in incident rates. There is some variation here by incident type: the more confrontational incidents such as assault and fighting or verbal offence have higher r-squared figures than those for minor incidents (although these are still low). This is particularly interesting, as the more confrontational incidents have the lowest communality in the factor analysis: the counts of other incidents explain much less of the variation in assaults and fighting, for instance, than in disruptive incidents. This might suggest that many pupils from a range of social backgrounds infringe the rules, but that assaults and other confrontational incidents are more concentrated amongst pupils with specific demographic and personal characteristics (e.g. those with higher depression scores and not living with both parents).[137]

I also find that behaviour is fairly persistent over three years in the same context (same schools, same peers) in terms of both rank order and the absolute level of behaviour. This persistence, and the relationship of behaviour with background characteristics, both suggest that behaviour is not simply due to random shocks, but constitutes a stable personality trait. Similarly, Segal (2008) uses behaviour data collected two years apart and at different schools, finding that behaviour is persistent across this time. However, the transition matrices she presents for each category show that a substantial fraction of those identified of having (or not having) problematic behaviour in the first period are identified differently in the second period. Thus even if

[137] Note that I cannot infer causality from these associations, but they may provide evidence as to where to look for causal factors affecting behaviour, and may also provide a guide as to where to target resources in the absence of better data on behaviour problems.

on average behaviour in one period and in one context is strongly correlated with that in later periods and contexts, this is not deterministic. I observe a similar effect when looking at behaviour over three years: despite the similarity of context, behaviour scores in the first half term become a much weaker predictor of behaviour through time.

If pupils were equally likely to misbehave in different contexts based on their characteristics then this analysis would suggest that (all else being equal) schools with fewer boys; fewer FSM pupils; fewer pupils with low prior attainment or low ability; fewer white pupils; and fewer pupils from disrupted families, would experience fewer behavioural incidents. In particular, given the skewness of the distribution of incidents, it appears that removing the worst behaved 10% of pupils would more than halve the number of incidents. However, pupils' misbehaviour may depend on their peers, another aspect of context which I have not been able to take into account due to limited information on class assignment. It is possible that if the worst-behaved 10% of pupils were excluded, then another 10% of pupils would take their place by behaving worse at school. Alternatively, the presence of the worst 10% might actually increase the rate of incidents for the next decile: from the descriptions of behaviour incidents, it is clear that many involve more than one pupil, and that one pupil's actions may cause others to become involved. This is most clearly the case with fighting, but also with disruptive behaviour in lessons. This implies that by excluding a minority of pupils from a school, either through selective admissions or by permanently excluding pupils, schools could greatly improve the average level of behaviour.[138] It is therefore difficult to extrapolate the effects of removing poorly behaved pupils from the specific context in which behaviour is measured.

Pupils who misbehave appear to do so with different teachers. However, there are significant differences in behaviour rates by time of day, day of the week, and in different subjects. This suggests that misbehaviour in schools is partly driven by context, and that it could be reduced by modifying the context. Importantly, a pupil's relative rank is largely maintained throughout high-incident and low-incident times of the day and week, and to a lesser extent in different subjects. This is interesting because it suggests that rank order stability might remain even where absolute levels of noncognitive traits vary. I have data for only a limited range of contexts: the databases only record incidents which occur at school, so the variation in context relies on timing and different activities whilst at school. I do not have any information on how pupils

[138] Schools may already do this through selective admissions, see e.g. Gibbons and Silva (2011).

behave outside of school, or at different schools, so I cannot reject that the behaviour observed is to some extent situation-specific. However, the situation in which I measure behaviour will have much in common with many later contexts (e.g. further education and training, and the workplace), and there is likely to be less endogenous selection into these schools than into a typical adult workplace because pupils of this age do not have much control over their environments (Borghans et al., 2008). Moreover, the evidence I have presented suggests that in this context pupils' rank order of behaviour incidents stays relatively constant across different teachers, subjects, times of day and days of the week, even when absolute levels of behaviour fluctuate. Thus although we might expect individuals' expressed noncognitive traits to vary by activity, at least within the school there appears to be strong rank order stability in the tendency to misbehave.

This does not mean that behaviour is not malleable. In policy terms, we should be most interested in the *absolute* standard of behaviour, and here it is clear that context is important. In particular, the overall framework schools use for behaviour management is probably very important to reducing poor behaviour, promoting good behaviour, and reducing the impact of poor behaviour on other pupils and teachers (DfE, 2012a). However, the rank-order stability of behaviour suggests that the personality traits of individual pupils are significant, and so intervention at a pupil level should also be considered. In schools which already have good behaviour systems, providing greater resources and effective support for pupils with the most severe behavioural issues could greatly decrease the number and severity of behaviour incidents. This may be what the designation of EBD SEN is for: to highlight that some pupils are in need of additional support, and to target resources towards them. However, in many cases schools will not receive additional funding for this work, and there is evidence that provision for pupils with emotional and behavioural SEN varies greatly between schools and may not be effective (Taylor, 2013; Jonas, 2011; Atkinson, Lamont, & Downing, 2007; Keslair, Maurin & McNally, 2011). Since in-school provision for pupils with behaviour problems is relatively limited, this suggests that I have been observing *normative* change (changes that are typical of a given population in the absence of intervention) and standard behaviour given the context, which still leaves room for nonnormative change as a result of intentional efforts to modify behaviour (Borghans et al., 2008). International evidence on well-evaluated behavioural interventions suggests that there are programmes available which have substantial impacts on behaviour (e.g. Durlak et al., 2011). Such programmes are evidence based, and use detailed curriculum materials and trained staff. Given the extreme concentration of incidents amongst a small percentage of pupils, targeting resources towards these pupils and their families may be more efficient than attempts at universal prevention, and even expensive interventions could be cost effective. Introducing them into mainstream schools could be a cost-effective way of improving behaviour for this targeted group of pupils, improving attainment for many others too by reducing the negative externalities imposed on them in the classroom.[139] Improving behaviour at school could be valuable not just for the purposes of academic attainment, although this is a major motivation. As Bowles et al. (2001) suggest, schools help to develop pupils' noncognitive skills, so by becoming better at inducing good behaviour they may promote pupils' longer term noncognitive capacities.

[139] Special schools for pupils with EBD SEN may already offer this sort of targeted support, but I do not have any behavioural data for schools of this type.

Chapter 4: Figures and Tables

Figure 4.1: Cumulative frequency of incidents per pupil
N pupils: 3,284; N incidents: 45,493

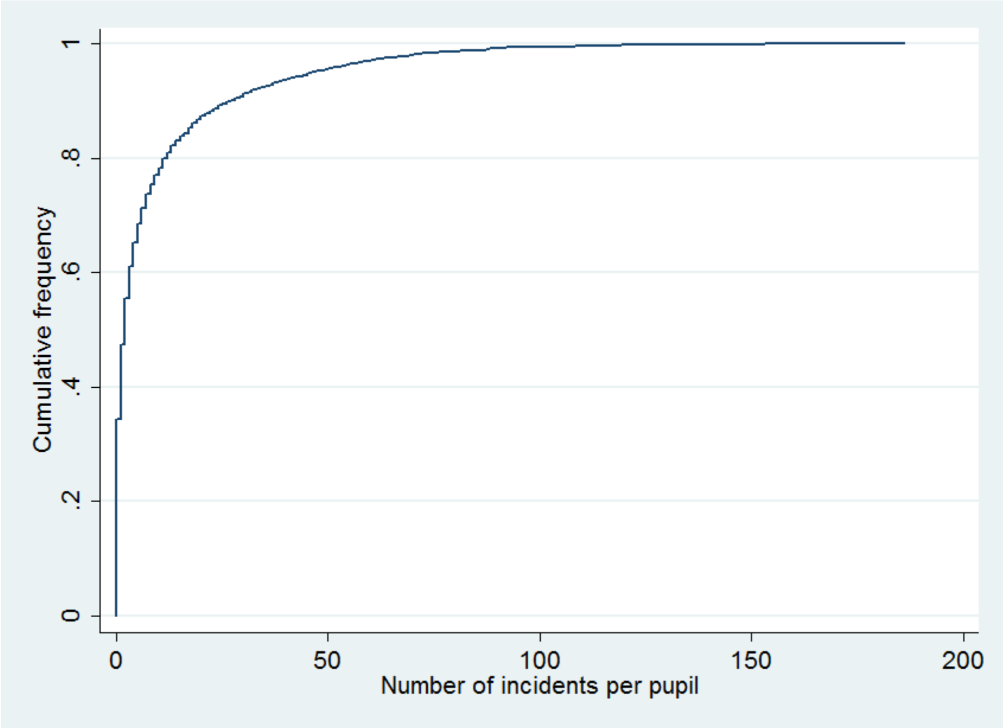


Figure 4.2 Cumulative frequency: number of incidents per pupil per day
N pupils: 3,284; N incidents: 45,493

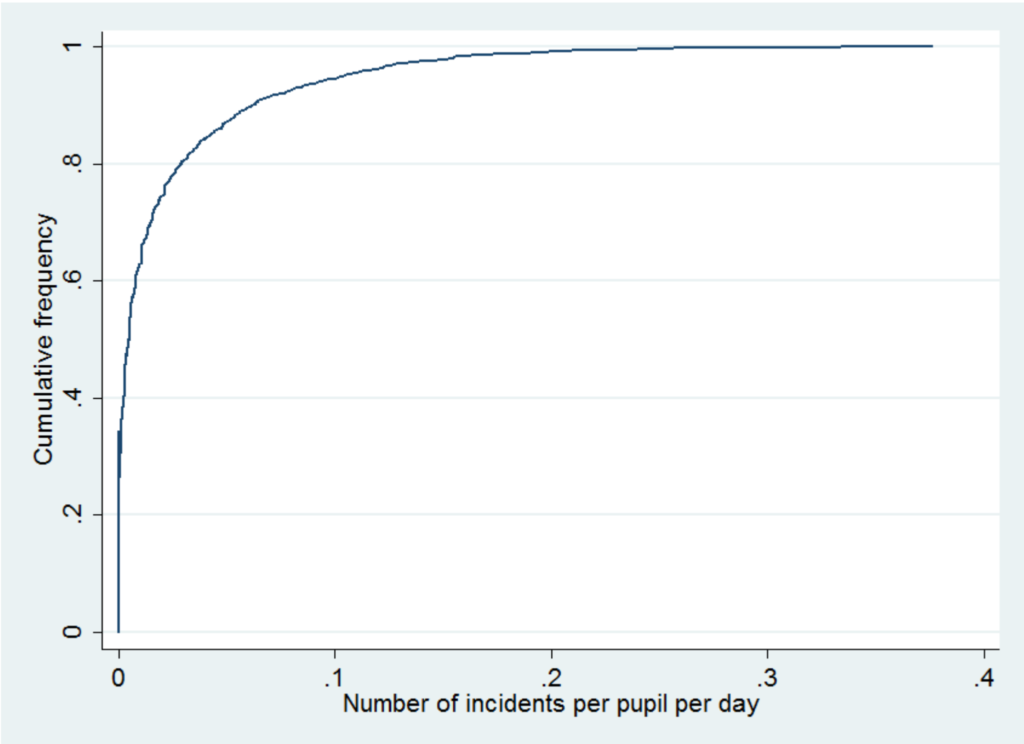


Figure 4.3: Mean behaviour incident rate by yeargroup
N pupils: 3,284; N incidents: 45,493. 95% confidence intervals shown in red

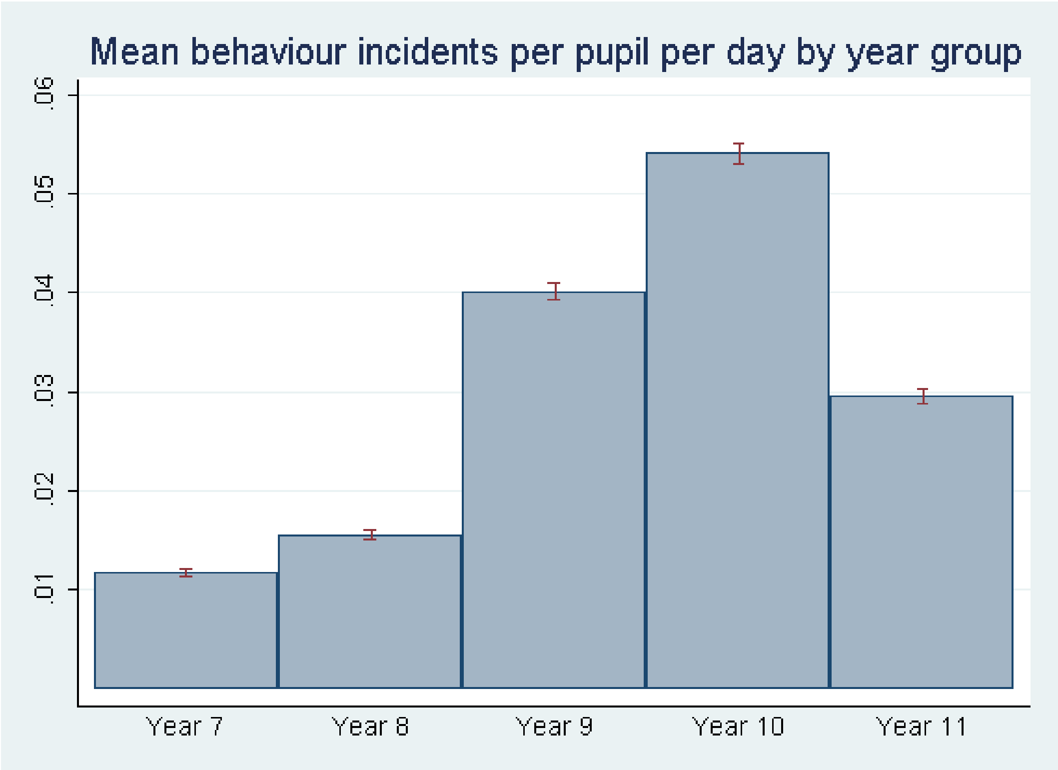


Figure 4.4: Persistence of behaviour through time
N pupils: 2,702 – sample is restricted to pupils with at least 5 terms of data

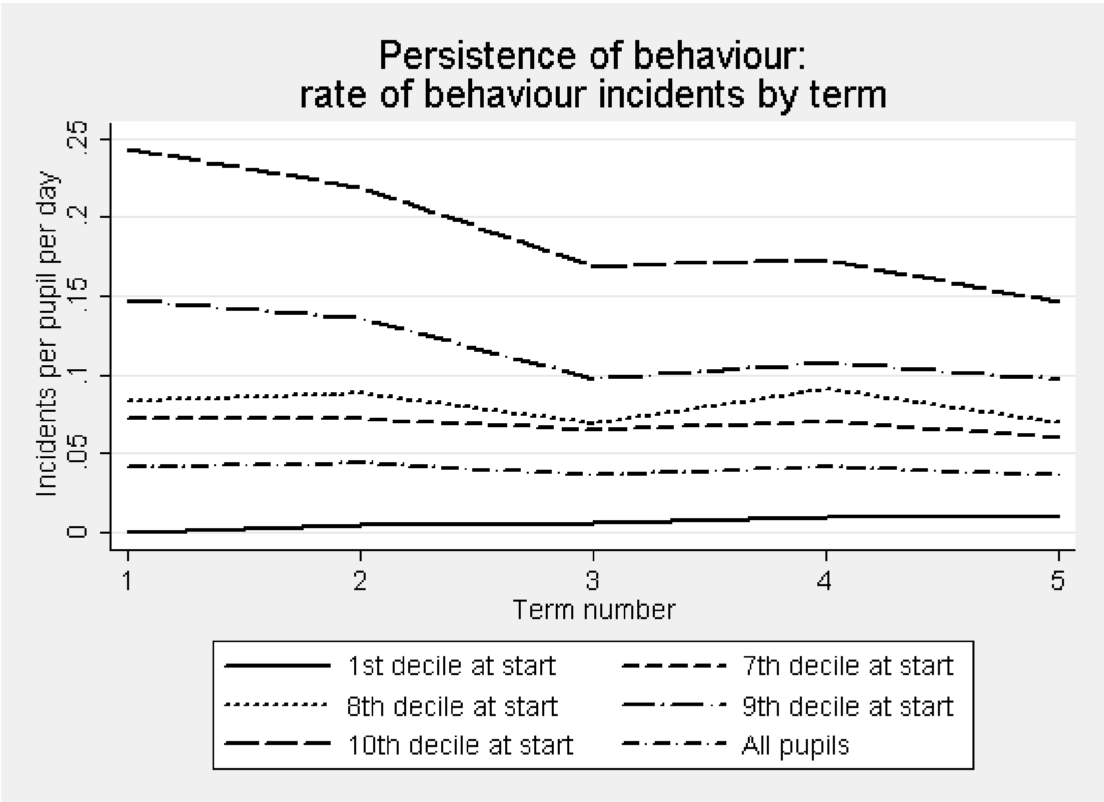


Figure 4.5: Persistence of behaviour incidents – predicting future incidents
Figure plots incidence rate ratios from negative binomial regressions of the number of incidents in later half terms on the number of incidents in the first half term a pupil appears in the data, see notes to Table 4.14 for details and sample sizes at each point.

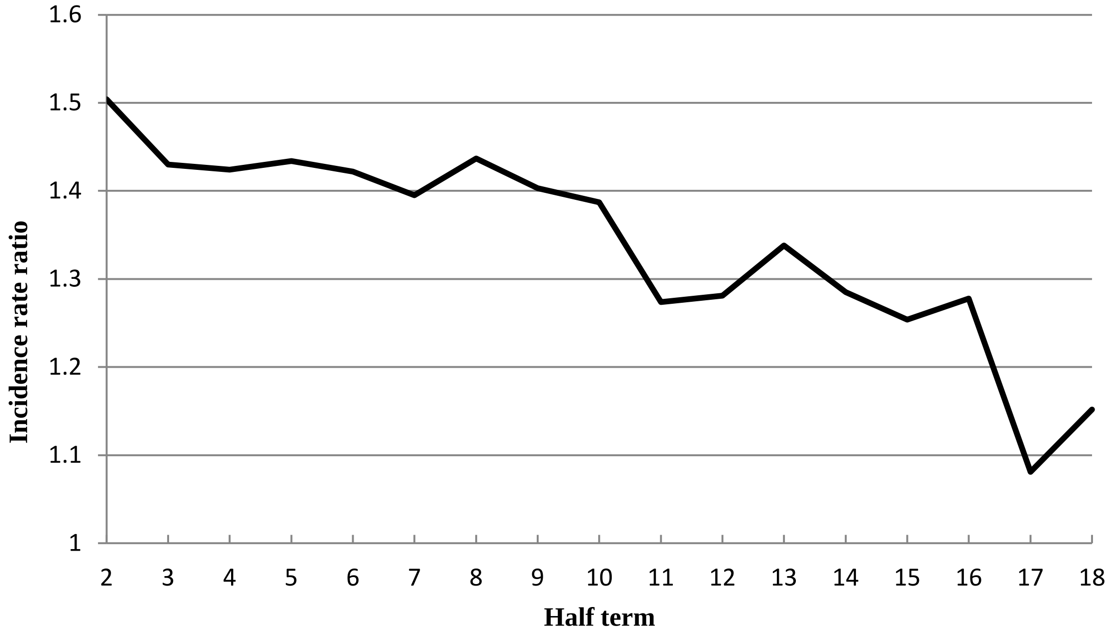


Figure 4.6: Incidents per pupil and number of teachers involved
N pupils: 915; N incidents: 6,325

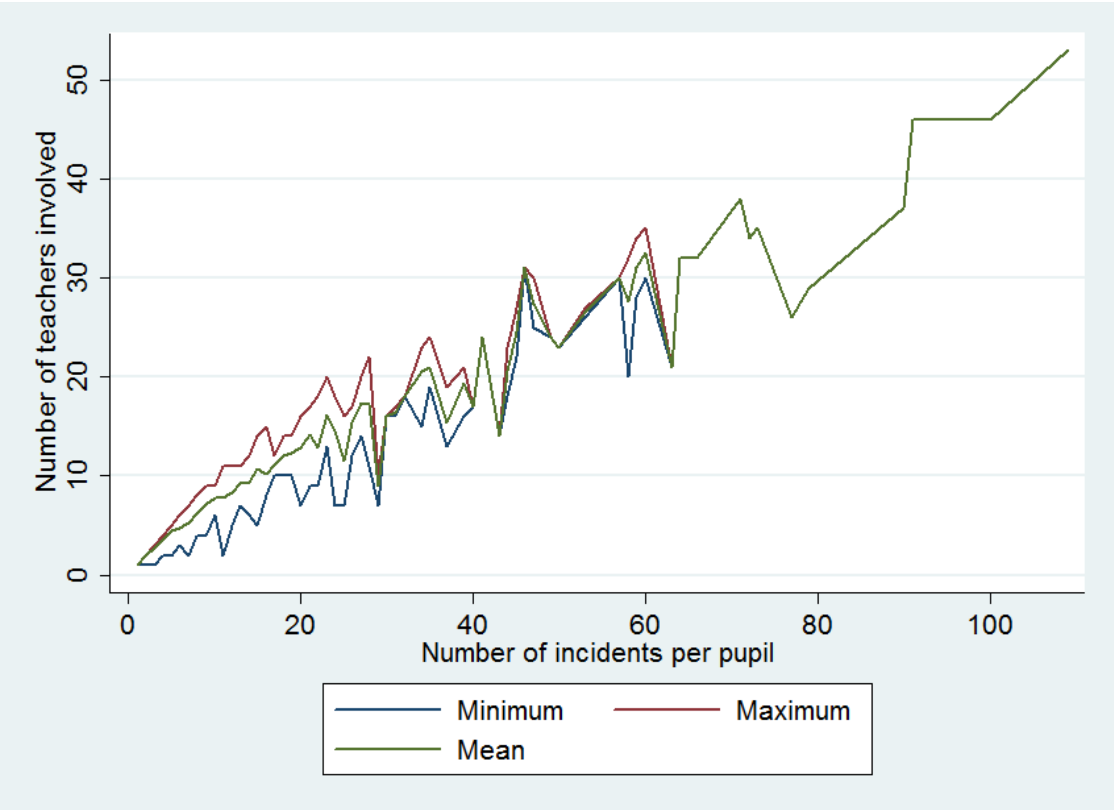


Figure 4.7: Mean behaviour incidents per pupil per day by day of the week
N pupils: 3,284; N incidents: 44,668; bars show 95% confidence intervals

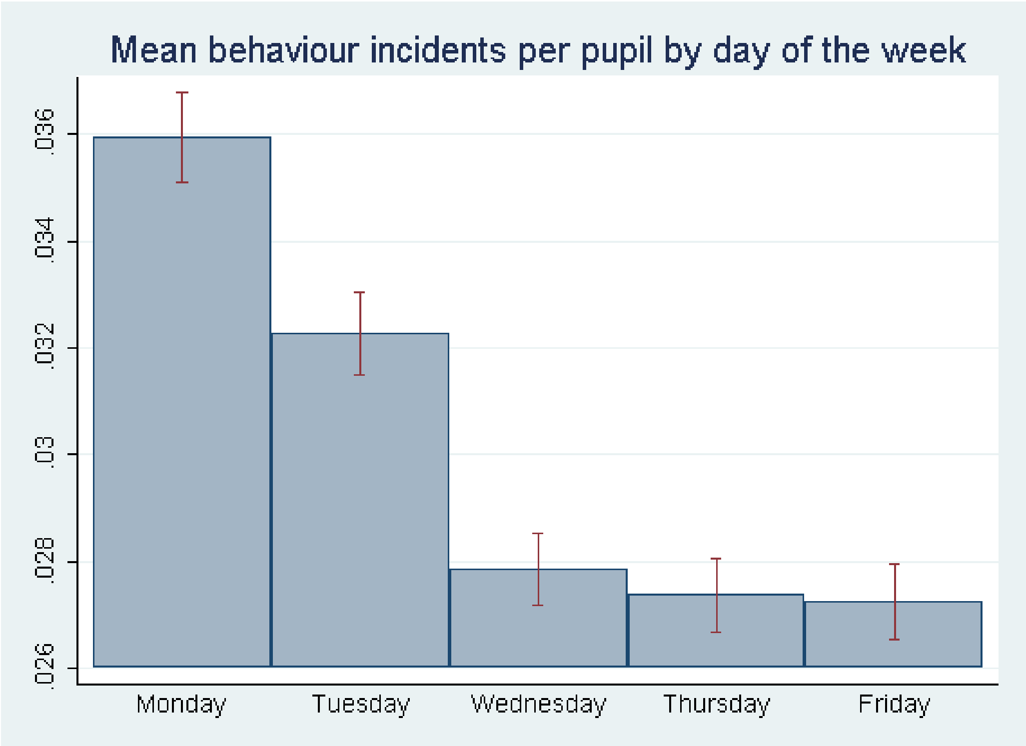


Figure 4.8: Behaviour incidents per pupil by lesson timing
N pupils: 1,096; N incidents: 14,634; bars show 95% confidence intervals

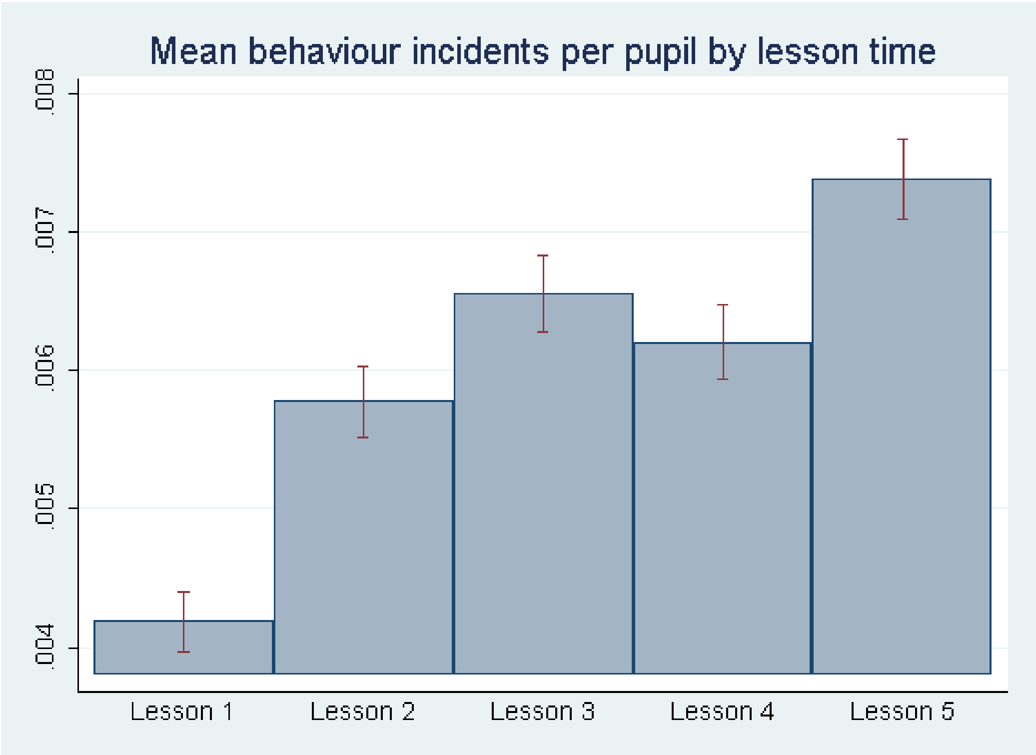


Table 4.1: Comparison of schools with behaviour information with all state-funded mainstream secondary schools in England

	England average	All 4 schools	School 1	School 2	School 3	School 4
Fraction male	0.51	0.5134	0.4413	0.556	0.5497	0.4969
Fraction with Special Educational Needs	0.21	0.1969	0.4326	0.3694	0.1581	0.1694
Fraction eligible for Free School Meals in past six years	0.27	0.3026	0.2316	0.2388	0.1405	0.3950
IDACI score: neighbourhood deprivation measure	0.22	0.2865	0.2383	0.2395	0.1781	0.3491
Fraction attained level 4 in Key Stage 2 English (2006)	0.79	0.7725	0.7679	0.6951	0.8165	0.7595
Fraction attained level 4 in Key Stage 2 maths (2006)	0.76	0.7455	0.756	0.649	0.791	0.733
Fraction attained level 4 in Key Stage 2 science (2006)	0.87	0.8815	0.8393	0.8293	0.906	0.8805
Fraction white ethnicity	0.82	0.8752	0.8983	0.8842	0.9157	0.8526
Number of permanent exclusions per pupil	0.0035	0.0025	0.0014	0.0013	0.0063	0.0010
Number of fixed-term exclusions per pupil	0.1145	0.1075	0.1343	0.0630	0.1223	0.1102
Number of fixed-term exclusion sessions per pupil	0.6155	0.8106	0.6557	0.4357	0.8973	1.2538
Size of school (pupils)	950		700	762	1120	989

Notes: Figures are for 2008-09 except where stated otherwise. Figures for all four schools are an average of the four schools' figures weighted by the number of pupils at each. Figures for England are for all state-funded mainstream secondary schools, except for IDACI and Key Stage 2 scores which are reported for the whole population. SEN figures from 2009 data (chapter 1 tables from DCSF, 2009a); figures on gender composition, ethnicity and median school size from DCSF (2009b). The average school size for England reported is the median size for maintained secondary schools: 901-1000 pupils. School size for the four included schools is that in 2009. IDACI score from APHO (2011), derived from deprivation data from the Department for Communities and Local Government for 2010 applied to 2009 population estimates. Key Stage 2 data from DCSF (2009c); 2006 data reported because the median (and mode) cohort of pupils in the behaviour databases sat KS2 exams in 2006. FSM eligibility over six years is from DfE (2012c), based on school censuses from January 2006 to March 2012. Exclusions from 2007-08 NPD pupil-level data for all maintained non-selective secondary schools in England; none of the combined four schools' exclusion rates is significantly different from national average.

Table 4.2: Behaviour incidents by type

Type of incident	N		Percent
missing categorisation or other	**8,494**		**18.67**
severity not specified		*5,406*	
other (minor)		*1,868*	
other (severe)		*1,220*	
disruption	**12,486**		**27.45**
defiance	**7,935**		**17.44**
defiance		*7,393*	
persistent defiance		*542*	
lateness	**4,211**		**9.26**
lateness		*4,156*	
persistent lateness		*55*	
no equipment	**3,035**		**6.67**
no equipment		*2,978*	
persistent no equipment		*57*	
verbal abuse	**2,064**		**4.54**
verbal abuse - not specified		*254*	
verbal abuse towards pupil		*277*	
rudeness/insolence towards teacher		*485*	
verbal abuse towards teacher		*1,048*	
truancy/leaving lesson without permission	**1,525**		**3.35**
assault, aggression or force	**1,368**		**3.01**
assault - not specified		*26*	
assault on teacher		*39*	
assault on pupil		*814*	
fighting		*489*	
missed detention	**539**		**1.18**
inadequate work	**746**		**1.64**
uniform/jewellery	**536**		**1.18**
damage to property	**516**		**1.13**
bullying	**464**		**1.02**
bullying - not specified		*199*	
bullying of pupil		*19*	
bullying of teacher		*8*	
bullying physical		*104*	
bullying verbal		*134*	
dangerous behaviour	**414**		**0.91**
swearing	**252**		**0.55**
smoking	**212**		**0.47**
out of bounds	**178**		**0.39**
racist incident	**163**		**0.36**
no homework	**152**		**0.33**
no homework		*81*	
persistent lack of homework		*71*	
theft	**79**		**0.17**
threatening behaviour	**51**		**0.11**
threatening behaviour - not specified		*10*	
threatening behaviour to student		*18*	
threatening behaviour to teacher		*23*	
illicit substances	**28**		**0.06**
mobile phone	**23**		**0.05**
forging notes/failure to inform parents	**22**		**0.05**
TOTAL	**45,493**		**100**

Table 4.3: Pupil characteristics

	All pupils					10% of pupils with most incidents			
	Mean	**SD**	**Min**	**Max**	**N**	**mean**	**SD**	**N**	**p-value**
Number of incidents per pupil (in dataset)	13.85	36.16	0	476	3,284	66.76	76.07	321	0.000
Number of school days per pupil (in dataset)	451.8	221.7	9	954	3,284	424.2	190.6	321	0.313
Number of incidents per pupil per day	0.04	0.11	0	1.50	3,284	0.18	0.24	321	0.000
Assault or fighting: N incidents	0.45	1.26	0	24	3,284	2.32	2.77	321	0.000
Defiance: N incidents	2.42	5.97	0	62	3,284	14.93	11.46	321	0.000
Disruption: N incidents	3.80	15.11	0	263	3,284	19.61	35.75	321	0.001
Lateness or truancy: N incidents	1.75	9.91	0	198	3,284	6.10	21.58	321	0.039
Dangerous behaviour or property damage: N incidents	0.31	0.87	0	11	3,284	1.67	1.85	321	0.000
Verbal offence: N incidents	0.88	2.19	0	24	3,284	5.22	4.20	321	0.000
Other types of incident (minor): N	1.66	4.48	0	57	3,284	5.74	9.18	321	0.000
Fraction male	0.51	0.50			3,284	0.71	0.46	321	0.000
Fraction with Special Educational Needs (SEN)	0.20	0.40			3,275	0.47	0.50	319	0.000
Fraction with emotional or behavioural disorder SEN	0.03	0.16			3,121	0.10	0.30	319	0.002
Age in September 2009	14.59	2.27	11	20.08	3,278	15.19	2.18	319	0.058
Born in autumn	0.25	0.43			3,278	0.24	0.43	319	0.710
Born in winter	0.23	0.42			3,278	0.27	0.44	319	0.170
Born in spring	0.26	0.44			3,278	0.25	0.43	319	0.701
Born in summer	0.26	0.44			3,278	0.24	0.43	319	0.274
Year of birth	1994.5	2.31	1989	1998	3,278	1993.9	2.22	321	0.062
Fraction eligible for Free School Meals in past 6 years	0.30	0.46			3,276	0.50	0.50	319	0.000
IDACI score: neighbourhood deprivation measure	0.29	0.18	0	0.96	3,247	0.32	0.17	319	0.000
Mean Key Stage 2 score (combined English, maths and science)	4.51	0.67	1	5.74	2,710	4.15	0.69	267	0.000

	All pupils					10% of pupils with most incidents			
	Mean	**SD**	**Min**	**Max**	**N**	**mean**	**SD**	**N**	**p-value**
Fraction white ethnicity	0.88	0.33			3,237	0.93	0.25	320	0.002
Fraction Bangladeshi or Pakistani ethnicity	0.05	0.22			3,237	0.00	0.06	320	0.000
Fraction black	0.01	0.12			3,237	0.01	0.11	320	0.749
Fraction other Asian ethnicity	0.02	0.15			3,237	0.02	0.15	320	0.791
Fraction other and mixed ethnicity	0.03	0.18			3,237	0.03	0.17	320	0.596
Developed ability score at start of secondary school	96.80	12.54	52	141	1,212	89.23	11.25	111	0.000
Depression symptoms score	9.65	6.53	0	44	1,303	13.50	8.00	139	0.000
Anxiety symptoms score	9.30	5.90	0	28	1,300	10.40	5.88	139	0.121
Self-reported behaviour score	12.10	5.50	0	34	1,304	15.62	5.38	139	0.000
Teacher-reported behaviour score	7.60	5.98	0	33	1,306	15.55	5.13	139	0.000
Pupil reports living with neither parent	0.07	0.25			1,296	0.14	0.35	138	0.033
Pupil reports living with one parent	0.46	0.50			1,296	0.55	0.50	138	0.004
Pupil reports living with both parents	0.47	0.50			1,296	0.30	0.46	138	0.000

(Pupil characteristics continued)

Notes: Each characteristic (e.g. gender) was regressed on a dummy for pupils' behaviour quantile (10^{th} decile versus all other pupils), along with dummies for school, with regressions clustered by class membership. The p-values on the coefficients on each dummy are reported in the fourth column for the top 10% of pupils, e.g. these pupils are significantly more likely to be male (71% male versus 51% male for all students). Note that the figures for 'All pupils' include the 10% of pupils with the most behaviour incidents separately detailed in the last four columns.

Table 4.4: Fraction of pupils explaining fraction of incidents

	Percentage of pupils explaining this percentage of incidents								
Percentiles of behaviour incidents	All incidents	Disruption	Defiance	Lateness or truancy	Verbal offence	Assault or fighting	Dangerous behaviour or damage	Other (minor - have got category)	Missing category
1%	0.15	0.12	0.12	0.10	0.12	0.12	0.12	0.15	0.14
5%	0.42	0.34	0.30	0.20	0.30	0.25	0.28	0.46	0.48
10%	0.94	0.67	0.61	0.43	0.61	0.52	0.55	1.01	1.04
25%	2.89	2.07	1.92	1.32	1.86	1.62	1.61	3.02	3.14
50%	7.98	5.72	5.36	4.04	5.24	4.69	4.60	8.07	8.51
75%	17.84	12.49	11.72	9.22	11.81	10.11	9.71	17.02	18.89
90%	31.97	21.53	20.04	15.31	20.04	16.44	14.31	27.71	31.88
95%	41.72	27.50	25.70	17.63	24.27	18.69	15.87	33.46	39.48
99%	58.55	35.39	34.71	20.12	27.80	20.46	17.09	38.46	51.52
All incidents	67.88	38.10	37.06	21.15	28.63	20.86	17.33	39.98	55.13
N pupils	3,284	3,284	3,284	3,284	3,284	3,284	3,284	3,284	3,284
N incidents	45,493	12,486	7,943	5,736	2,890	1,472	1,009	5,463	8,494
% incidents in this category	100.00	27.45	17.46	12.61	6.35	3.24	2.22	12.01	18.67

Notes: This table shows the concentration of behaviour incidents, e.g. that 10% of all behaviour incidents are accounted for by only 0.94% of pupils, 50% of incidents are due to 7.98% of pupils, and so on. This is another way of representing the skewness of behaviour incidents: a very small number of pupils are responsible for the large majority of incidents, while one third of pupils have no incidents at all. This is calculated by working out the fraction of incidents explained by the fraction of pupils within each school (to allow for differential reporting thresholds by school), then taking a weighted average of the school figures to generate an overall number. Running the same exercise on each school separately or with all data pooled results in very similar results. For instance, the percentage of pupils explaining 25% of incidents varies between 2.3% and 6.1% of pupils across schools. Incidents have been placed into broad categories based on the more detailed category in Table 4.2. The category of ‘Other (minor)’, includes inadequate work or homework, no equipment, incorrect uniform, and other infrequent and minor incidents which I have a categorisation for. ‘Missing category’ covers the 8,494 incidents for which I do not have any type information.

Table 4.5: Predicting the number of behaviour incidents per pupil

	(1)	(2)	(3)	(4)	(5)
Boy	0.540***	0.535***	0.758***	0.345***	0.934***
SE	(0.135)	(0.148)	(0.148)	(0.096)	(0.119)
IRR	1.717	1.708	2.135	1.412	2.544
FSM	0.578***	0.575***	0.479***	0.153	0.319***
SE	(0.054)	(0.065)	(0.110)	(0.117)	(0.119)
IRR	1.783	1.778	1.614	1.165	1.376
SEN	0.781***	0.504***	0.387***	-0.339***	
SE	(0.088)	(0.072)	(0.126)	(0.126)	
IRR	2.183	1.655	1.472	0.712	
White	0.475***	0.581***	0.676***	0.232	0.502***
SE	(0.133)	(0.123)	(0.187)	(0.156)	(0.157)
IRR	1.608	1.789	1.967	1.261	1.653
IDACI score of neighbourhood	0.680***	0.647***	0.926**	0.139	0.541
SE	(0.186)	(0.191)	(0.396)	(0.296)	(0.381)
IRR	1.973	1.910	2.524	1.149	1.717
Winter born	-0.031	-0.042	-0.343**	-0.379***	-0.394***
SE	(0.076)	(0.084)	(0.134)	(0.124)	(0.130)
IRR	0.969	0.959	0.710	0.684	0.674
Spring born	-0.148*	-0.276***	-0.298**	-0.262**	-0.250
SE	(0.081)	(0.090)	(0.128)	(0.103)	(0.153)
IRR	0.862	0.759	0.743	0.770	0.778
Summer born	-0.130	-0.209**	-0.384**	-0.468***	-0.507***
SE	(0.090)	(0.099)	(0.179)	(0.115)	(0.155)
IRR	0.878	0.812	0.681	0.626	0.602
Mean KS2 score		-0.457***	-0.368***	0.017	-0.221**
SE		(0.062)	(0.092)	(0.109)	(0.098)
IRR		0.633	0.692	1.017	0.802
EBD SEN				0.554***	
SE				(0.187)	
IRR				1.740	
Depression score				0.039***	0.118***
SE				(0.012)	(0.015)
IRR				1.040	1.126
Anxiety score				-0.106***	-0.076***
SE				(0.016)	(0.016)
IRR				0.900	0.927
Self-reported behaviour score				0.092***	
SE				(0.019)	
IRR				1.096	
Teacher-reported behaviour score				0.152***	
SE				(0.010)	
IRR				1.165	
Does not live with parents				0.250*	0.680***
SE				(0.151)	(0.166)
IRR				1.285	1.974
Lives with one parent				0.239***	0.562***
SE				(0.086)	(0.116)
IRR				1.270	1.755
Developed ability score at age 11				-0.010*	-0.015***
SE				(0.006)	(0.006)
IRR				0.990	0.985
N	3,226	2,683	1,041	1,041	1,041
N clusters	72	69	52	52	52

Notes: Table reports results of negative binomial regressions predicting the number of incidents per pupil, reporting coefficients, standard errors and incidence rate ratios (IRRs), with the exposure set to the number of days a pupil appears in the dataset (number of school days). Regressions include dummies for school; standard errors are clustered by class group. Each column represents a separate regression.

Table 4.6: Predicting the number of behaviour incidents per pupil: boys

	(1)	(2)	(3)	(4)	(5)
FSM	0.515***	0.550***	0.365**	0.009	0.301*
SE	(0.064)	(0.083)	(0.146)	(0.142)	(0.170)
IRR	1.673	1.733	1.440	1.009	1.351
SEN	0.790***	0.542***	0.508***	-0.384***	
SE	(0.114)	(0.093)	(0.152)	(0.147)	
IRR	2.203	1.719	1.662	0.681	
White	0.246*	0.259**	0.488**	0.281*	0.338
SE	(0.129)	(0.130)	(0.236)	(0.155)	(0.211)
IRR	1.278	1.295	1.630	1.325	1.402
IDACI score of neighbourhood	1.056***	1.063***	1.038**	0.614*	0.765*
SE	(0.246)	(0.327)	(0.477)	(0.370)	(0.460)
IRR	2.876	2.894	2.822	1.848	2.148
Winter born	0.070	0.065	-0.201	-0.299*	-0.216
SE	(0.106)	(0.118)	(0.168)	(0.155)	(0.170)
IRR	1.073	1.067	0.818	0.742	0.806
Spring born	-0.100	-0.153	-0.210	-0.140	-0.074
SE	(0.094)	(0.116)	(0.176)	(0.137)	(0.207)
IRR	0.905	0.859	0.811	0.869	0.928
Summer born	-0.045	-0.097	-0.437**	-0.325**	-0.434**
SE	(0.091)	(0.107)	(0.180)	(0.140)	(0.176)
IRR	0.956	0.907	0.646	0.723	0.648
Mean KS2 score		-0.466***	-0.377***	0.133	-0.197
SE		(0.080)	(0.120)	(0.173)	(0.163)
IRR		0.627	0.686	1.142	0.822
EBD SEN				0.580***	
SE				(0.213)	
IRR				1.787	
Depression score				0.005	0.088***
SE				(0.017)	(0.018)
IRR				1.005	1.092
Anxiety score				-0.122***	-0.060***
SE				(0.018)	(0.019)
IRR				0.885	0.942
Self-reported behaviour score				0.123***	
SE				(0.021)	
IRR				1.131	
Teacher-reported behaviour score				0.154***	
SE				(0.014)	
IRR				1.167	
Does not live with parents				0.222	0.633***
SE				(0.188)	(0.188)
IRR				1.249	1.884
Lives with one parent				0.121	0.453***
SE				(0.108)	(0.132)
IRR				1.129	1.572
Developed ability score at age 11				-0.011	-0.017*
SE				(0.008)	(0.009)
IRR				0.989	0.984
N	1,658	1,369	522	522	522
N clusters	66	64	51	51	51

See notes to Table 4.5.

Table 4.7: Predicting the number of behaviour incidents per pupil: girls

	(1)	(2)	(3)	(4)	(5)
FSM	0.652***	0.642***	0.730***	0.340**	0.468***
SE	(0.101)	(0.107)	(0.173)	(0.144)	(0.172)
IRR	1.919	1.899	2.075	1.406	1.597
SEN	0.739***	0.443***	0.287	-0.332**	
SE	(0.116)	(0.123)	(0.234)	(0.167)	
IRR	2.094	1.558	1.332	0.718	
White	0.769***	1.039***	0.908***	0.139	0.584**
SE	(0.253)	(0.159)	(0.351)	(0.309)	(0.296)
IRR	2.157	2.825	2.480	1.149	1.793
IDACI score of neighbourhood	0.444	0.411	1.031*	-0.397	0.511
SE	(0.271)	(0.255)	(0.624)	(0.453)	(0.580)
IRR	1.558	1.508	2.804	0.672	1.668
Winter born	-0.138	-0.125	-0.394*	-0.477***	-0.508***
SE	(0.106)	(0.110)	(0.219)	(0.179)	(0.173)
IRR	0.871	0.883	0.674	0.621	0.602
Spring born	-0.165	-0.347**	-0.348	-0.403***	-0.355*
SE	(0.141)	(0.140)	(0.217)	(0.153)	(0.193)
IRR	0.848	0.707	0.706	0.668	0.701
Summer born	-0.212	-0.266	-0.293	-0.522***	-0.492**
SE	(0.161)	(0.181)	(0.271)	(0.153)	(0.221)
IRR	0.809	0.766	0.746	0.593	0.611
Mean KS2 score		-0.422***	-0.436***	-0.143	-0.364**
SE		(0.090)	(0.131)	(0.149)	(0.160)
IRR		0.655	0.646	0.867	0.695
EBD SEN				0.956	
SE				(0.628)	
IRR				2.600	
Depression score				0.071***	0.155***
SE				(0.017)	(0.019)
IRR				1.074	1.168
Anxiety score				-0.093***	-0.095***
SE				(0.024)	(0.018)
IRR				0.911	0.910
Self-reported behaviour score				0.067**	
SE				(0.027)	
IRR				1.069	
Teacher-reported behaviour score				0.153***	
SE				(0.013)	
IRR				1.165	
Does not live with parents				0.249	0.610**
SE				(0.256)	(0.293)
IRR				1.282	1.841
Lives with one parent				0.369***	0.613***
SE				(0.124)	(0.152)
IRR				1.447	1.846
Developed ability score at age 11				-0.010	-0.011
SE				(0.007)	(0.009)
IRR				0.990	0.990
N	1,568	1,314	519	519	519
N clusters	70	66	51	51	51

See notes to Table 4.5.

Table 4.8: Predicting the number of incidents by behaviour category

	All incidents	Assault/ fighting	Defiance	Disruption	Lateness/ truancy	Dangerous behaviour	Verbal offence	Other (minor)	Missing category
Boy	0.535***	1.242***	0.371**	0.718***	0.045	1.039***	0.574***	0.397***	0.407**
SE	(0.148)	(0.181)	(0.176)	(0.177)	(0.127)	(0.125)	(0.188)	(0.089)	(0.172)
IRR	1.708	3.462	1.449	2.050	1.046	2.827	1.776	1.488	1.503
FSM	0.575***	0.589***	0.661***	0.527***	0.638***	0.648***	0.674***	0.612***	0.325***
SE	(0.065)	(0.134)	(0.066)	(0.090)	(0.110)	(0.126)	(0.084)	(0.097)	(0.081)
IRR	1.778	1.803	1.938	1.694	1.892	1.912	1.962	1.844	1.385
SEN	0.504***	0.637***	0.578***	0.575***	0.451***	0.482***	0.662***	0.280***	0.512***
SE	(0.072)	(0.090)	(0.078)	(0.088)	(0.123)	(0.104)	(0.101)	(0.107)	(0.105)
IRR	1.655	1.892	1.782	1.777	1.570	1.619	1.939	1.322	1.668
White	0.581***	0.195	0.825***	0.647***	0.504**	0.415**	0.601***	0.235**	0.653***
SE	(0.123)	(0.171)	(0.196)	(0.165)	(0.212)	(0.182)	(0.172)	(0.107)	(0.147)
IRR	1.789	1.215	2.282	1.910	1.655	1.515	1.825	1.265	1.922
IDACI score of neighbourhood	0.647***	0.737**	0.777***	0.734***	0.598*	0.373	0.906***	0.962***	0.387
SE	(0.191)	(0.320)	(0.277)	(0.267)	(0.328)	(0.255)	(0.258)	(0.262)	(0.242)
IRR	1.910	2.090	2.174	2.084	1.818	1.452	2.475	2.616	1.472
Winter born	-0.042	-0.068	-0.044	-0.059	0.003	0.035	-0.092	0.032	-0.080
SE	(0.084)	(0.128)	(0.126)	(0.099)	(0.181)	(0.129)	(0.117)	(0.093)	(0.110)
IRR	0.959	0.934	0.957	0.943	1.003	1.036	0.912	1.032	0.923
Spring born	-0.276***	-0.105	-0.339***	-0.354***	-0.127	-0.076	-0.244*	-0.191**	-0.368***
SE	(0.090)	(0.120)	(0.115)	(0.109)	(0.148)	(0.165)	(0.128)	(0.078)	(0.113)
IRR	0.759	0.900	0.712	0.702	0.881	0.927	0.784	0.826	0.692
Summer born	-0.209**	-0.245	-0.163	-0.198*	-0.312*	-0.255*	-0.241**	-0.191*	-0.166
SE	(0.099)	(0.158)	(0.115)	(0.115)	(0.185)	(0.132)	(0.123)	(0.106)	(0.114)
IRR	0.812	0.782	0.849	0.820	0.732	0.775	0.786	0.826	0.847
Mean Key Stage 2 test score	-0.457***	-0.486***	-0.549***	-0.537***	-0.520***	-0.534***	-0.381***	-0.179***	-0.425***
SE	(0.062)	(0.082)	(0.067)	(0.077)	(0.085)	(0.109)	(0.089)	(0.057)	(0.071)
IRR	0.633	0.615	0.578	0.584	0.595	0.586	0.683	0.836	0.654
N	2,683	2,683	2,683	2,683	2,683	2,683	2,683	2,683	2,683
N clusters	69	69	69	69	69	69	69	69	69

(Notes on next page)

Notes: Table reports results of negative binomial regressions predicting the number of incidents per pupil, reporting coefficients, standard errors and incidence rate ratios (IRRs), with the exposure set to the number of days a pupil appears in the dataset (number of school days). Regressions include dummies for school attended; standard errors are clustered by class group. The first column is the same as column 2 of Table 4.5, for comparison. Incidents have been placed into broad categories based on the more detailed category in Table 4.2. Each column represents a separate regression, with the number of incidents of a particular type as the outcome.

Table 4.9: Value of r-squared from linear regressions predicting incidents per pupil per day

	Basic	Full
All incidents	0.055	0.059
Assault	0.068	0.131
Defiance	0.074	0.115
Disruption	0.036	0.043
Lateness or truancy	0.016	0.020
Dangerous behaviour or property damage	0.069	0.075
Verbal offence	0.092	0.122
Other (minor)	0.037	0.048
Missing categorisation	0.047	0.049
N pupils	2,683	1,041
N clusters	69	52

Notes: Table reports r-squared values from OLS regressions predicting the number of incidents per pupil per day in different categories. The row heading gives the outcome category of the incident, and each cell represents a different regression. The first column (basic) represents the same specifications as column 2 of Table 4.5; the second column (full) uses the same specifications as column 5 of Table 4.5. Dummies for the school attended are excluded here because they contribute substantially to the baseline r-squared. The outcome in each case is the number of incidents (in that category) per pupil per day; this takes into account the amount of time a pupil appears in the dataset.

Table 4.10: Spearman rank correlation coefficients between incident rates by type of incident

N pupils: 3,284

	Assault	Defiance	Disruption	Lateness or truancy	Dangerous behaviour or damage	Verbal offence	Other (minor)
Assault	1						
Defiance	0.483	1					
Disruption	0.479	0.669	1				
Lateness or truancy	0.248	0.479	0.526	1			
Dangerous behaviour or property damage	0.449	0.537	0.467	0.330	1		
Verbal offence	0.478	0.621	0.617	0.400	0.461	1	
Other types of incident (minor)	0.293	0.472	0.509	0.467	0.325	0.416	1

Notes: Table presents Spearman rank correlation coefficients for pairwise comparisons between behaviour incidents per pupil per day by incident type. This table includes all pupils. In every pairwise comparison $p<0.0001$, including when Bonferroni adjustments are made for multiple testing.

Table 4.11: Correlation coefficients between incident counts by type: pupils with at least one incident of any kind

N pupils: 2,229

	Assault	Defiance	Disruption	Lateness or truancy	Dangerous behaviour or property damage	Verbal offence	Other (minor)
Assault	1						
Defiance	0.408	1					
Disruption	0.387	0.596	1				
Lateness or truancy	0.145	0.400	0.474	1			
Dangerous behaviour or property damage	0.394	0.495	0.396	0.248	1		
Verbal offence	0.410	0.560	0.548	0.312	0.398	1	
Other types of incident (minor)	0.143	0.315	0.374	0.398	0.210	0.270	1

Notes: Table presents Spearman rank correlation coefficients for pairwise comparisons between behaviour incidents per pupil per day by incident type. This table includes only pupils with at least one incident of any type. In every pairwise comparison $p<0.0001$, including when Bonferroni adjustments are made for multiple testing.

Table 4.12: Relationship between incident types

	Assault or fighting	Defiance	Disruption	Lateness or truancy	Dangerous behaviour/ damage	Verbal offence	Other types of incident (minor)
Assault		0.186***	0.219***	0.026	0.227***	0.293***	0.040
SE		(0.043)	(0.039)	(0.050)	(0.043)	(0.054)	(0.037)
IRR		1.204	1.245	1.027	1.255	1.341	1.041
Defiance	0.026**		0.087***	0.088***	0.071***	0.094***	0.037***
SE	(0.012)		(0.008)	(0.009)	(0.020)	(0.016)	(0.011)
IRR	1.026		1.091	1.092	1.074	1.098	1.038
Disruption	0.003	0.017		-0.015***	0.004	0.008	0.006**
SE	(0.006)	(0.016)		(0.003)	(0.005)	(0.007)	(0.003)
IRR	1.003	1.017		0.985	1.004	1.008	1.006
Lateness or truancy	0.001	0.017***	-0.010***		-0.021***	-0.003	0.010***
SE	(0.012)	(0.005)	(0.002)		(0.006)	(0.005)	(0.003)
IRR	1.001	1.017	0.990		0.979	0.997	1.011
Dangerous behaviour/damage	0.379***	0.478***	0.300***	0.166***		0.218***	0.212***
SE	(0.074)	(0.059)	(0.045)	(0.060)		(0.065)	(0.052)
IRR	1.460	1.613	1.350	1.181		1.244	1.236
Verbal offence	0.228***	0.321***	0.208***	0.118***	0.075**		0.100***
SE	(0.033)	(0.031)	(0.029)	(0.035)	(0.032)		(0.024)
IRR	1.256	1.379	1.231	1.125	1.078		1.105
Other (minor)	0.017	0.060***	0.059***	0.053***	0.066***	0.056***	
SE	(0.019)	(0.016)	(0.014)	(0.013)	(0.019)	(0.019)	
IRR	1.018	1.062	1.061	1.055	1.068	1.058	
N pupils	3,284	3,284	3,284	3,284	3,284	3,284	3,284
N clusters	74	74	74	74	74	74	74

Notes: Table presents coefficients, standard errors, and incidence rate ratios for negative binomial regressions of pupil-level counts of behaviour incidents in one category on the counts of behaviour incidents in all other categories and dummies for school. Each column represents a separate regression, with the outcome variable listed at the head of each column. For example, the first column gives coefficients of a negative binomial regression of incidents of assault on incidents involving defiance, disruption etc. Regressions are clustered by class/cohort, and exposure is set to number of school days a pupil appears in the dataset. The sum total of incidents missing categorisation is omitted in these regressions to avoid confounding the interpretation of the other variables.

Table 4.13: Factor analysis of incident types

	Factor 1	Factor 2	Uniqueness
Assault or fighting	0.2368	0.3236	0.8392
Defiance	0.7284	0.3481	0.3483
Disruption	0.7398	-0.2596	0.3852
Lateness or truancy	0.5529	-0.4506	0.4912
Dangerous behaviour or damage	0.4436	0.4715	0.5808
Verbal offence	0.6481	0.3486	0.4584
Other types of incident (minor)	0.7797	-0.4156	0.2194

Notes: Table shows factor loadings and unique variances for a principal factor analysis of incident types measured as daily incident rates by category for 3,284 pupils.

Table 4.14: Persistence of behaviour – predicting later incidents

Outcome: number of behaviour incidents during half term number

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
N incidents during half term 1	0.408***	0.358***	0.353***	0.360***	0.352***	0.333***	0.362***	0.339***	0.327***	0.242***	0.248***	0.291***	0.251***	0.226***	0.245***	0.078	0.142**
SE	(0.049)	(0.039)	(0.034)	(0.028)	(0.027)	(0.031)	(0.029)	(0.046)	(0.032)	(0.038)	(0.044)	(0.050)	(0.029)	(0.051)	(0.058)	(0.116)	(0.072)
IRR	1.504	1.430	1.424	1.434	1.422	1.395	1.437	1.403	1.387	1.274	1.281	1.338	1.285	1.254	1.278	1.081	1.152
N	3,205	3,181	3,164	3,146	3,121	2,725	2,697	2,664	2,456	2,269	2,248	1,366	1,348	1,326	1,155	1,139	1,009
N clusters	73	71	71	71	71	68	68	68	65	55	53	33	33	32	31	31	26

Notes: Table shows results of negative binomial regression of the number of behaviour incidents during later half terms on the number of incidents during the first half term a pupil appears in the dataset. Each column represents a separate regression. Regressions include dummies for school; gender; SEN; FSM eligibility; ethnic group; deprivation score of neighbourhood; and season of birth. Exposure is set to the number of school days in the outcome half term. Standard errors are clustered by class group. A full academic year contains 6 half terms, so predictions are made up to three years after baseline.

Table 4.15: Persistence of behaviour – Spearman rank correlation coefficients by half term

	Rate of incidents in half term number																
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Incident rate in half term 1	0.53	0.53	0.55	0.53	0.55	0.51	0.50	0.51	0.48	0.46	0.46	0.49	0.42	0.41	0.38	0.24	0.24
N pupils	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013	1,013
Incident rate in half term 1	0.63	0.58	0.60	0.53	0.53	0.57	0.56	0.52	0.51	0.36	0.36	0.43	0.37	0.40	0.37	0.23	0.24
N pupils	3,205	3,181	3,164	3,146	3,121	2,725	2,697	2,664	2,456	2,269	2,248	1,366	1,348	1,326	1,155	1,139	1,009

Notes: Table shows Spearman rank correlation coefficients for pairwise comparisons between behaviour incidents per pupil per day in the first half term pupils appear in the dataset, and each subsequent half term. The first row restricts observations to those who have values for all half terms 1-18; the second allows the composition of the sample to vary by period. In every pairwise comparison $p<0.0001$, including when Bonferroni adjustments are made for multiple testing.

Table 4.16: Persistence of behaviour – transition out of 10th decile
N pupils: 239

	% in 8th decile or lower	% in 9th decile	% in 10th decile	% missing: censored	% missing: left school	% missing: finished school	% missing: all reasons
Half term 1	0	0	100	0	0	0	0
Half term 2	26.46	18.83	54.71	0.89	0	0	0.89
Half term 3	30.59	19.63	49.77	1.78	0.89	0	2.67
Half term 4	25.12	20.93	53.95	1.78	2.67	0	4.44
Half term 5	42.06	15.42	42.52	1.78	3.11	0	4.89
Half term 6	39.34	20.38	40.28	2.22	4.00	0	6.22
Half term 7	38.33	21.11	40.56	9.78	5.78	4.44	20.00
Half term 8	34.09	21.02	44.89	9.78	7.56	4.44	21.78
Half term 9	41.52	20.47	38.01	10.22	9.33	4.44	24.00
Half term 10	43.56	23.31	33.13	12.89	10.22	4.44	27.56
Half term 11	57.64	10.42	31.94	20.89	10.67	4.44	36.00
Half term 12	55.63	16.20	28.17	20.89	11.11	4.89	36.89
Half term 13	51.16	15.12	33.72	36.89	12.89	12.00	61.78
Half term 14	43.21	27.16	29.63	37.33	14.22	12.44	64.00
Half term 15	55.70	21.52	22.78	37.78	14.67	12.44	64.89
Half term 16	54.41	19.12	26.47	42.67	14.67	12.44	69.78
Half term 17	69.12	16.18	14.71	42.67	14.67	12.44	69.78
Half term 18	63.79	15.52	20.69	46.22	15.56	12.44	74.22

Notes: Table shows the transition between deciles of pupils whose behaviour puts them in the 10th decile in their yeargroup and school during the first half term in which they appear in the dataset. There are six half terms in a year. Missing data may be due to school-level censoring (panel provided by the school ends); pupils finishing compulsory schooling; or pupils leaving school for other reasons. Note that the table does not take into account whether pupils move down a decile then move up again – all that it records is the percentage originally in the top 10% who are also in the top 10% in that half term. The number of pupils in the 10th decile at baseline is 239; this actually represents about 7% of the 3,284 pupils in the dataset, due to a lumpy distribution of incidents at baseline. The percentages reported in the first panel use the number of pupils for whom data is still available as the denominator; the percentages reported in the second panel are the percentage of the original 239 who are missing data.

Table 4.17: Behaviour incidents by time of day and day of the week

	(1)	(2)
Lesson 2		0.322***
SE		(0.077)
IRR		1.380
Lesson 3		0.444***
SE		(0.066)
IRR		1.559
Lesson 4		0.397***
SE		(0.058)
IRR		1.487
Lesson 5		0.572***
SE		(0.052)
IRR		1.772
Other time of day		1.047***
SE		(0.082)
IRR		2.849
Tuesday	-0.126***	-0.086
SE	(0.040)	(0.054)
IRR	0.881	0.918
Wednesday	-0.248***	-0.347***
SE	(0.045)	(0.063)
IRR	0.781	0.707
Thursday	-0.268***	-0.327***
SE	(0.037)	(0.056)
IRR	0.765	0.721
Friday	-0.252***	-0.269***
SE	(0.038)	(0.046)
IRR	0.777	0.764
N pupils	2,683	1,034
N	1,296,961	1,978,128
Clusters	78	41

Notes: This table presents regressions of the number of incidents per pupil per day (column 1) or the number of incidents per pupil per lesson (column 2), on dummies for day of the week and time of day. Regressions also include pupil characteristics; dummies for school, half term, school*half term, the week of half term, and days directly before and after holidays. Included are incidents which can be assigned a specific date (44,668 of 45,493 incidents), excluding incidents such as no homework or missed detentions. Column 1 includes all schools. Time of day information is only available for two schools (14,634 incidents). The ‘other times’ category includes morning and afternoon registrations, break and lunch times, the transitions between lessons, and before and after school, and represents a longer time period than the hour available for each lesson.

Table 4.18: Spearman rank correlation coefficients between incident rates by day of the week
N pupils: 3,284

	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	1				
Tuesday	0.7248	1			
Wednesday	0.7070	0.7309	1		
Thursday	0.7102	0.7291	0.7361	1	
Friday	0.7202	0.7184	0.7235	0.7226	1

Table 4.19: Spearman rank correlation coefficients between incident rates by time of day
N pupils: 1,096

	Lesson 1	Lesson 2	Lesson 3	Lesson 4	Lesson 5	Other times of day
Lesson 1	1					
Lesson 2	0.7129	1				
Lesson 3	0.6824	0.7062	1			
Lesson 4	0.7052	0.7177	0.6979	1		
Lesson 5	0.6865	0.7045	0.7005	0.6922	1	
Other times of day	0.6569	0.6819	0.6537	0.6817	0.6600	1

Notes: Tables present Spearman rank correlation coefficients for pairwise comparisons between behaviour incidents per pupil per day (Table 4.18) and incidents per pupil per lesson (Table 4.19). In every pairwise comparison in both tables $p<0.0001$, including when Bonferroni adjustments are made for multiple testing.

Table 4.20: Relative frequency of incidents by subject

Denominator \ Numerator	English	Maths	Science	Performing Arts	Physical Education	Technology	Humanities
English	1	2.23	2.15	0.72	0.77	0.77	1.02
Maths	0.45	1	1.11	0.37	0.45	0.30	0.63
Science	0.47	0.90	1	0.35	0.38	0.36	0.53
Performing Arts	1.39	2.68	2.86	1	0.99	1.02	1.51
Physical Education	1.29	2.22	2.60	1.01	1	1.28	1.57
Technology	1.30	3.38	2.76	0.98	0.78	1	1.15
Humanities	0.98	1.60	1.88	0.66	0.64	0.87	1

Notes: Table shows relative frequency of behaviour incidents during different subject lessons, based on known timetable frequency of subject by yeargroup and school. Numbers represent the ratio of incidents by subject, with the column heading giving the numerator and the row heading the denominator, e.g. there are 2.23 as many incidents recorded during maths lessons than during English lessons, and 0.66 as many incidents in performing arts as in humanities lessons. Performing arts includes dance, drama and music; humanities includes history, geography and religious studies. Other subjects were not included because of small sample sizes (fewer than 100 incidents per yeargroup and school), or because of lack of information about timetable frequency. Subject information was available for 23,460 incidents. Ratios were calculated by school and yeargroup, then averaged. Note that this does not take into account the day of the week or the time of day when the lesson was scheduled. The following broad categories of incident are included in the calculation: assaults/fighting; defiance; disruption; dangerous or destructive behaviour; and verbal offence. This is to avoid the problem of certain incidents being more relevant to some subjects than others, e.g. forgetting kit will occur in PE but not English, and some incidents are not obviously disruptive, e.g. incorrect uniform.

Table 4.21: Spearman rank correlation coefficients of behaviour incident rates by subject
N pupils: 3,287

	English	Maths	Science	Performing Arts	Physical Education	Technology	Humanities
English	1						
Maths	0.5208	1					
Science	0.5326	0.5073	1				
Performing Arts	0.3832	0.4150	0.3941	1			
Physical Education	0.3931	0.4054	0.4242	0.3696	1		
Technology	0.3894	0.3952	0.4253	0.3014	0.3678	1	
Humanities	0.4563	0.4678	0.4705	0.4397	0.3564	0.3652	1

Notes: Tables present Spearman rank correlation coefficients for pairwise comparisons between behaviour incidents per pupil per lesson by subject. In every comparison $p<0.0001$, including when Bonferroni adjustments are made for multiple testing.

Chapter 5: Day and time patterns in behaviour at secondary school

Introduction

Rich countries have spent several decades increasing the resources provided to state schools, but this has largely failed to produce improvements in educational attainment, even when quality variables are taken into account (Hanushek, 2003; Wössmann, 2003). Public spending on education in the UK increased by over 5% per year in real terms during the 2000s, but is likely to decline substantially to 2014-15 for may schools (Chowdry & Sibieta, 2011; DfE, 2013). Improving the efficiency of time spent at school could therefore be an important means to improving attainment, particularly if this does not involve increased spending. One potential channel is through improving pupil behaviour at schools, which has been a focus of the UK coalition government's education policy (DfE, 2011). Misbehaving pupils not only affect their own progress, but also impose negative externalities on other pupils, taking up teacher time and reducing the time that can be spent on learning (Lazear, 2001). Imposing discipline is also time consuming and unpleasant for teachers, who may leave schools with poor behaviour. Pupils with poor behaviour may also induce other pupils to behave poorly, further reducing the effectiveness of lesson times.

I present evidence of variation in pupil behaviour in secondary schools by time of day and day of the week. Differences in educational performance are usually analysed in terms of variation between individuals, schools, and countries. When differences within individuals through time are examined this is over long periods, e.g. on the development of cognitive and noncognitive skills over the life course (Borghans, Duckworth, Heckman & ter Weel, 2008). Nevertheless, wellbeing and other psychological variables vary over the week and during the course of the day, so it is plausible that educational productivity might vary too. Investigating behaviour incidents in schools is interesting for a number of reasons. First, behaviour incidents are an indirect indicator of educational productivity: disruptive and problematic behaviour in the classroom makes teaching and learning more difficult (Lazear, 2001; Lavy, Passerman, & Schlosser, 2012). Moreover, quite aside from any impact pupils' behaviour might have on their peers, pupils' tendency to misbehave may be an indication of their own capacity to learn at a point in time, both because this is a direct measure of not being engaged in educational activities, and because less engaging lessons may be causing pupils to misbehave more. We can therefore assume that behaviour incidents will be negatively related to educational productivity at an individual and a class level. Second, behaviour at school is an interesting measure of noncognitive skills. It has been shown to be related to the possession of other noncognitive traits such as self-control (Coolidge, DenBoer & Segal, 2004; Petrides et al., 2004). The context and incentives pupils face, and the sort of behaviour valued, are also similar to those in the workplace (Segal, forthcoming; Bowles, Gintis & Osborne, 2001; Roberts, Harms, Caspi & Moffit, 2007). Thus observed daily and weekly patterns of behaviour within schools may be indicative of patterns also present in workplaces, and may therefore be informative on productivity during the working week. Much of the literature on time of day and day of the week effects relies on factory settings where output can be easily measured, but the school setting might better approximate productivity effects in the sort of office-based or service jobs which dominate developed economies. Third, behaviour can be seen as an indicator of wellbeing: poor behaviour in adolescence is often related to psychological distress (Dodge, 2006). Fourth, behaviour incidents may provide a better measure than is available in much of the literature on productivity over the week, as it is less subject to endogenous selection and observation.

This is the first paper to demonstrate a link between pupils' behaviour incidents and scheduling patterns. Day of the week effects have been observed in a range of areas of human activity, including labour productivity (Yao, Dresner & Zhu, 2010); injury rates (Card & McCall, 1996); absence from work (Ichino & Moretti, 2009); cardiovascular mortality (Evans et al., 2000; Chenet, McKee, Leon, Shkolnikov, & Vassin, 1998; Willich et al., 1994; Anson & Anson, 2000); stock market returns (Pettengill, 2003); job satisfaction (Taylor, 2006); and wellbeing (Csikszentmihalyi & Hunter, 2003; Bryson & MacKerron, 2012; and for a review of the field see Bryson & Forth, 2007). These tend to find that Mondays (or the first day of the working week, which varies across countries) are worse on many dimensions: productivity is lower, absence is higher, injury rates are higher, stock market returns are lower, subjective wellbeing is lower, and cardiovascular mortality is higher. Time of day effects have also been observed: injury rates at work are higher at night (Fortson, 2004); while subjective wellbeing is low at the start of the day and rises to reach a peak in the evening (Csikszentmihalyi & Hunter, 2003; Stone et al., 2006; Bryson & MacKerron, 2012). However, it is difficult to infer causality from this literature for a number of reasons.

First, much of the research investigating the impact of scheduling on productivity and affect suffers from incomplete data or selection bias: with survey data, people may be able to choose when to respond, so the day and time of the observation may be endogenous. For instance, we would expect that the days on which respondents were available for a survey interview to be related to their working patterns (Taylor, 2006). My dataset covers every day when a school is open, so there is little selective reporting of this kind: pupils cannot choose whether their incidents are recorded. Second, many papers investigating time and day effects are unable to disentangle the impact of time or day from the impact of activities typically performed at those times. Activities and conditions of work may vary over the day or week (e.g. there may be a higher workload at the beginning of the week), and controlling for activities typically attenuates the observed time patterns (Stone et al., 2006; Csikszentmihalyi & Hunter, 2003). Relatedly, workers or firms may already be organising the working day around employees' preferences or capacities, which would lead to an underestimate of the impact of time and day on productivity. Workers may also have selected into firms based on their time preferences. By comparison, pupils in my dataset have no control over their timetables; there is no variation in timetable or workload through the school week; and during compulsory schooling pupils have little choice over what subjects they study, who teaches them, and who their classmates are.[140] They are also unlikely to have had much choice over the school they attend.[141] Thus in most cases the only margin on which they can affect their activities is through absenting themselves altogether. This makes the measurement context importantly different from most workplaces, where employees have sorted into that specific employment (and been selected by employers); workload and tasks may differ through the week; and workers may have discretion over the hours they choose to work (even if this is simply a choice amongst shifts). Likewise, once teachers have chosen to work full time or part time and timetables have been set, they will not have much discretion over the lessons they teach. There is plausibly little endogeneity in the timetabling of subjects: schools need to staff every lesson in the week, and block timetabling by yeargroup and department makes strategic manipulation of timetabling difficult. In any case, the patterns I observe hold for each subject

[140] In these schools, pupils have no choice over the subjects they study aged 11-14 and limited choice at 15-16.

[141] There is likely to be strong selection into schools by parental characteristics, but I provide evidence in Chapter 4 that the schools for which I have data are fairly typical of state secondary schools in England (Table 4.1).

considered individually, so it is unlikely that subject scheduling is driving the overall results. We can therefore take the estimates of time and day effects as evidence of the impact of the time slot itself, rather than of the subjects being taught. Third, it may be difficult to compare different times of day in other contexts because of the different durations of activities and timing of breaks. In my data I can compare the rates of behaviour incidents across lessons because I know that pupils are engaging in similar activities in equally long blocks of time. Finally, I have associated data available which allows me to conduct robustness checks and explore the causal mechanisms for the patterns I find. Thus my identification of day of the week and time of day effects may be more robust than those reported in other contexts.

I am therefore addressing the following question:

Is pupils' behaviour affected by the time of day and the day of the week?

I use a dataset of behaviour incidents in four schools over six academic years to examine the impact of scheduling on the number of behaviour incidents. I find that pupil disruption differs by day of the week and time of day, with Mondays having significantly more behaviour incidents than later days, and the last lesson of the day having significantly more incidents than earlier lessons. There are a number of possible explanations for this pattern, which I investigate with the data available, including a companion dataset on pupils' psychological state collected for the UK Resilience Programme Evaluation. I consider the following explanations:

1. Pupil wellbeing varies by time of day and day of the week, and this is reflected in differences in behaviour.
2. Conditional on being in school, pupils behave uniformly across the week and the day. However, pupil lateness, absences, and exclusions vary by day of the week, and this drives the impact on behaviour incidents (composition effects).
3. There is no actual difference in behaviour over the week, but teachers' ability or willingness to report incidents varies by day and time, and this reporting difference generates the patterns I see.
4. Teacher wellbeing has a strong weekly and daily pattern which affects their ability to engage pupils and control classes, and this generates differences in pupil behaviour through the week.
5. Teacher absences have a strong weekly and daily pattern; pupils are more likely to misbehave with supply teachers than with regular staff; so this generates weekly and daily differences in the behaviour incident rate.

6. Schools do not schedule subjects arbitrarily (endogenous timetabling), so the patterns I observe relate to other factors occurring over the school week, e.g. the hardest subjects or the weakest teachers are usually scheduled on Mondays or the last lesson of each day.

Note that these explanations are not mutually exclusive, and several could be contributing to the pattern I observe. I provide evidence that pupils' wellbeing varies across the week, so it is plausible that their behaviour might vary too. I suggest that it is unlikely that the patterns are entirely due to differences in reporting or endogenous scheduling, but that pupil absence and lateness, teacher and parent wellbeing, and teacher absence may all contribute to the observed pattern. I suggest that pupil behaviour and academic attainment could be improved by reorganising pupils' schedules. Specifically, the number of incidents could be reduced by exploiting differences in the incident patterns and rates over the week for different subjects. I cannot show that this would increase academic performance, but an improvement in behaviour would be valuable even without an accompanying improvement in attainment. Overall, my findings suggest that modifying the organisation of time within schools could be an important way of improving conditions for staff and pupils and might increase attainment.

The structure of the paper is as follows. First, I briefly set out the policy background and concerns about pupil behaviour in schools, and summarise the data I am using. I described the schools for which I have data and their behaviour management systems in the previous chapter (Chapter 4). I then outline the structure of the school year and the school day, and present descriptive statistics for the sample of pupils for whom I have behaviour data. Next I present figures and tables showing strong variation in behaviour incident rates by time of day and day of the week. Next I investigate several possible explanations for these patterns, before discussing my findings and placing them in the context of the literature on scheduling effects. I conclude that modifying timetables could produce real improvements in pupil behaviour.

Behaviour in English schools

Whether there are gains to be had from improvements in behaviour in English schools depends on whether it is currently a problem. Media reports of school-based behaviour paint a grim picture of rampant dangerous behaviour, but these are probably not representative.[142] A survey of headteachers from G8 countries found classroom disturbance to be the most frequently occurring behaviour problem across countries, with 54% of pupils in England attending schools with weekly or more frequent classroom disruption, and 23% with intimidation or verbal abuse from other students at least weekly (Miller et al., 2009). A survey by a UK teachers' union found that two-thirds of teachers had had to deal with a violent pupil in the current academic year, with the majority of respondents saying that they thought behaviour had got worse over the previous 5-10 years (ATL, 2012). Two-thirds of secondary school teachers responding to the 2012 Teacher Voice Omnibus survey said that they believed negative pupil behaviour was driving teachers out of the profession (NFER, 2012). However, 84% of secondary schools are rated as having good or outstanding standards of behaviour (Ofsted, 2012b), and the most common form of poor behaviour is low-level disruption of lessons, with serious violence very rare and carried out by a small number of pupils (Ofsted, 2005; DfE, 2012a; ATL, 2012). Nevertheless, such disruption can be irritating and can obstruct learning. The impact of poor behaviour on a pupil's own performance is difficult to measure. Gutman and Vorhaus (2012) use a cohort study which allows them to control for a range of pupil characteristics, finding that pupils who misbehave frequently achieve 178.80 points lower at age 16 then those with no behaviour problems, equivalent to three extra GCSEs at grade A* - a substantial difference, although even rich controls will not account for the endogeneity of misbehaviour and attainment.

Data and sample

The behaviour data came from four schools involved in the UKRP evaluation. I described the recruitment of LAs and schools to the UKRP evaluation in Chapter 1 ('Context and recruitment'). Chapter 4 described the schools which provided the behaviour data, and how they compared to all schools in England (Table 4.1; this is important for the external validity of my results). I also outlined the behaviour management systems of these schools ('Behaviour management'). I have detailed behaviour incident data available for four schools. The behaviour data and accompanying psychological data available are the same as described in Chapter 4, where I provided a detailed account of the types of incidents (e.g. Table 4.2). In addition, I obtained exclusions data from the NPD, which detail the start date, length of exclusion,

[142] See, for example, "NINETY primary pupils sent home every day for attacks in class: Shocking figures reveal rising school violence", Daily Mail Online, 25th July 2012.

and reason for exclusion for each pupil excluded between 2005/06 and 2009/10. One piece of information I do not have for the four schools with behaviour data is daily absence data. I do have this information for a fifth UKRP school with a similar timetable and similar demographic characteristics to the four behaviour schools. I also obtained schools' term dates and daily timetables from their websites.

Descriptive statistics

The School Year

Local Authorities are responsible for setting term and holiday dates for most schools[143]. In practice, term dates match up closely across England. Most schools start the year in early September and finish in mid-July, dividing the year into three terms. The autumn term runs from September to mid-December, with a week-long holiday at the end of October; the spring term starts in early January and usually ends at Easter, and there is a one-week half-term holiday in mid-February. The summer term usually starts after Easter and finishes in mid-July, with a one-week half-term holiday in late May. The summer holidays last roughly six weeks.

Table 5.1 describes the dates and academic years of data available. I have a panel of 6 academic years of behaviour data, from 7th September 2005 to 7th April 2011 inclusive. Each school has data available for a different range of dates. We can see from Table 5.2 that most academic years have roughly the same number of school days as the legal requirement of 190 days – where there are fewer this is due to missing data (e.g. I do not have complete data for 2010-11), and where this is higher this is due to the four schools having slightly different holiday or training dates: no school had more than 193 pupil days in an academic year. The autumn term is on average 10 days longer than the summer term, and 12 days longer than the spring term, because of the timing of holidays.

The number of days of the week also reflects the school timetable (Table 5.3): there will be no Saturdays and Sundays since pupils at these schools never go to school at weekends. There are fewer Mondays and Fridays in the dataset, because these days are more likely to form part of holidays or teacher training days on which staff attend school but pupils do not. Because there are no schooldays at weekends, every Friday will be before a gap of at least two days, and every Monday will come after a gap of at

[143] Specifically, for community, special and voluntary controlled schools.

least two days. Gaps of three days or more could be due to bank holiday weekends, additional staff training days, or school holidays.

The School Day

There are no legal requirements on the length or structure of the school day (DfE, 2012e). The governing body of each school decides the length and timing of sessions and of breaks. However, it is common to have lessons lasting approximately one hour. In the UKRP evaluation, 18 of 22 schools had lessons of one hour duration, with another two having lessons of 50 minutes. The remaining two schools had 80-minute lessons. The structure of the school day in the four schools I have behaviour data for is shown in Table 5.25. All four follow a similar structure of five hour-long lessons per day, with a 20-minute break after lesson 2 and a 40-50 minute lunch break after lesson 4. They also start and end at similar times.

Pupils

Table 5.4 gives details of all pupils in the dataset. There are 3,284 pupils in total. There is an average of 13.9 incidents per pupil during the time they are covered in the dataset, over an average period of 451.8 school days. Thus I have information on pupils' behaviour for an average of 2.38 academic years, with an average of 0.037 incidents per pupil per day. The full five years of compulsory secondary schooling would cover approximately 950 days (190 days per year over five years), so it is clear that the dataset does not cover the whole of secondary school for the majority of pupils.[144] Half of pupils are male; 20% have SEN; and in September 2009 they were aged between 11 and 20, with a mean age of 14.6, although pupils will only be included in the dataset between the ages of 11 and 16 as education beyond Year 11 (age 16) is not compulsory.[145] 30% of pupils have been eligible for FSM at some point in the last six years, and the average percentage of children living in poverty in pupils' neighbourhoods is 29%. 88% of pupils come from white backgrounds, with 5% from Pakistani or Bangladeshi backgrounds, and small numbers of pupils from other ethnic backgrounds. I show elsewhere that pupil and school characteristics for these four schools are broadly typical of non-selective secondary schools in England (Table 4.1). The bottom four rows give

[144] This could call into question my conclusions about the age at which pupils are most likely to be disruptive and where the greatest gains in rescheduling could be found. However, so long as adjacent cohorts of pupils are good counterfactuals for each other, as I have argued elsewhere, the data available should give a good idea of behaviour through secondary school.

[145] For example, pupils who were aged 16 and in Year 11 at the start of panel in September 2005 will be aged 20 by 2009, but will have left school (and the dataset) in July 2006.

values for additional data from the UKRP dataset on self- and teacher-reported behaviour, anxiety and depression scores, which are available for a subset of pupils (see Chapter 1 for information on these measures).

Scheduling and behaviour incidents

Behaviour incidents and the Day of the Week

Table 5.5 presents the number and rate of behaviour incidents by day of the week, along with the broad category of incident, for those incidents which can be assigned a specific date.[146] The first row gives the number of pupil x date observations: this is the number of pupils on roll at a school multiplied by the number of days of the school term. There are more pupil observations in the middle of the week (Tuesdays-Thursdays) than on Mondays and Fridays, because the latter are less likely to be school days.[147] The fourth row gives the total number of incidents by day. Here we see that there are more incidents recorded on Mondays than on Tuesdays, with fewer incidents on each subsequent day and the lowest number on Friday. This is despite there being fewer Mondays in the dataset. The second row gives the rate of incidents per pupil per day (the number of incidents divided by the pupil x date observations for that day): here we see that there is a higher rate of incidents on Mondays and that the rate declines through the week. The lower part of Table 5.5 shows that this pattern of fewer incidents later in the week holds for most categories of incident, although Monday is not the worst day for every category.

This pattern is shown clearly in Figure 5.1, with Wednesday-Friday having significantly fewer incidents per pupil than Monday and Tuesday. Wednesday, Thursday and Friday appear to have roughly the same rate of incidents per pupil, with overlapping confidence intervals. Figure 5.2 plots the number of behaviour incidents per pupil over the course of a half term, i.e. consecutive school weeks uninterrupted by a holiday. The first Monday in the half term is day 1 so that the first full 5-day week of the half term covers days 1-5; if a half term starts on a day other than a Monday these are coded as 0 (Friday), -1 (Thursday), -2 (Wednesday), and -3 (Tuesday), so that we

[146] Based on the categories and descriptions of incidents, the behaviour incidents which could not be assigned to a single date include: persistent lateness; persistent no equipment; forging notes or failing to inform parents; missed detention; and no homework and persistent lack of homework. This is because they will have happened over a period of several days or weeks (they represent an accumulation of incidents) or at a time other than the one recorded. These total 825 incidents, or 1.8% of incidents, and excluding them does not alter the results I obtain.
[147] Note that this number does not take into account a pupil's attendance at school: I do not know whether pupils go to school on a given day, only whether they are on roll and the school is open for pupils.

can see the development of the behaviour incident rate over days of the week and over weeks of the half term. Here we can clearly see the Monday-Friday pattern in the data, with most weeks having more incidents on Mondays with the rate then dropping off. We can also see an inverted u-shaped pattern in the incident rate over the course of the half term, with the first few days and the first full week of term having a lower incident rate than weeks 2-6, with the rate then dropping into weeks 7 and 8.[148] Both the day of the week and the week of term appear to matter: the first Monday of term has about as many incidents per pupil as the Friday of the fourth week of term.

Table 5.6 examines these patterns econometrically. The first column reports results from OLS regression with the number of incidents a pupil had on a particular day *1000 on the left hand side, and dummies for Tuesday-Friday on the right hand side. This specification contains no control variables. Here we see a strongly statistically significant reduction in behaviour incidents on Tuesdays relative to Mondays, with a larger decrease for Wednesday-Friday. Column 2 adds in dummies for days which were either the first school day after a gap of 4 or more days, or the last day before such a gap – these could be special days with non-standard activities, and could be generating the pattern we see for Mondays and Fridays. It also includes controls (dummies) for the week of half term (relative to week 1), the school, half term, and half term*school (to allow time effects to vary by school, as we might expect they would), as well as pupil fixed effects to estimate the within-pupil difference in behaviour incidents across the week. Adding these leaves the coefficients on Wednesday-Friday substantially unchanged, but we also see that the first day after a gap of 4 or more days reports 0.0053 fewer incidents per pupil (4.8 fewer incidents per day in a school with 900 pupils), and the last day before a gap has 0.0076 fewer incidents per pupil (6.9 fewer incidents). Weeks 0, 7 and 8 appear to have significantly fewer incidents than week 1, while weeks 2-4 have significantly more.

I start by using OLS for ease of interpretation; to demonstrate robustness to using different estimation techniques; and to enable me to present a specification with pupil fixed effects whilst including all observations in the dataset. However, the data are count data, and are probably better analysed using negative binomial regressions (the standard deviation of the outcome variable is much larger than the mean, so I cannot use

[148] Note that only the longest half terms, always in autumn, will last as long as 8 weeks; 49% of school-half term observations in this dataset have between 30 and 35 schooldays, or 6-7 weeks, and only 17% have more than this.

Poisson regressions). Column 3 includes the same variables as column 1, but uses negative binomial regression rather than OLS. The coefficient, standard error, and the incidence rate ratio (IRR) are reported for each day of the week. Again, all the coefficients on Tuesday-Friday are significantly negative, and the IRRs show that Tuesday has an incident rate about 0.90 that of Monday; while Wednesday-Friday have IRRs between 0.76 and 0.78. Column 4 adds dummies for days before and after gaps and for the week of term, and includes controls for pupil characteristics: dummies for gender, FSM entitlement, SEN, Key Stage 2 mean score (mean of English, maths and science scores in national tests at age 11); deprivation level of the pupil's postcode (IDACI – percentage of children in the pupil's postcode area living in poverty); ethnicity (five categories); and month of birth (relative age within each cohort). I include these characteristics because they are likely to be related to pupil behaviour (see Chapter 4). Because these characteristics are not available for the full sample, this reduces the sample size and the number of pupils in the regression, but does not change the results. Column 5 includes the same variables but estimates a random effects specification, which again leaves the coefficients on the day of the week unchanged.

There are probably two main mechanisms generating zeros in the data: pupils who are not involved in any behaviour incidents; and pupils who are absent from school.[149] For this reason, zero-inflated negative binomial regressions may give a better estimate of the impact of day of the week on behaviour incidents. Column 6 includes the same regressors as columns 1 and 3 (no control variables), but uses zero-inflated negative binomial regressions, using gender, SEN status, FSM eligibility, and mean Key Stage 2 score to predict zeros caused by absences. This makes very little difference to the coefficients and IRRs on the day of the week variables. Column 7 adds in pupil characteristics and week of term dummies, again making little difference to the coefficients on the days of the week.

Taken together, we can see that the impact of the day of the week on the rate of behaviour incidents is robust to a number of different specifications and approaches, and remains highly significant. This gives a Friday incident rate about 0.75-0.78 the Monday rate. In a school of 900 pupils, this would translate to 7.1 fewer incidents on Fridays relative to Mondays, with the rate on Mondays standing at about 32.4 incidents per day.

[149] I do not have daily absence data for these schools so cannot distinguish between absent pupils and well-behaved pupils.

Heterogeneity in the impact of day of the week

Heterogeneity by pupil characteristics. Figure 5.3 plots the rate of behaviour incidents over the half term as in Figure 5.2, but splits the sample by gender. Figure 5.4 and Figure 5.5 do the same, splitting the sample by SEN status and FSM eligibility respectively. Here we see that although these groups have different levels of behaviour incidents, with boys, SEN pupils and FSM pupils having higher rates of incidents, the patterns across the week and the half term are similar: all pupils, even non-SEN pupils who have low average rates of behaviour incidents, show a decline in behaviour incidents through the week and an inverted u-shape over the half term. Table 5.7 models these differences using zero-inflated negative binomial regressions with the same specification as column 7 of Table 5.6, but with the sample split by demographic characteristics. Column 1 gives the results for boys, column 2 for girls. Here we see that boys and girls have similar IRRs on Tuesdays relative to Mondays, but that boys experience a slightly greater decline on Wednesdays-Fridays. Columns 3 and 4 compare pupils who have special educational needs with pupils who do not. Pupils without SEN show a greater relative decline in incidents on each day from Tuesday to Friday, although as we can see from

Figure 5.4, their absolute decline in incident rates is smaller. Columns 5 and 6 compare pupils entitled to free school meals with those who are not. Pupils without FSM show a greater relative decline in incidents, but pupils with FSM have a greater absolute decline through the week. These results suggest that the weekly pattern of incidents is similar for all these groups of pupils, both those who have high rates of incidents and those with low rates.

Heterogeneity by yeargroup. Figure 5.6 presents the mean number of behaviour incidents per pupil per day by day of the week and by yeargroup. Here we see that Year 10 has the highest rate of incidents per pupil per day, followed by Year 9, then Year 11. Years 7 and 8 have very few incidents. Yet the pattern of fewer behaviour incidents later in the week appears to hold for all yeargroups. Table 5.8 reports the day of the week regression results by yeargroup, again using zero-inflated negative binomial regressions with the same specification as column 7 of Table 5.6.[150] Here we see that all yeargroups have a statistically significant reduction in incidents on Fridays relative to Mondays; and that most also show reductions on Wednesdays and Thursdays. Only

[150] The number of pupils sums to more than 3,284 because almost all pupils are included in the dataset for more than one academic year, so they will appear at least twice in different yeargroups.

Year 9 and Year 10 have statistically significant reductions on Tuesdays. These results suggest that the pattern of fewer behaviour incidents later in the week is common to all five yeargroups, although the mean incident rate, the magnitude of the reduction through the week, and the day with the fewest incidents varies.

Heterogeneity by school. Table 5.9 presents regression results separately for each school, using zero-inflated negative binomial regressions with the same specification as column 7 of Table 5.6. Here we see that there is some variation in the magnitude of the decline in incident rates over the course of the week, and in the day which appears to have the fewest incidents, but that in all four schools Fridays have significantly fewer incidents than Mondays; in three of four schools Wednesdays or Thursdays have significantly fewer incidents, and in two schools Tuesdays also have fewer incidents. Thus the overall pattern of fewer incidents later in the week seems to hold for each school separately.

Overall, there is little heterogeneity in the pattern of fewer incidents later in the week, with both high- and low-incident groups, and all schools, showing a similar pattern. However, the magnitude of the reduction may vary by group, and the day with the lowest incident rate may also vary.

Behaviour incidents and the Time of Day

Table 5.10 presents the number and rate of behaviour incidents by type and time of day. Information on time of day is available for two schools, which reduces the sample to 14,634 incidents. The time of day is split into the five lessons of the day, and 'other times': the latter includes morning and afternoon registrations, break and lunch times, the transitions between lessons, and before and after school. It also includes a small number of observations which could not be assigned to a single time of day, because they were recorded as happening 'all day', 'generally' or in multiple lessons. Here we see that the first lesson of the day has the fewest incidents (1,455 incidents, or 0.0042 incidents per pupil), and the last lesson the most (2,568 incidents, or 0.0074 incidents per pupil), and this holds for most of the categories of incident, notably disruption and defiance. It is also noticeable that the number of incidents occurring at other times of day is higher than that of lesson 5. This is partly because all lessons have roughly the same length – lasting about one hour – while all other times of day aggregated constitutes more than two hours, so at a constant rate of incidents per hour we should expect more than twice the number of incidents in 'other'.[151] Dividing the number of incidents at other times by two, we get 2,079 incidents, or 0.006 incidents per pupil outside of lessons each day: roughly equal to the rate of incidents during lessons 2 and 4. The variation in the type of incident by timing is also worth noting: disruption is much lower at 'other' times of day (presumably because there are fewer organised activities to disrupt), while two-thirds of lateness and truancy happen during other times of day (mainly morning and afternoon registration). 36% of dangerous behaviour or damage and 51% of assaults and fighting happen outside of lessons. The overall pattern is shown in Figure 5.7, which shows mean behaviour incidents per pupil per lesson by time of day, along with 95% confidence intervals showing that both lesson 1 and lesson 5 have significantly different incident rates from lessons 2-4.

Table 5.11 gives the mean number of incidents per pupil per time period, by time of day and day of the week. The number of pupil x time period observations is the same for every time of day because during compulsory schooling a pupil will not have free lessons during the day.[152] The bottom row gives the average across the week for each time period. The last column shows the strong effect of day of the week on the number of behaviour incidents, with Mondays having more incidents than later days. However, the time of day pattern is not the same across all days of the week: notably, the worst lesson on Mondays is actually lesson 3, not lesson 5, and the worst lesson on Fridays is lesson 2. This is shown in Figure 5.8, which plots the mean number of incidents per pupil per lesson across the 25 lessons in a week, excluding non-lesson times (Monday lesson 1 = 1; Monday lesson 2 = 2; Tuesday lesson 1 = 6, etc.). The first lesson of each day is marked by a diamond. Here we can see two things. First, the day of the week pattern is evident, even though we are only using two schools of the four schools: Mondays and Tuesdays have more incidents per pupil than Wednesday-Friday. Second, there is an upward trend during the course of the day. The pattern is not entirely consistent across days of the week, but the first lesson of the day always has the fewest incidents.

Table 5.12 presents similar specifications to Table 5.6, but with the addition of time of day variables, and using only the two schools for which I have this information.

---

[151] Since 'other times' includes the categories 'before school' and 'after school' it is not possible to say exactly how much time this would cover.

[152] There may be a few half days which have not been taken into account here, e.g. on the last day at the end of the autumn or summer term. This would tend to reduce the rate of behaviour incidents in lesson 5, the last lesson of the day, which would bias against my finding that lesson 5 has the most incidents.

We can see that lessons 2-5 all have significantly more behaviour incidents than lesson 1 (the comparison group), regardless of the specification used. The day of the week pattern is also visible.[153] Interestingly, the addition of time of day variables into this regression makes no difference to the strength of the day of the week pattern (specifications not shown). Moreover, the addition of controls does not substantially affect the size or significance of the coefficients on the time of day variables. The largest difference in hourly incident rates is between lesson 1 and lesson 5: there are 1.77 times as many incidents during lesson 5 as during lesson 1, or, in a school of 900 pupils, 6.7 incidents during lesson 5 and 3.8 during lesson 1.

Heterogeneity in the impact of time of day

Heterogeneity by pupil characteristics. Figure 5.9 plots the mean number of incidents per pupil per lesson over the 25 lessons in a week as in Figure 5.8, but splits the sample by gender. Figure 5.10 and Figure 5.11 split the sample by SEN status and FSM eligibility respectively. Despite the differences in the overall incident rate, we can see that the pattern across the week is similar for all groups, although it is more marked in the groups with higher rates of behaviour incidents (boys, SEN and FSM pupils). For all groups there is an increase over the day, and a reduction over the week. Table 5.13 models these differences using zero-inflated negative binomial regressions with the same specification as column 7 of Table 5.12, with the sample split by demographic characteristics. Here we see that boys have a relatively greater increase in incidents through the day than girls; pupils with SEN have a relatively greater increase than pupils without SEN; and FSM pupils show a relative increase the same or smaller than pupils without FSM, depending on the lesson (although we can see from Figure 5.11 that the absolute increase is larger for FSM pupils). This suggests that the daily pattern of incidents is similar for all these groups of pupils, even those with low incident rates, although the magnitude of the growth rate may differ.[154]

---

[153] The coefficients in the OLS specifications are smaller than when the full sample is used because the outcome variable is now the rate of incidents per pupil per lesson, rather than the rate of incidents per pupil per day.

[154] I could adjust the p-values for multiple testing here. However, since I am primarily concerned with identifying any possible differences there might be due to scheduling rather than ensuring that I only identify highly significant effects, I do not use this. That is, I am less concerned with false positives than with false negatives, as this analysis is preliminary and should be viewed as exploratory work on which to base an RCT, not the final word on the impact of scheduling. In addition, since there appears to be little or no difference in the pattern between pupils with different demographic characteristics; increasing the p-values would therefore not change the results.

Heterogeneity by yeargroup. Figure 5.12 shows the mean number of behaviour incidents per pupil per lesson by time of day and yeargroup. Years 9 and 10 have the highest rates of incidents overall, and appear to be driving the pattern of more incidents later in the day. Year 11 pupils actually have fewer incidents in lessons 2 and 4 than in lesson 1, with lessons 3 and 5 having the highest rates. Years 7 and 8 again have low levels of incidents. Lesson 1 has the lowest incident rate for all yeargroups, but otherwise it is difficult to detect a clear pattern. Table 5.14 presents regression results by yeargroup, again using zero-inflated negative binomial regressions with the same specification as column 7 of Table 5.12. The pattern is messier here than for the full sample, but for all yeargroups lessons 5 and 3 have a higher rate of incidents than lesson 1, and lesson 4 has a higher rate for all yeargroups except Year 11.[155] The rate of incidents at 'other times' is much higher for Years 10 and 11, and most of this is due to lateness at registration. Because this analysis is only run on two schools, and these have a shorter panel than the full sample, the analyses by yeargroup may be less reliable: rather than having the same pupils in different yeargroups over a period of up to five years, many of these results come from different pupils in the same school, which can be problematic if cohorts differ. Nevertheless, I find the same overall pattern of more incidents later in the day.

Heterogeneity by school. Table 5.15 presents regression results separately for each school. Both schools show a general pattern of increasing rates of behaviour incidents through the day. The largest difference is in the rate of incidents at 'other times'. This is due to School 1 recording lateness in the behaviour dataset, which accounts for 49% of the incidents which take place at other times in this school (almost all during morning or afternoon registration), while School 3 rarely records lateness in this way. Excluding lateness makes the two schools' results look more similar. Overall, both schools appear to have the same pattern of increasing rates of behaviour incidents through the day.

Thus the increase in behaviour incidents through the day appears to hold for different groups of pupils, different age groups, and for both schools. However, there are differences in the magnitude of the increase for each lesson.

[155] Although note the small number of Year 11 pupils for whom data is available.

Mechanisms and Robustness Checks

There appear to be strong daily and weekly patterns in behaviour incidents. Are these due to genuine differences in behaviour? There are a number of possible explanations for these patterns, and I will present evidence as to the plausibility of several of these.

Psychological state by day of the week

Pupils' behaviour over the week may reflect differences in their wellbeing. In this section I use a companion dataset collected for the UK Resilience Programme Evaluation to show that pupils' psychological wellbeing also varies by day of the week, which suggests that the measured behaviour pattern is real. Figure 5.13 presents a bar graph with 95% confidence intervals for the mean self-reported behaviour score by day of the week, measured using the child-reported Goodman Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). Figure 5.14 does the same for depression scores, measured using the Children's Depression Inventory (Kovacs, 2003), and Figure 5.15 presents anxiety scores, measured using the Revised Children's Manifest Anxiety Scale (Reynolds & Richmond, 1985). All these measures are pupil-reported, using questionnaire inventories with at least 20 items, which are scored and summed to give a scale measure (see Chapter 1 for details). For all three of these measures a higher score indicates more or more severe symptoms. The exact pattern of symptoms across the week varies slightly between measures, but in all three cases the score on Monday is significantly higher (worse) than on Wednesday and Thursday, and usually Friday too: pupils appear to be more depressed, more anxious, and report behaving worse on Mondays than later in the week. This is a very similar pattern to that observed in the behaviour incidents data. As a comparison, Figure 5.16 presents responses to another question asked in the questionnaire with little emotive content: 'how many bedrooms do you have at home?'. This is not something we would expect to vary across the week unless pupils with different characteristics were being surveyed on different days, and indeed, there are no significant differences in responses by day of the week.[156]

These relationships are explored in regressions similar to the ones reported above for the behaviour incidents, but using OLS rather than negative binomial regressions as

[156] The inventories for behaviour, depression and anxiety ask children to report on how they have been behaving or feeling over the past two weeks or the past month, but subjective wellbeing measures are known to reflect both short- and long-term factors (Kozma, Stone & Stones, 2000), so it is likely that their recall and perceptions will be influenced by how they are currently feeling or behaving.

the dependent variables can be modelled as continuous variables.[157] One difference is in the time variables used in the controls. This psychological dataset is a panel generated from annual surveys of pupils, so there is much less time variation in the available data. By contrast, the behaviour incidents data from school databases covers every day of the school year, so we have data for every day that pupils were registered at the school. Thus in the psychological dataset 69% of pupil-observations come from June or July, so although there is variation in the day of the week pupils completed questionnaires there is little variation in the time of year when they did so. I therefore control for the data collection period when the surveys were filled in, e.g. June-July 2009. This also means that any identified day of the week effects may be due to selective reporting – a common problem for other surveys which find differential happiness or productivity by day of the week (Bryson & Forth, 2007), but not a problem with the behavioural incidents dataset. One way of (partially) controlling for this is by restricting the sample reported in regressions to pupils who have at least two observations for the outcome reported, and who provided data on at least two different days of the week during the three years they were followed up. This means that the pupil fixed effects specification should estimate the impact of day of the week within pupils, controlling for the data collection period. Restricting the sample in this way does not substantially change the results, but does reduce the sample size by approximately 1,550 pupils and 3,000 observations. Moreover, the pupils who are excluded have self-reported behaviour scores, depression scores, and anxiety scores approximately 0.20 SD lower on average than those who are included (table not shown). So restricting the sample may generate some positive selection in the sample used, excluding pupils with the worst wellbeing scores, but it also allows a cleaner identification of the impact of day of the week. Despite these obvious deficiencies, the psychological data offers an interesting comparison to the behaviour database, particularly because of the similarity of the context: the four schools which provided behaviour incident data were drawn from the 22 schools which are featured in the psychological dataset, and 1,308 pupils from the behaviour incident database also appear in the psychological dataset.[158] Unfortunately there is no information available about the time of day when pupils completed the questionnaires.

[157] Running the regressions with Poisson or negative binomial regressions does not change the results.
[158] Other pupils from these four schools were not surveyed as part of the UKRP evaluation because they were not in the relevant cohorts.

Table 5.16 summarises the self-reported behaviour, depression and anxiety scores by day of the week for the restricted sample. Mondays have the highest (worst) mean score for each measure, although Wednesdays and Thursdays appear to be the best days of the week rather than Fridays. Table 5.17 reports the regression results for the self-reported behaviour score, standardised to have a mean of 0 and a standard deviation of 1. This shows that pupils surveyed on Tuesdays-Fridays report significantly better behaviour than those surveyed on Mondays. Column 1 reports the results without any controls; column 2 adds in pupil characteristics and controls for the data collection period; column 3 adds in school fixed effects (22 schools); column 4 uses class fixed effects; and column 5 includes data collection period controls and pupil fixed effects. Unlike the regressions on behaviour incidents, here the size and significance of the coefficients on the days of the week do change with the specification used: the day of the week effects are largest when no controls are included, and are reduced when school, class or pupil fixed effects are included in columns 3-5. This suggests that the day pupils completed the questionnaire was not random.[159] Nevertheless, even when pupil fixed effects are included in column 5, estimating the within-pupil difference in score on different days of the week, we still find a significant effect, with Tuesday and Wednesday showing a 0.06 SD reduction in behaviour scores relative to Monday; Thursday a 0.10 SD reduction; and Friday a 0.09 SD reduction. Table 5.18 presents the same specifications with the standardised depression score as the outcome variable. Here the effect of day of the week is smaller, but is still weakly significant for Tuesdays, Wednesdays and Fridays in some specifications, and highly significant for Thursdays. The size of the coefficients range from -0.046 SD for Tuesday, to -0.093 SD for Thursday.[160] Table 5.19 presents the same regressions for the anxiety score. Here we find an impact of day of the week of up to -0.077 SD on Thursdays relative to Mondays. As a placebo check, Table 5.20 presents the same results for the number of bedrooms pupils report having in their homes: there is no consistent relationship between day of the week and this variable, and all coefficients are very small.

These results from the dataset of psychological variables support the hypothesis that there is genuinely a difference between pupils' wellbeing and behaviour on

[159] It could be that including school and class fixed effects results in overcontrolling: most pupils completed the questionnaire on the day their class was scheduled to do so, so the day of the week will be associated with class assignment and school.
[160] The evaluation of the UK Resilience Programme found a short-run programme effect of 0.15 SD on depression scores (see Chapter 1), so these day of the week effects are sizeable relative to the measured programme impact.

different days of the week. Three aspects of these patterns in particular support the hypothesis. First, pupils' reports of their own behaviour differ by day, which gives me two separate data sources (school databases and pupil surveys) which suggest that pupil behaviour differs over the week. Second, the number of bedrooms pupils report having at home does not differ by day of the week, suggesting that this is not simply a reporting effect due to which pupils would respond more or less positively to all questions on particular days due to selection or non-randomness of the survey dates, but rather a response pattern which contains some information. Third, the patterns in the behaviour database and in these pupil surveys are very similar: Mondays are the worst days for behaviour incidents, self-reported behaviour, and self-reported depression and anxiety scores, with subsequent days showing an improvement, particularly Wednesday-Friday. This pattern also matches that found in other studies: Csikszentmihalyi and Hunter (2003) find that adolescent happiness is lowest on Mondays, rising through the week to peak on Saturday. This all suggests that pupils feel and behave worse on Mondays than later in the week. However, we do not know the mechanism – it could be due to differences arising from pupils, or could be due to teachers treating pupils differently and causing them to behave worse or be unhappy – both would be consistent with the observed patterns.[161]

Pupil lateness, absence and exclusion from school

A second possible driver of the observed behaviour patterns is pupil lateness, absence, and exclusion from school, primarily affecting pupil composition. If pupils are more likely to be late on Mondays and Tuesdays, and lateness is recorded as a behaviour incident or causes disruption, then we would expect to see more incidents earlier in the week. However, it is clear from Table 5.5 that although lateness and truancy do decline through the week, other categories of incidents such as disruption and defiance decline to a greater extent, so lateness is not driving the results. Similarly, lateness and truancy do not seem to be driving the time of day pattern: lateness and truancy rise through the day, but less so than incidents in other categories such as disruption and defiance (Table 5.10). Running the regressions shown in Table 5.6 and Table 5.12 but excluding lateness and truancy does not change the results (tables not shown).

---

[161] We should also consider reverse causality, e.g. the poor behaviour of some pupils could cause their classmates to be less happy on Mondays.

However, many incidents of lateness might be recorded in schools' attendance databases rather than in the behaviour database. If pupils are more likely to be absent earlier in the day or later in the week, this could result in lower rates of behaviour incidents at these times through composition effects.[162] Since I do not have daily data on pupil attendance, I cannot distinguish between pupils who are absent and pupils who are well-behaved, as neither will have any behaviour incidents. Absence rates are low overall (6.5% at state secondary schools in 2010-11, ranging between 6.9% and 8.1% in these four schools; DfE 2012d) but if the most disruptive pupils are absent this could have a large effect.[163] I do have daily absence data available for a fifth UKRP school with similar timetables and similar demographic characteristics to these four. Table 5.22 presents the fraction of pupils absent, late, or excluded for each registration session (two sessions per day: morning and afternoon), and Table 5.23 puts these into regressions. Here we see that absence is higher in the afternoon, and is higher on Fridays. Splitting this into authorised and unauthorised absence, we see that authorised absence follows the same pattern of overall absence, while unauthorised absence is also higher on Fridays but does not differ between mornings and afternoons. Lateness is much higher in the morning, and may be lower on Thursdays and Fridays. The fraction of pupils excluded for each session is about the same through the day and week. The final column of each table gives the daily exclusion rate for the four schools in the behaviour database for the same academic year.[164] The rate is higher than for the fifth school because the pupils are older, but again the exclusion rate looks constant through the week.

Could these patterns explain the pattern in behaviour incidents? Higher (authorised) absence rates in the afternoon would not explain the daily behaviour pattern, as more absences should reduce the afternoon incident rate through excess zeros and probable positive selection of pupils.[165] The absence pattern over the week also fails to account for the drop in incidents from Monday-Thursday. However, it could help explain the difference between the pattern from pupil-reported data (where Thursday is

---

[162] Pupils have a certain amount of discretion over authorised absences as well as unauthorised absences; it is parents who authorise absences, and many will do this even without good grounds such as illness.
[163] If behavioural incidents are subject to 'escalation', with one misbehaving pupil causing many others to misbehave, the effects of lateness or absence could be particularly large.
[164] Daily data on exclusions is available from the National Pupil Database.
[165] Indeed, Cortes, Moussa and Weinstein (2012) use the variation in 9th grade students' absence rates over the day as an instrumental variable for class size (and disruption): with 7 periods a day, they find that period 1 has the lowest attendance rate, and period 4 the highest, with a decline in attendance through the afternoon.

the best day of the week) and the behaviour pattern (roughly constant Wednesday-Friday): if Fridays are worse days than Thursdays, but more pupils are absent, the ‘false zeros’ in the behaviour data would bias down the Friday rate, while the pupil-reported measures would represent the wellbeing of the pupils actually present. Lateness in the morning could improve behaviour in the first lesson of the day. However, lateness to registration does not necessarily result in lateness to classes: the attendance data distinguishes between lateness to registration and lateness ‘after registers closed’ (i.e. arriving after the lesson begins), and the latter accounts for only 61 of 1,870 incidents of lateness. Thus even pupils who are late will usually be at school in time for lessons to begin. Higher rates of lateness in the morning are also partly offset by lower morning absence rates. There is also a lower rate of lateness on Thursdays and Fridays, which are the best days for behaviour. Exclusion rates are constant through the week in both samples. Moreover, dropping days when pupils were excluded and rerunning the main regressions on day of the week and time of day effects (Table 5.6 and Table 5.12) does not change the results (tables not shown), so it is unlikely that exclusions are driving the patterns.

Thus lateness, absence and exclusions cannot fully explain the day and time patterns in behaviour, although they may contribute to some of the observed effects.

Selective reporting or misreporting by teachers

A third possibility is that there is no difference in behaviour through the week, but reporting differences generate the incident pattern. For instance, dates and times might not be accurately reported; rather, they might reflect when incidents were recorded rather than when they happened, or have dates and times assigned arbitrarily. However, when descriptions of incidents are available and give information about the day and time of an incident, these match up with those recorded. Only if measurement error was systematically related to other characteristics would it generate daily and weekly patterns; classical measurement error should simply introduce noise. One possibility is that when the date of an incident is missing this is set to the Monday of the week it happened. However, it is possible to leave the date blank, so this would not be necessary; and this would not account for the higher rate of incidents on Tuesdays relative to Wednesdays-Fridays. Perhaps more plausible is that teachers’ ability or willingness to report incidents varies by day and time. This could take two forms: first, that teachers misreport the time or day of incidents because they wait until they have a break in their teaching schedule to report incidents, then record the time at which they report the incident as the time at which the incident happened. A second form could be that some teachers record dates and times accurately but are less inclined to report incidents earlier in the day and later in the week, due to differential fatigue or other reasons.[166] If misreporting were of the former kind, we would expect to see many more incidents in lessons 2, 4 and 5, as break time, lunch time, and the end of school follow these lessons, and possibly more incidents reported in lesson 1 of Tuesdays-Fridays (carried over from lesson 5 of the previous day). We would also expect to see more incidents on Tuesdays-Thursdays, as some teachers would report the previous days' incidents while others would report that day's incidents. However, these are not the patterns we see: lessons 2 and 4 have relatively low rates of incidents (usually lower than lesson 3), and Mondays have the highest incident rates. I cannot offer evidence on the second form of selective reporting, as I have no independent way of assessing when incidents occur. However, those incidents serious enough to involve call-outs of senior members of staff and withdrawal of pupils from lessons are reported by staff other than the teacher first involved and are very likely to be accurately recorded. There are significantly fewer of these on Thursdays and Fridays than on Mondays and Tuesdays, and significantly fewer in lessons 1 and 2 than in lessons 3-5 (table not shown). Thus it seems unlikely that differences in teachers' propensity to report incidents are driving the results.

As a test of whether teachers were reporting more or less negatively by day of the week, I used the teacher-reported Goodman SDQ to estimate the impact of day of the week on teachers' reports of pupils' behaviour (Table 5.21). There is likely to be even more selection in when teachers filled in these questionnaires than in when pupils completed surveys, as teachers were not constrained to doing so during the school week or during particular time slots. Nevertheless, there does not seem to be any strong pattern in teacher-reported behaviour scores by day of the week, although those who filled in the questionnaires on Sundays appear to have given pupils slightly higher (worse) scores.

---

[166] For instance, Bloom and Van Reenen (2007) report that a firm's scores on an international management survey were significantly higher when managers were interviewed later in the week and/or earlier in the day. They treat this as reporting bias on the part of the interviewees and so control for time and day in their analysis. However, there could be several processes generating this pattern: interviewees' mood is certainly one, but interviews were not scheduled at random and it is possible that different days and times of the week could reflect different managerial situations (as with variation in tasks over the week). In addition, one of the two scorings for each interview was conducted by the interviewer at the time of the interview, so their mood could also have a role.

Teacher wellbeing

A fourth possibility is that teachers record incidents accurately, but their own emotional state has a strong weekly and daily pattern and this influences pupils' behaviour. For instance, teachers who are tired or feeling low may be less able to deal effectively with a mildly disruptive pupil, allowing the situation to escalate. Similarly, they may teach less engagingly, resulting in more disruption by bored pupils. I do not have evidence on the wellbeing of teachers in this sample, but there are a number of studies which map the diurnal and weekly patterns of wellbeing of adults (e.g. Taylor, 2006; Bryson & MacKerron, 2012; Stone et al., 2006). If the wellbeing of teachers in my sample follows a similar pattern, this could help to explain the weekly pattern in behaviour incidents: teachers may be happier later in the week, and this could lead to better behaviour amongst pupils. However, this would not explain the time of day pattern: these studies suggest that working people get happier through the day, so we might expect fewer incidents later in the day. Tiredness might be important: if tiredness decreases towards lunchtime then increases again (Stone et al., 2006), and teachers' lunchtimes fall around late morning, then by lesson 5 they may be at their most tired. However, none of these patterns fully explains the pattern of behaviour incidents I have found, so although it is plausible that teachers' state of mind may contribute to this pattern, pupils' wellbeing may well do so too.

Teacher absences

Teacher absences could be contributing to the pattern of incidents. Many studies in a variety of settings find that absences from work are higher on Mondays (Ichino & Moretti, 2009; Bryson & Forth, 2007; Miller, 2008). I do not have data on teacher absences for this sample, but if teachers are more likely to be absent on Mondays, and if cover staff are less able to engage or control classes, then some of this pattern may be generated by teacher absences.[167] It seems less plausible that teachers are more likely to be absent at different times of day. For one school I have data on the teachers involved in incidents. Using this, I can compare the proportion of incidents involving cover staff that occur on different days of the week and at different times of day. I find that there is a significantly higher fraction of incidents involving cover staff on Mondays (8.1%) than on Tuesdays (5.2%), Wednesdays (6.6%) and Fridays (4.9%). There is no clear

[167] When a teacher is absent their lesson is taken by another teacher, usually agency supply teachers or cover staff employed by the school for this purpose. Lessons are not cancelled.

pattern by time of day, with only lesson 3 (11%) having a significantly different fraction of incidents involving cover teachers than lesson 1 (5.9%). However, the sample size is smaller here (N incidents=6,227), and there is likely to be substantial measurement error: I cannot include lessons with no behaviour incidents as I do not know who is teaching a class when there are no incidents. If teacher absences were driving the behaviour pattern we might also expect higher incident rates in winter, when employee absences tend to be higher (Ercolani, 2006). Mean incident rates are lower in December-February than in September-November, although the lowest rates are in April-July. So although teacher absences might be contributing to the Monday effect, they do not fully explain the incident patterns.

Endogenous timetabling

I have assumed throughout that different subjects are timetabled throughout the week in an arbitrary way. I know that all pupils will have a full timetable, and that full-time teachers will have to teach most lessons, so there is little room for selecting timetable slots. Moreover, scheduling lessons for at least five yeargroups will mean that inevitably some pupils will have lessons at times which are less desirable. Nevertheless, it is possible that schools are already scheduling subjects to optimise pupil attainment, taking into account time and day impacts and any other factors. Note that if day and time effects are real and schools are organising timetables optimally, this should *reduce* the size of the observed patterns, as schools will already be compensating for the negative impacts of afternoon or Monday lessons. The only situation in which endogenous timetabling would undermine my identification of day and time patterns would be if there were in fact no patterns, but (all four) schools scheduled particular subjects or teachers at specific times, and these were associated with higher or lower incident rates. For instance, if schools scheduled the most difficult or incident-prone subjects and the weakest teachers for Mondays and for the last lesson of each day, this could generate the incident patterns we see. Table 4.20 in Chapter 4 (‘Relative frequency of incidents by subject’) suggests that incident rates vary by subject. For example, there are 1.3-1.4 times as many incidents in English lessons as in physical education (PE) or performing arts lessons, and 2.2 times as many incidents in maths as in English. If maths and sciences are standardly scheduled late in the day and early in the week, this could be generating the behaviour patterns we see.

I cannot directly test whether different subjects are scheduled at different times, because I only have information on the subject studied for lessons when a pupil is involved in a behaviour incident; I have no information on the subject timetabled when there are no incidents. Nevertheless, I can set the incident count to zero whenever an incident occurs in a maths or science lesson and check whether the patterns remain.[168] Table 5.24 presents results from regressions which exclude incidents of maths and science. Here we see the same day of the week and time of day patterns, with only slight differences in the absolute and relative sizes of the coefficients. Thus it would appear that maths and science, the two subjects most prone to behaviour incidents, are not solely responsible for the results. I can also run the same regressions with each subject separately, finding very similar results for all subjects: an increase in incidents through the day and a decline through the week (table not shown). This suggests that the patterns I observe are not driven by endogenous timetabling of high-incident subjects, although I cannot test whether timetabling of individual teachers is responsible.

Discussion

I suggest above that the observed behaviour patterns are not due to misreporting or endogenous timetabling, but are likely to be related to pupil wellbeing, and that pupil absence and lateness, teacher and parent wellbeing, and teacher absence may also contribute. However, it is not clear what mechanisms are responsible for the 'Monday effect' on wellbeing and behaviour, which is also observed in many other contexts. One reason could be that pupils have to readapt to school activities and discipline each Monday after a relatively free weekend, and this takes some time. This is similar to suggestions that First World War factory workers had lower output (and higher injury rates) on Mondays because of a loss of coordination due to lack of practice of skills over the weekend (cited in Bryson & Forth, 2008; see also Vernon, 1921). Although the academic and self-control skills required at secondary school might seem to be of a very different type than those used in munitions factories, it is known that pupils lose academic skills over holiday periods (Lindahl, 2001), and the need to rehabituate oneself to school might prove difficult for teenagers. An example of this is the decreased rate of lateness at the end of the week (Thursdays and Fridays), which we could attribute to the establishment of a routine. Of course, this applies equally to teachers, who may be less able to teach effectively and control behaviour on Mondays (and possibly pupils' parents, who may have their own work-related problems).

[168] I also set the incident count to zero if I do not know the subject scheduled at that time. This will attenuate the estimate of the time patterns for all other subjects because it will also set to zero some incidents which happened in other subjects. Including these incidents does not change the results.

However, one striking result from my data is the lower rate of incidents on days which are immediately after holidays of at least 4 days' duration: these days are better than the average Monday, and about as good or better than the average Tuesday. Moreover, weeks 0 and 1 of each half term appear to have fewer incidents on average than weeks 2-4. If the Monday effect was entirely due to lack of habituation, I would expect as many or more incidents at the beginning of each half term.[169]

A second (additional) reason could be tiredness after the weekend due to the change in routine, for teachers, pupils, and pupils' parents. Sleep delay on Friday and Saturday nights – going to sleep and getting up at a later hour than during the week – is a common pattern in both adolescents and adults, and RCT evidence on artificially delaying sleeping times suggests that this has an effect on performance. Yang and Spielman (2001) delayed young adults' Friday and Saturday sleep schedules by 2 hours, while a control group followed their normal routines. The delayed sleep group showed greater insomnia on Sunday nights, and poorer cognitive performance and lower mood on Mondays. Wolfson and Carskadon (1998) surveyed 13-19 year old students, finding that those reporting less sleep each night or a larger weekend sleep delay had worse school grades, and reported greater daytime sleepiness and depressive mood. They concluded that most adolescents did not get enough sleep, and that this interfered with functioning during the day. Similarly, Carrell, Maghakian and West (2011) found that first year university students with courses randomly scheduled for the first period of the day did worse in *all* classes that day, and recommend that school start times should be delayed to account for adolescent sleep preferences. Thus lack of sleep on Sunday nights could contribute to poor performance on Mondays.

It is not just changes in sleeping routines which distinguishes the weekend from other days and could spill over onto Mondays: the Monday increases in cardiovascular mortality in Moscow and Scotland have been attributed to weekend binge drinking (Chenet et al., 1998; Evans et al., 2000), while Willich et al. (1994) suggest that the stress of the new working week might be contributing in Germany. While it is unlikely that many students in this sample will be heavy drinkers by the age of 16, 34% of pupils aged 14 in the full UKRP sample reported drinking alcohol outside of their families, and 12% reported drinking more than once a month, and this may be more likely at

[169] There might be reporting issues at the beginning of term, especially at the beginning of academic years, e.g. if the school IT systems are not up and running or new staff are unfamiliar with the system. However, the week of half term pattern is observed throughout the year, not just at the beginning of the academic year or at the start of whole terms when new staff and new systems are most likely to arrive.

weekends. Weekends may also be a time of enhanced family stress for many children. Only about 47% of this sample live with both of their parents, and 37% of the full UKRP sample ate a meal with their family twice a week or less often when aged 13-14, both of which could be indicative of relatively low family cohesion. Confrontations which occurred between pupils over the weekend could also affect pupils' mood on Mondays.[170] Importantly, I see the Monday effect in the behaviour of all pupils, even in those groups with low incident rates. This suggests that mechanisms common to most pupils may be responsible, such as weekend sleep deprivation, or factors affecting the mood of teachers and parents, including sleep deprivation, family events, and alcohol consumption.

Note that this tiredness – due to the 'hangover' from weekend activities – is different from the cumulative fatigue workers might experience as a result of days or weeks of long shifts. For instance, Vernon (1921, Chapter 3) notes an improvement in total output as a result of reductions in hours in the working week.[171] Similarly, Folkard and Lombardi (2006) find that consecutive day (and particularly) night shifts result in an increase of injury risk. As a result of this, the middle of the week might be the most productive, as workers are practised but not yet tired. However, this is in a context of heavy 12-hour shifts, and starting from a baseline working week of 66 hours: it is unlikely that school pupils face such extreme demands. State school teachers work 40-50 hours per week on average (Green, Machin, Murphy, & Zhu, 2008), so may have more room for an effect. Such cumulative fatigue may in part explain the pattern of incidents over the half term: incidents rise over the first few weeks of the half term, then plateau until week 6, which usually marks the end of the half term.[172] Thus teachers may become more tired through each half term, reducing their ability to teach engagingly and control behaviour, and pupils may become increasingly bored.

A third explanation of the day of the week effect operates through (the perception of) the disutility of work or school. Several studies find that mood improves through the working week (Taylor, 2006; Bryson & MacKerron, 2012; Csikszentmihalyi & Hunter, 2003), and this may be related to the proximity to the weekend. As Monday is the day furthest from the next weekend break and has the highest volume of work ahead, this

---

[170] Though there are actually fewer incidents of assault and fighting on Mondays than on subsequent days (Table 5.5).

[171] This is for women turning fuse bodies (heavy work) in 1915-1917, and men sizing fuse bodies (very heavy work) over the same period.

[172] Incidents fall again into weeks 7 and 8, but most half terms do not last longer than 6 weeks, so this drop may be due to special events happening around Christmas or at the end of the summer term.

may be the lowest day. Anticipation of the next holiday could also contribute to the inverted U-shape we see in incidents over the half term: incidents might first increase due to fatigue, then decrease due to holiday anticipation. Anticipatory anxiety about the week ahead may also contribute to the Monday effect. Bryson and MacKerron (2012) show that happiness declines on Sunday evenings (unlike on other days), and this may be why we see a negative Sunday effect on teachers' reports of pupils' behaviour (Table 5.21). This is known as 'Sunday neurosis' (Areni & Burger, 2008), and may continue into Monday morning. Wellbeing would then improve through the week as realised events are not as bad as anticipated.[173] Both of these effects could contribute to a lack of motivation on Mondays for both teachers and pupils (and again, pupils' parents).

The day of the week patterns I observe appear very similar to those in most of the literature on day of the week. I find that Wednesday, Thursday and Friday are all equally good days for behaviour and wellbeing. There may be a slight dip in wellbeing on Fridays, but this is not clear. Vernon reports a typical pattern of a rise in industrial productivity to the middle of the week, then a fall towards the end, although the 'best' day in this pattern varies, and days 3-5 often look very similar (Vernon, 1921, pp27-31). MacKerron and Mourato (2010) find that happiness is higher on Thursdays than on Mondays-Wednesdays, with a further increase on Fridays. Csikszentmihalyi and Hunter (2003) find a rise in happiness through the week from Monday to Saturday, although adjacent days' levels may not differ significantly. Taylor (2006) finds that job satisfaction and mental health scores are worse on Mondays and better on Fridays and Saturdays relative to Tuesday-Thursday, which all score about the same. Thus the main difference in my findings is that Fridays are not usually significantly better than Thursdays. However, the closest population to the one I observe is the sample of children aged 12-18 in Csikszentmihalyi and Hunter, who find that Friday is no better than Thursday on average for children's wellbeing, and that the large end-of-week uptick comes only on Saturday. This may be because young people anticipate upcoming days less than adults. However, if pupil absence is higher on Fridays this would bias down the incident rate, which would mean the pattern would look closer to the inverted U-shape in industrial productivity asserted by Vernon (1921). Nevertheless, the differences I find between the incident rates and psychological wellbeing on

[173] Torges, Stewart, and Nolen-Hoeksema (2008) show that depression and anxiety are higher and wellbeing is lower *before* the death of a hospice-based loved one than one month after the death. This suggests that in some circumstances the anticipation of an adverse event can be worse than the experience of the event itself.

Wednesdays-Fridays are small and generally not statistically significant, so it is not clear what the pattern at the end of the week really is.

It seems unlikely that the day of the week is important in itself. Rather, the day in the context of the normal working week and the sequence of activities through the week are what might determine its importance. Thus cardiovascular deaths peak on Mondays in Scotland but on Sundays in Israel (Evans at al., 2000; Anson & Anson, 2000), the first days of their respective working weeks. This may not be the case for the time of day – there may be biological mechanisms causing particular times of day to be more or less favourable to productivity or happiness, even once activity and rest schedules have been taken into account. Nevertheless, the time of day effect I observe seems more difficult to explain. I do not have information on psychological wellbeing for this sample by time of day, but if this is similar to the pattern found by other studies, then the relationship between incidents and wellbeing observed for the day of the week effect (with worse mental health being associated with more incidents) is reversed. I find that behaviour incidents increase through the day, while the literature tends to find that wellbeing increases through the day. For instance, Csikszentmihalyi and Hunter (2003), Bryson and MacKerron (2012) and Stone et al. (2006) all find that happiness and other positive emotions rise through the day, particularly during the working week. It also seems odd that starting a week would have a negative effect on behaviour, but starting a day would have a positive effect, unless different mechanisms are at work in each case (e.g. fatigue from lack of sleep is felt all day, but fatigue from boredom or work is felt in the afternoon). Yet my findings appear robust, and are in line with anecdotal evidence from teachers, who commonly report that the last lesson of the day has the most incidents and that even well-behaved classes can display challenging behaviour at this time. Interestingly, I have not heard stories from teachers about a day of the week effect. This could be because the time of day effect is much larger: Lesson 1 has only 56% as many incidents of Lesson 5, while Fridays have 78% as many incidents as Mondays. Larger effects are more likely to be 'visible to the naked eye', so teachers will have had more chance to notice them (Cohen, 1969).

One explanation could be increasing fatigue through the day, for both pupils and teachers. Stone et al. (2006) use the Day Reconstruction Method, in which respondents record their experiences of the previous day with a detailed questionnaire, to look at diurnal patterns of emotion in a sample of employed women.[174] They find that tiredness followed a v-shaped pattern: decreasing to about 12 noon, then increasing to the end of the day. The pattern remained even when activities were controlled for. If tiredness followed a similar pattern for the pupils and teachers in my sample, adjusting for a slightly earlier start time, we might expect tiredness to start increasing during lesson 3, after break time. This could contribute to the increased rate of incidents during lessons 3-5. Related to this is ‘spell fatigue’: fatigue, monotony and reduced effort resulting from uninterrupted spells of work within shifts. In the context of munitions factories, Vernon (1921) presents evidence that frequent short breaks relieve spell fatigue and increase productivity. Similarly, Folkard and Lombardi (2006) find that the type and length of shifts matter more to injury rates than total weekly hours of work, with frequent breaks undermining other day and week patterns by keeping injury rates low. Pupils at these schools face a maximum of 2.5 hours of uninterrupted focus, if we include registration and assembly times as well as lesson times. This is clearly much shorter than the long (up to 6 hour) spells the munitions workers studied by Vernon had to endure, but could still be a strain. However, this does not seem to fit the observed pattern: there is a sharp increase in incidents between lessons 1 and 2, but lesson 3 (after a break) actually has more incidents on average than lesson 4. Of course, being at school, even during a lunch break, is still not free time, and we could view the ‘spell’ as lasting the length of a school day. ‘Fatigue’ can include boredom as well as physical tiredness: part of the impact of shift length on factory workers was attributed to monotony rather than actual tiredness, and for pupils who do not generally enjoy academic lessons the fifth lesson of the day might prove particularly challenging.[175]

An important factor might be how break times are spent. One theory commonly advanced by teachers is that pupils are more restless after a break, particularly if they have been physically active, so that lesson 5 (after lunch) and lesson 3 (after morning break) would have worse behaviour, as indeed they do. Since there is no variation in the timing of lessons and breaks between the four schools I cannot test the impact of breaks directly. However, if boys are more likely to be active during break times then we might expect this pattern to be more pronounced for boys than girls, but this does not appear to

[174] The recruitment strategy oversampled employed teachers, nurses, and telemarketers. Since most teachers in my dataset are women, these results may be particularly relevant.

[175] Alternatively, ‘eveningness’ (greater alertness in the evening) is a common adolescent trait (as opposed to ‘morningness’; Preckel et al., 2012; Kirby & Kirby, 2006). If misbehaviour requires a certain amount of energy or alertness, and pupils are simply too tired to misbehave in the morning, we might expect incident rates to rise through the day while average academic learning rose with it.

be case: from Table 5.13 we can see that the difference in the rate increases for boys in lessons 3 and 5 relative to lessons 2 and 4 does not differ much from those of girls. Moreover, Mahar (2011) surveys the cross sectional and experimental literature on the impact of physically active breaks on elementary school students, and finds small to moderate *improvements* on attention-to-task as a result of physical activity.[176] An alternative hypothesis is that it is the unstructured nature of breaks which some pupils find difficult, resulting in boredom, confrontation and disruptive behaviour.[177]

Changes in disruptive behaviour through the day might not be related to wellbeing in a simple way. Egloff, Tausch, Kohlmann and Krohne (1995) find that the pleasantness and activation (engagement) components of positive affect show different patterns through the day. In common with most of the literature, positive affect in terms of pleasantness increases linearly through the day, but engagement rises to early afternoon then falls into the evening. The idea that ‘flow’ contributes substantially to wellbeing is well established in positive psychology (Csikszentmihalyi, 1999; Seligman & Csikszentmihalyi, 2000).[178] If pupils are happier in the afternoon, but less engaged, we would expect to see more behaviour incidents. Moreover, negative emotions may show diurnal patterns. Stone et al. find that impatience increases substantially in the afternoon, which again could provoke worse behaviour. Changes in inhibition may also be important: elsewhere I show that pupils’ anxiety scores are negatively related to the number of behaviour incidents they are involved in – a bit of anxiety might usefully inhibit poor behaviour (see Chapter 4). Stone et al. (2006) find that worry and depression diminish through the day, which might suggest diminished inhibition in the afternoon. Similarly, Bryson and MacKerron (2012) show that feelings of relaxation increase substantially during the morning of a working day, remaining roughly constant through the afternoon. Thus even if pupils are happier in the afternoon, if they are not engaged in their work and are less inhibited this could result in more behaviour incidents.

I have not been able to show that poor behaviour has an impact on academic attainment, although other evidence suggests that it does (Gutman & Vorhaus, 2012; Lavy, Passerman, & Schlosser, 2012; Segal, forthcoming). Poor behaviour as an

[176] There is even a positive impact of physical activity on children with attention deficit hyperactivity disorder (Gapin, Labban & Etnier, 2011).

[177] This was the speculation of a former head teacher, who commented that his school had shortened the length of lunch time because of frequent difficult behaviour at the end of lunch breaks.

[178] ‘Flow’ can be roughly defined as the state of engagement in an activity, often from tasks which are neither too easy nor too challenging, and during which time passes unnoticed.

indicator of poor wellbeing may be interesting in itself, given that happiness may be associated with higher productivity (Oswald, Proto & Sgroi, 2009). Behaviour incidents are also disruptive and unpleasant for teachers and other pupils (NFER, 2012), and may be an impediment to effective teaching. Some solutions will involve whole-school changes in behaviour policies (DfE, 2012a), or specific interventions to target the poor behaviour and emotional wellbeing of individual pupils (Durlak et al., 2011), but one approach would be to rearrange timetables around pupils' propensity to misbehave. There are consistent differences in the rate of behaviour incidents per lesson by subject, and some subjects are more critical to overall academic attainment than others, notably English and maths. If schools were able to schedule more active lessons such as performing arts and PE earlier in the week and later in the day, allowing subjects such as English and maths to be taught in lessons 1 or 2, or later in the week, we could see a reduction in the number of incidents during the most critical subjects as well as an overall reduction in incidents. This might not be possible for all yeargroups, but schools could prioritise those yeargroups or classes with the worst behaviour or the most to lose from poor academic attainment. Alternatively, if sleep deprivation is a major factor in the Monday effect, delaying the start of school on Monday mornings could help reduce disruption (Carrell, Maghakian & West, 2011). The middle weeks of each half-term appear to have higher rates of behaviour incidents than the earliest weeks, so it might also be worth shifting activities to take this into account. For instance, using week 4 (a very disruptive week) for trips or other special activities might optimise the use of less disruptive time at the beginning of the half-term. Lazear (2001) suggests that the optimal class size may vary by the disruptiveness of students, or the subject taught. If timetabling subjects at particular times is difficult, schools could reduce class sizes in the afternoon (and increase them in the morning), or increase class sizes later in the week, to take account of differential disruptiveness. Likewise, they could look at disruptiveness by subject to assess whether smaller classes in English, maths and science might be justified.

The results I obtain come from descriptive rather than experimental data, and it is important to estimate the impact of rescheduling on academic attainment before widely recommending rescheduling to schools. It is plausible that lessons with good behaviour are lessons in which pupils learn more, but if good behaviour is partly due to tiredness or absence, for instance, we would not necessarily want to schedule difficult academic lessons during time slots when pupils were more likely to be tired or absence, as this might have a negative impact on academic performance. There are two pieces of research that could be carried out to investigate the impact of scheduling on academic attainment. First, linking GCSE results to the data used in this paper, the quasi-exogeneity of scheduling could be used as a natural experiment to test the impact of behaviour (and possibly other factors) on academic attainment at 16.[179] Second, implementing a randomised controlled trial in which school schedules were modified to promote the academic attainment of particular year groups (subject to each school's staffing constraints), would allow the estimation of the causal impact of scheduling on academic attainment, allaying concerns that the observed scheduling was not in fact random (as is often the concern with observational data and with the identification strategies in natural experiments). The RCT could use randomisation by cohort within school to ensure that rescheduling was feasible, and that appropriate control groups were available.[180] For example, the intervention could schedule active lessons such as PE and performing arts for period 5 on each day, and English and maths lessons during periods 1 and 2. If successful, this would enable us to see the impact of scheduling on behaviour and academic attainment.[181]

Conclusions

I find strong day of the week and time of day effects in behaviour incidents in secondary schools. Mondays have the highest rate of incidents, followed by Tuesdays, with substantially lower rates on Wednesdays-Fridays. In terms of lesson timing, the first lesson of the day has the lowest rate of incidents, with the rate rising through the day and the highest rate during lesson 5. I investigate mechanisms for the impact using additional data, and from this I conclude that the pattern is not simply due to selective reporting or misreporting, and is not due to endogenous timetabling. I suggest that the weekly pattern is related to pupil wellbeing, particularly tiredness and boredom, and that pupil absence and lateness, teacher and parent wellbeing, and teacher absence may also contribute to the weekly and daily patterns.

This paper adds to the literature on time of day and day of the week effects. Many papers investigating such effects are unable to account for the selection into working at different times by employees and workplaces; the endogeneity of activities, the duration

[179] GCSE results for many pupils in the data only became available in the NPD in spring 2013, so I was not able to include it in this thesis.
[180] This was the randomisation approach used by Stallard et al. (2012).
[181] I have discussed the implementation of just such an RCT with Eric Maurin of the Paris School of Economics.

of tasks, and of labour demand through the week; the endogeneity of measurement availability (when individuals must respond); and the difficulty of observing productivity or correlates of productivity at an individual level. By comparison, I know that all pupils have five lessons a day and should be at school five days a week, and that they have little choice in the activities they engage in during this time. Likewise, once timetables have been set, teachers must teach a regular schedule of lessons. Subject scheduling appears to be exogenous, and in any case is not responsible for the daily and weekly patterns in behaviour, as these hold for each subject individually. My data covers every pupil on roll for every day when these schools are open, so there is little selective reporting, and since I have schools' timetables I also know that pupils engage in similar activities for equally long blocks of time. Thus my identification of day of the week and time of day effects may be more robust than those reported in other contexts. In addition, much of the literature relies on factory settings where output can be easily measured; the school setting might better approximate productivity effects in the sort of office-based or service jobs which dominate developed economies. My findings suggest that modifying the organisation of time within schools could be an important way of improving conditions for staff and pupils and might increase attainment. Specifically, scheduling subjects with higher relative incident rates and those which are more critical to academic success at low-incident times of the day and the week could reduce the overall number of incidents as well as improving attainment in the subjects where it makes most difference to pupils and schools. Delaying the start of school on Mondays could be considered. However, we would need to examine the impact of scheduling on academic attainment before recommending this policy to schools, preferably through designing and implementing an RCT. Other strategies such as whole-school changes in behaviour policies, and targeted interventions for pupils with problematic behaviour, should also be adopted.

Chapter 5: Figures and Tables

Figure 5.1: Mean behaviour incidents per pupil per day
N incidents = 44,668; bars show 95% confidence intervals

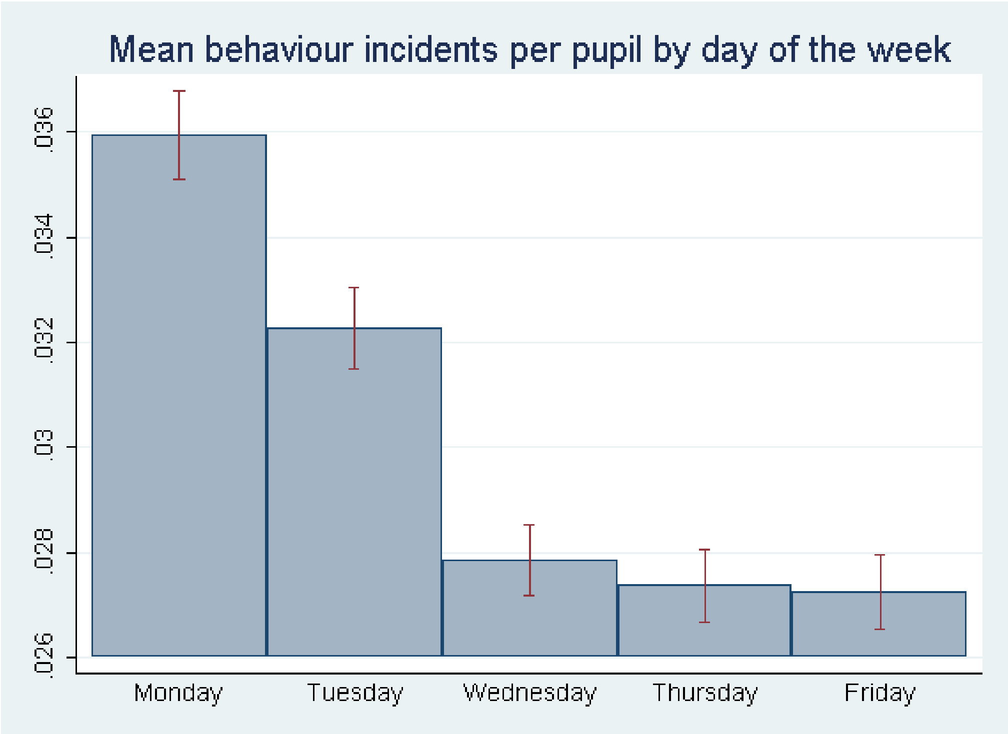


Figure 5.2: Mean behaviour incidents per pupil per day, by day of half term
The first Monday of the half term is set to 1. Days within the same week are linked, and Mondays are marked with a diamond.

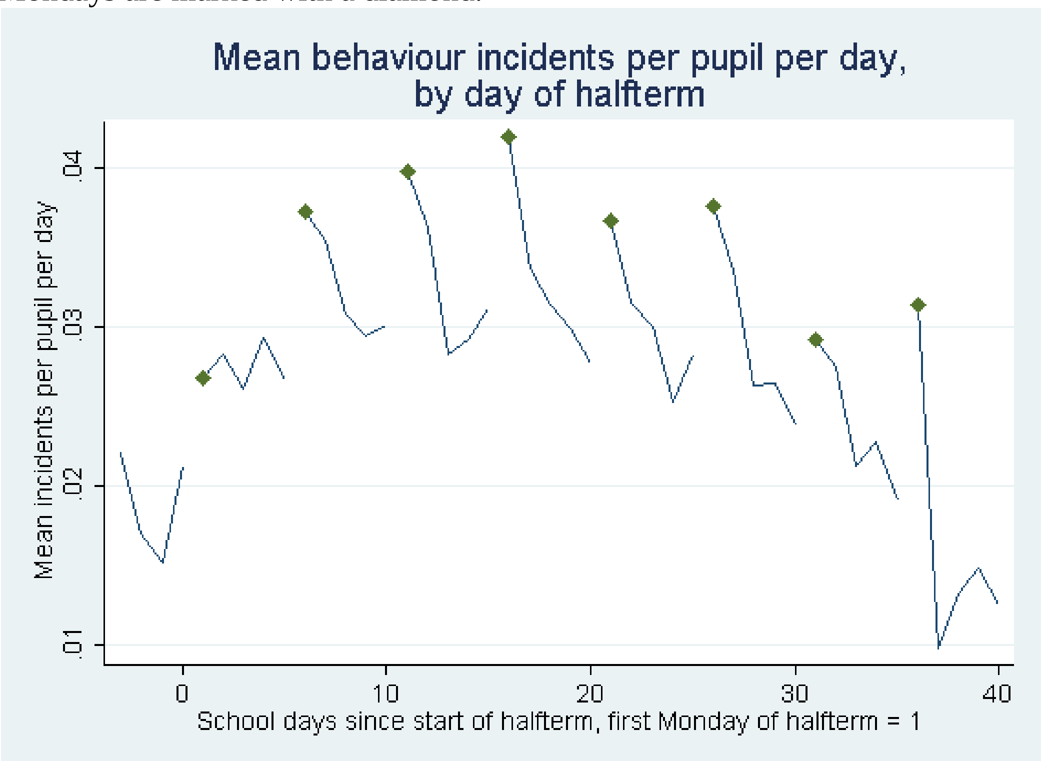


Figure 5.3: Mean behaviour incidents per pupil per day, by day of half term and gender

Boys are represented by solid lines, solid diamonds are Mondays; girls are dashed lines, hollow diamonds are Mondays.

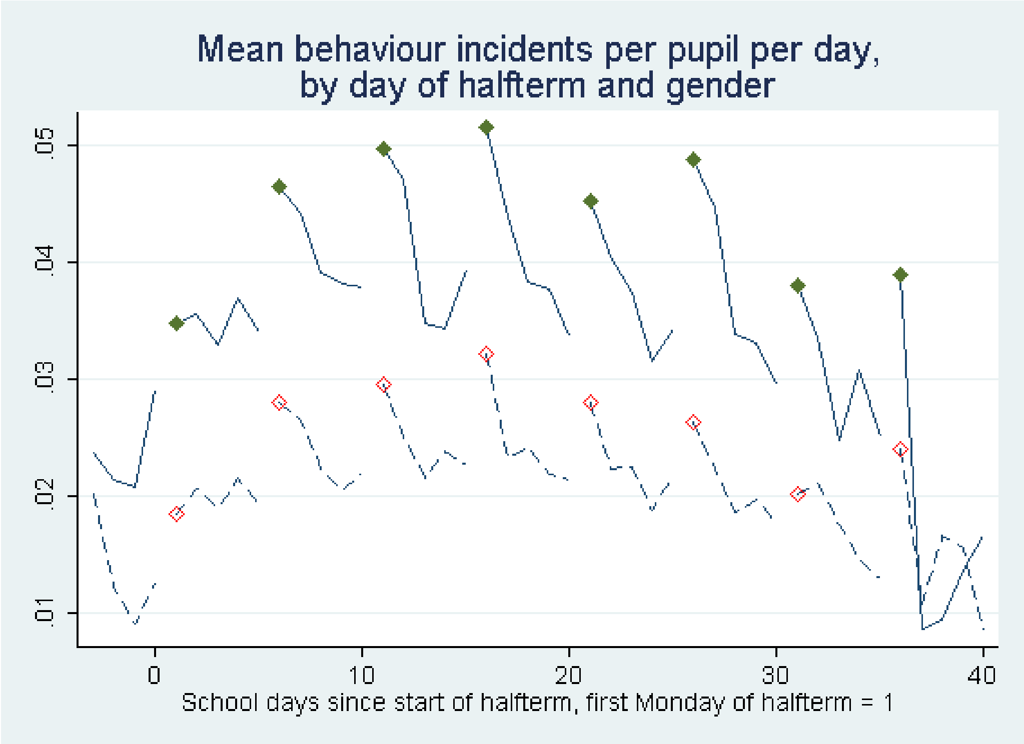


Figure 5.4: Mean behaviour incidents per pupil per day, by day of half term and SEN status

SEN pupils are represented by solid lines, solid squares are Mondays; non-SEN pupils are dashed lines, hollow squares are Mondays.

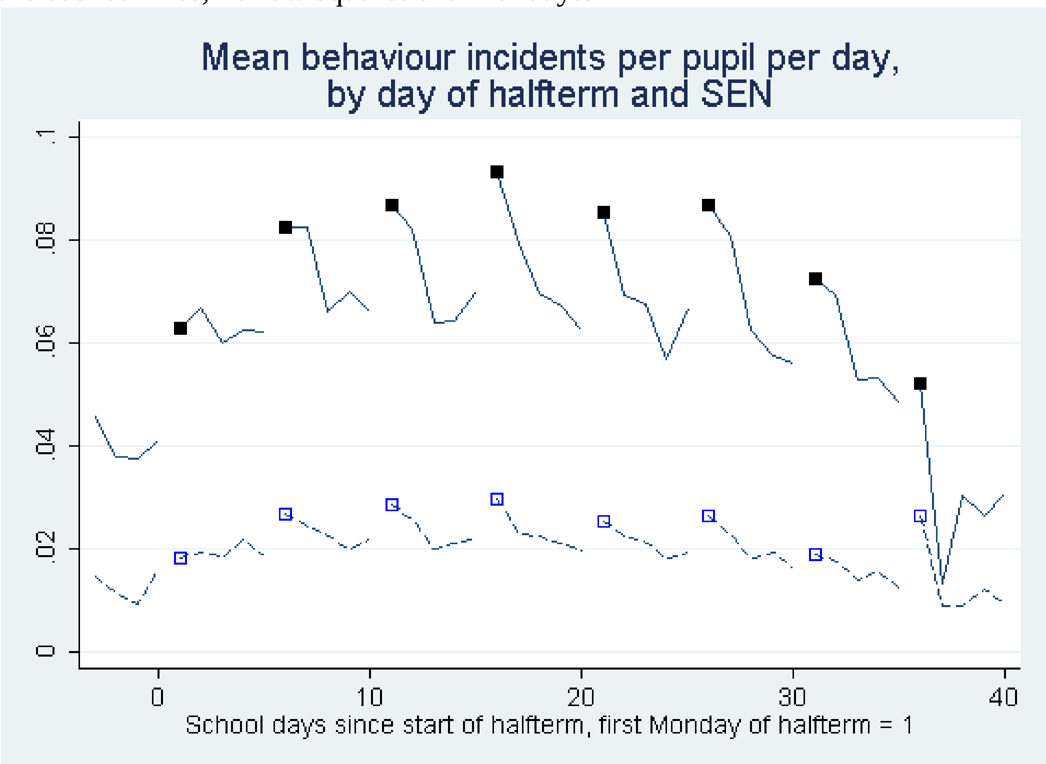


Figure 5.5: Mean behaviour incidents per pupil per day, by day of half term and FSM status
FSM pupils are represented by solid lines, solid triangles are Mondays; non-FSM pupils are dashed lines, hollow triangles are Mondays

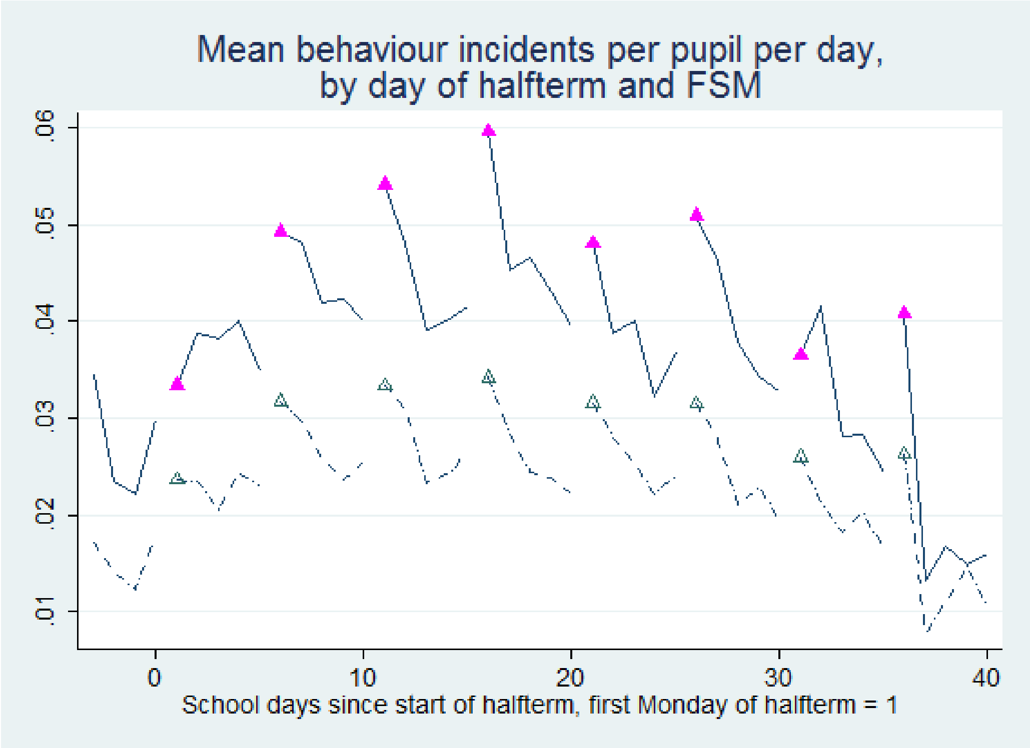


Figure 5.6: Mean behaviour incidents per pupil by yeargroup and day of the week

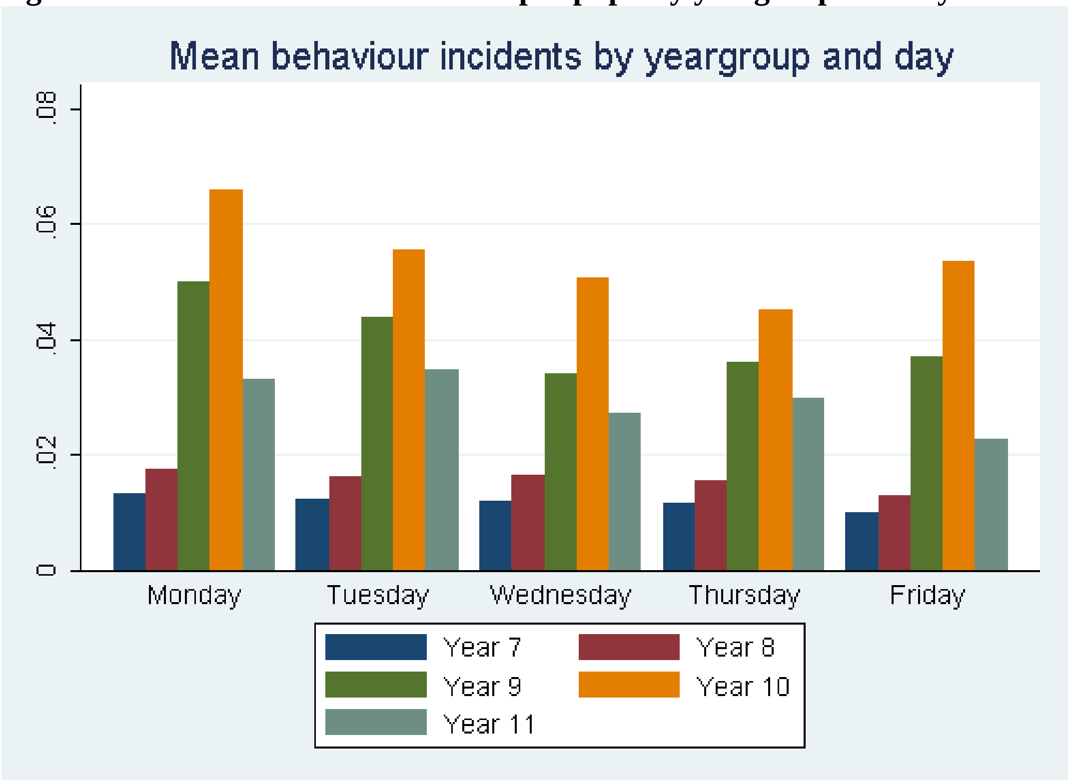


Figure 5.7: Behaviour incidents per pupil by time of day, lesson times only
Figure shows 95% confidence intervals

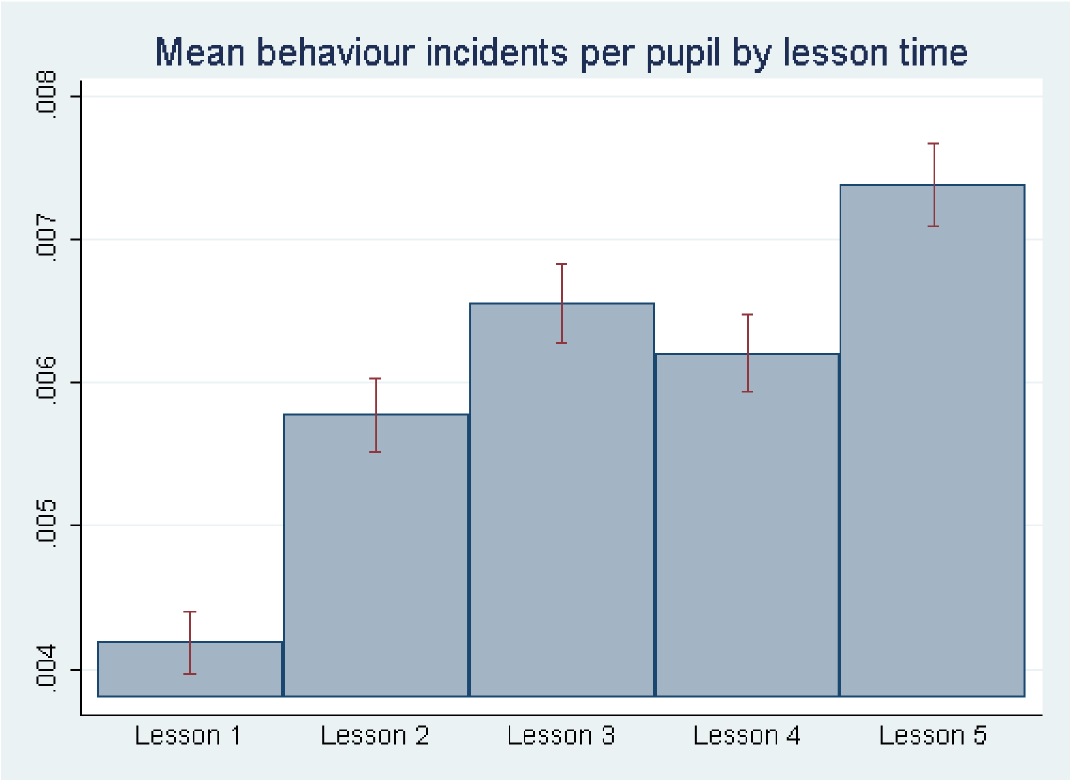


Figure 5.8: Behaviour incidents per pupil by time of day over the week
Monday lesson 1 is set to 1. Lessons within the same day are linked, and Lesson 1 of each day is marked with a diamond.

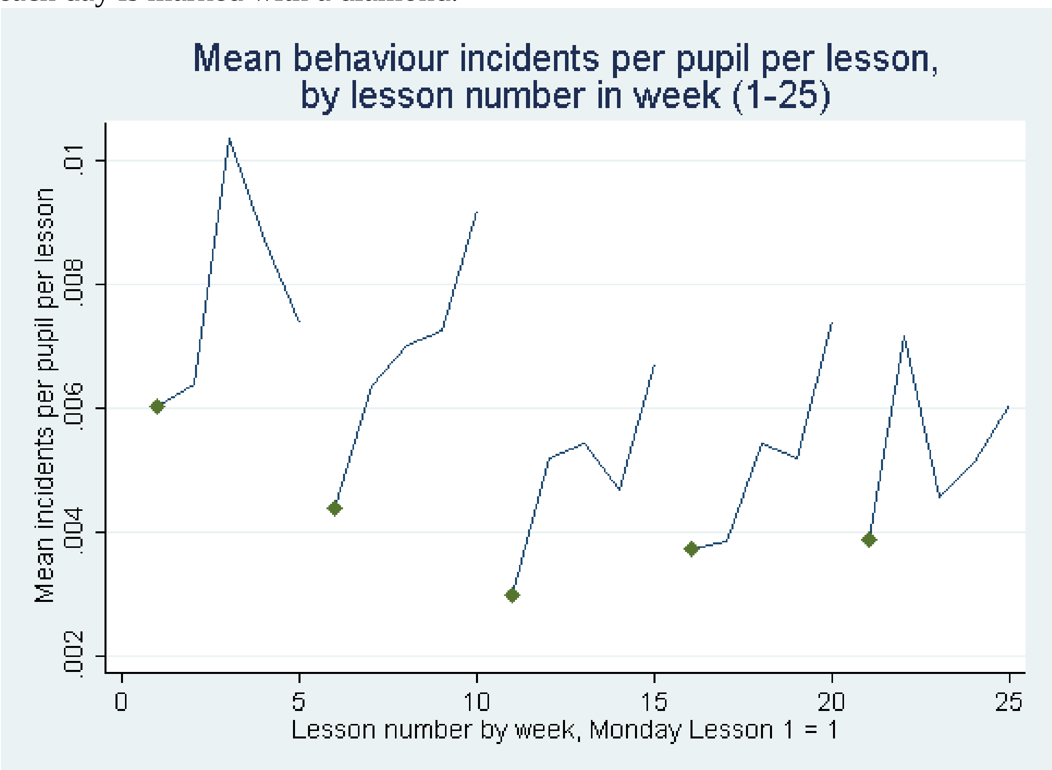


Figure 5.9: Behaviour incidents per pupil by time of day over the week and by gender

Boys are represented by solid lines, solid diamonds represent Lesson 1 of each day; girls are dashed lines, hollow diamonds represent Lesson 1.

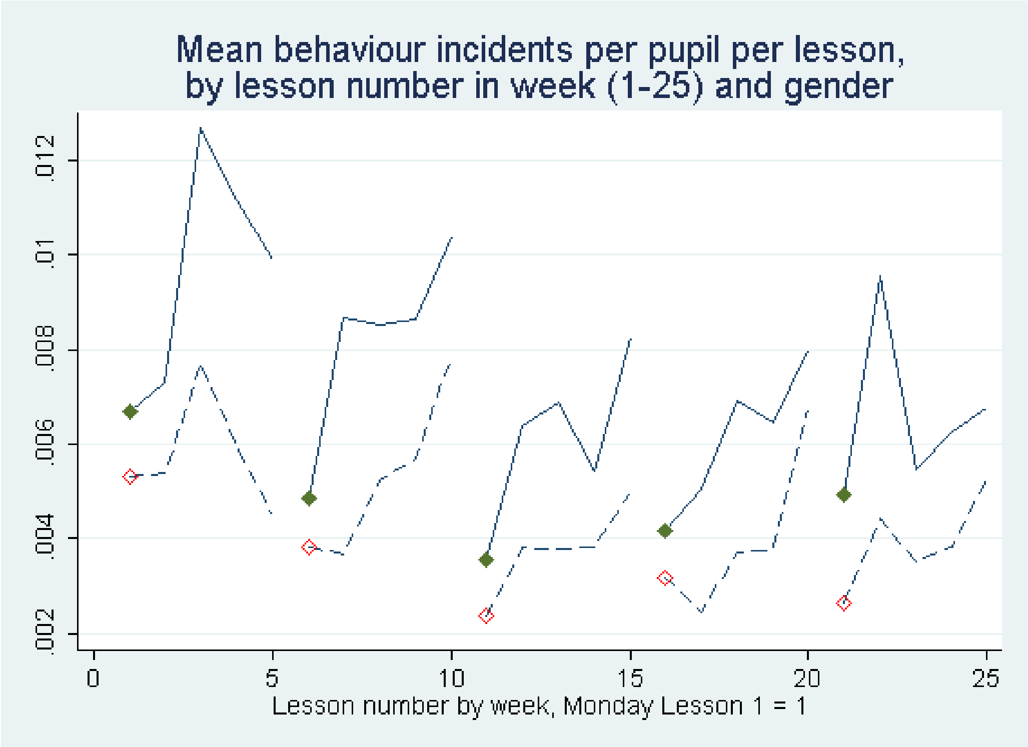


Figure 5.10: Behaviour incidents per pupil by time of day over the week and by SEN status

SEN pupils are represented by solid lines, solid squares represent Lesson 1 of each day; pupils without SEN are dashed lines, hollow squares represent Lesson 1.

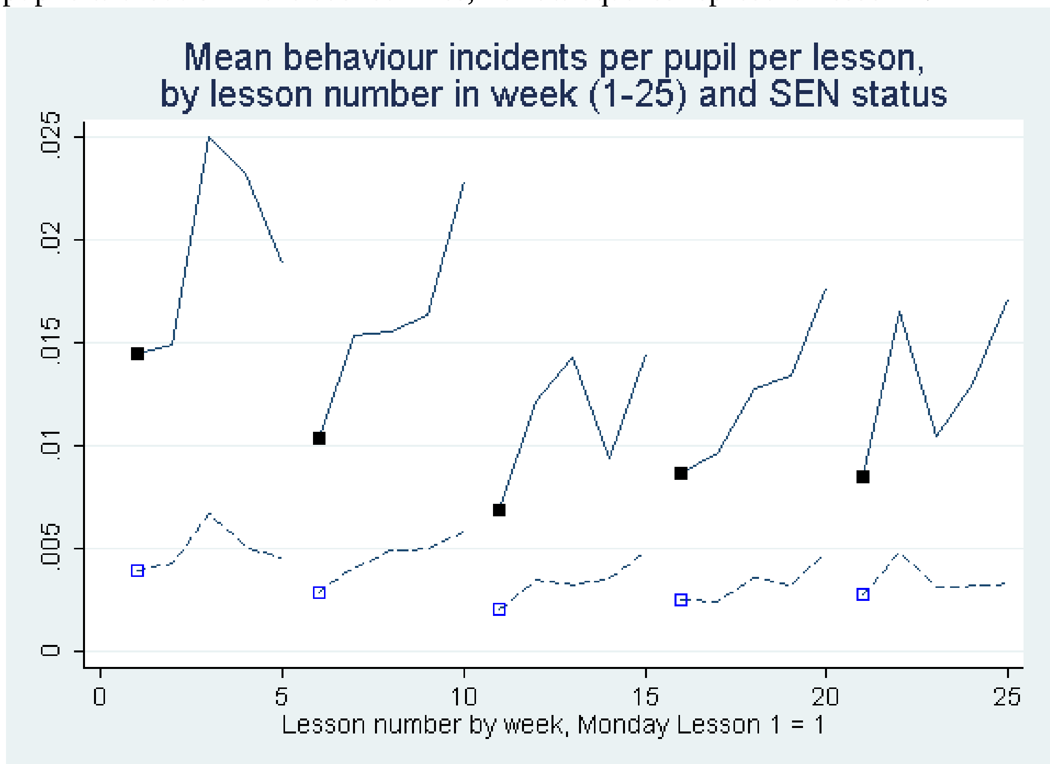


Figure 5.11: Behaviour incidents per pupil by time of day over the week and by FSM

SEN pupils are represented by solid lines, solid squares represent Lesson 1 of each day; pupils without SEN are dashed lines, hollow squares represent Lesson 1.

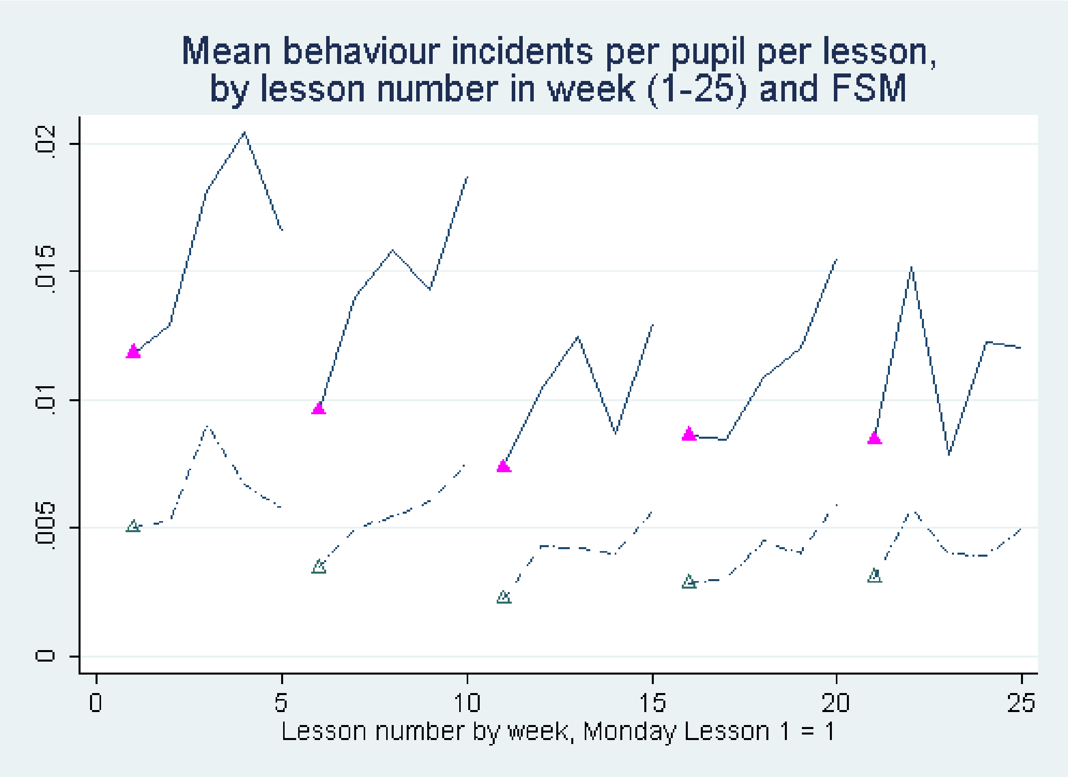


Figure 5.12: Mean number of behaviour incidents per pupil per lesson by time of day and yeargroup

Most pupils are in the dataset for more than one academic year, so most will appear at least twice in the graph in different yeargroups.

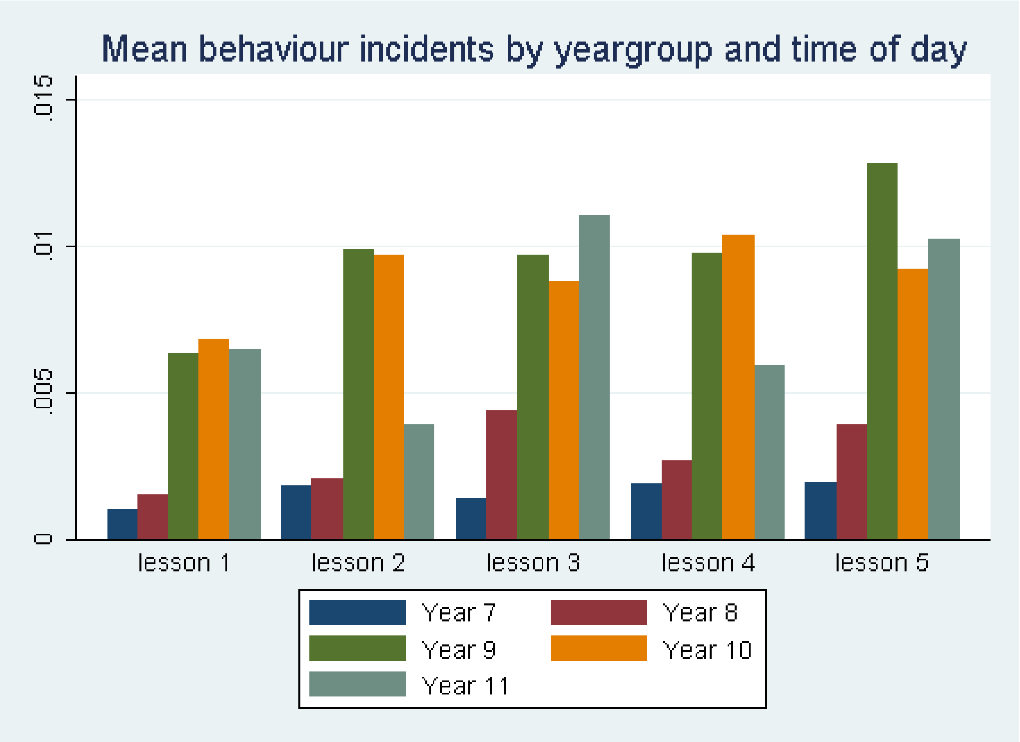


Figure 5.13: Self-reported behaviour score by day of the week
Means and 95% confidence intervals; a higher score indicates worse behaviour

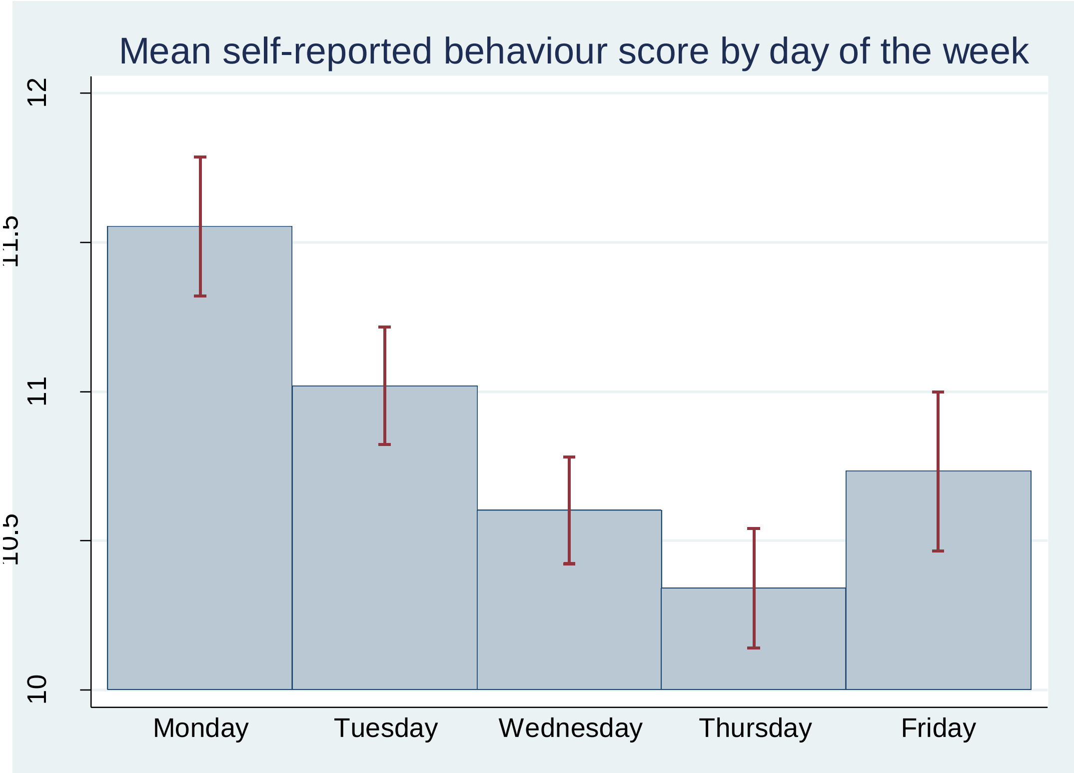


Figure 5.14: Depression score by day of the week
Means and 95% confidence intervals; a higher score indicates worse symptoms

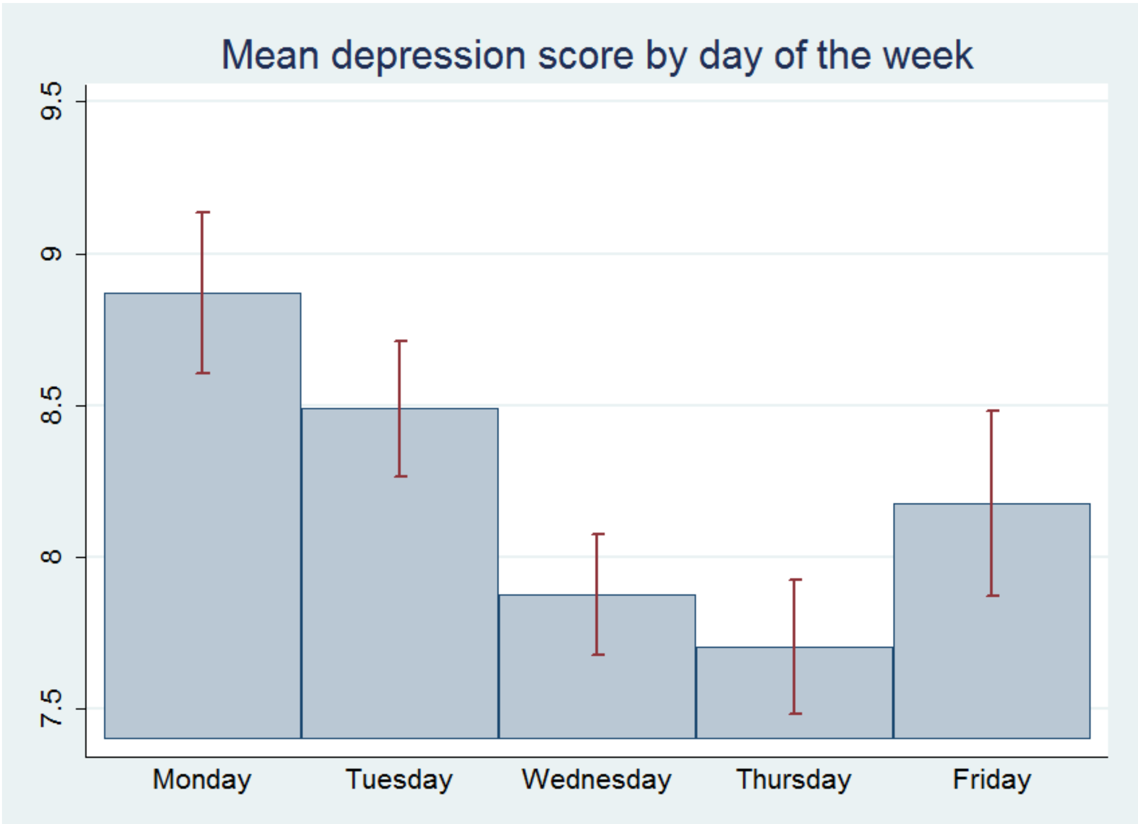


Figure 5.15: Anxiety score by day of the week
Means and 95% confidence intervals; a higher score indicates worse symptoms

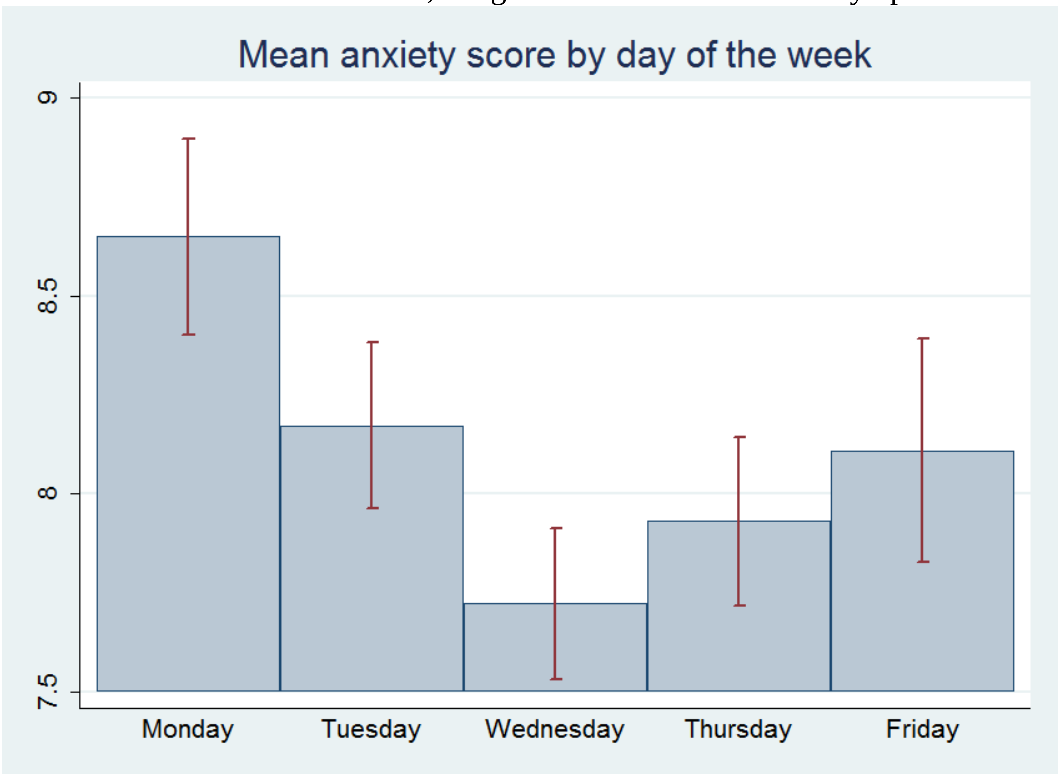


Figure 5.16: Number of bedrooms at home by day of the week
Means and 95% confidence intervals

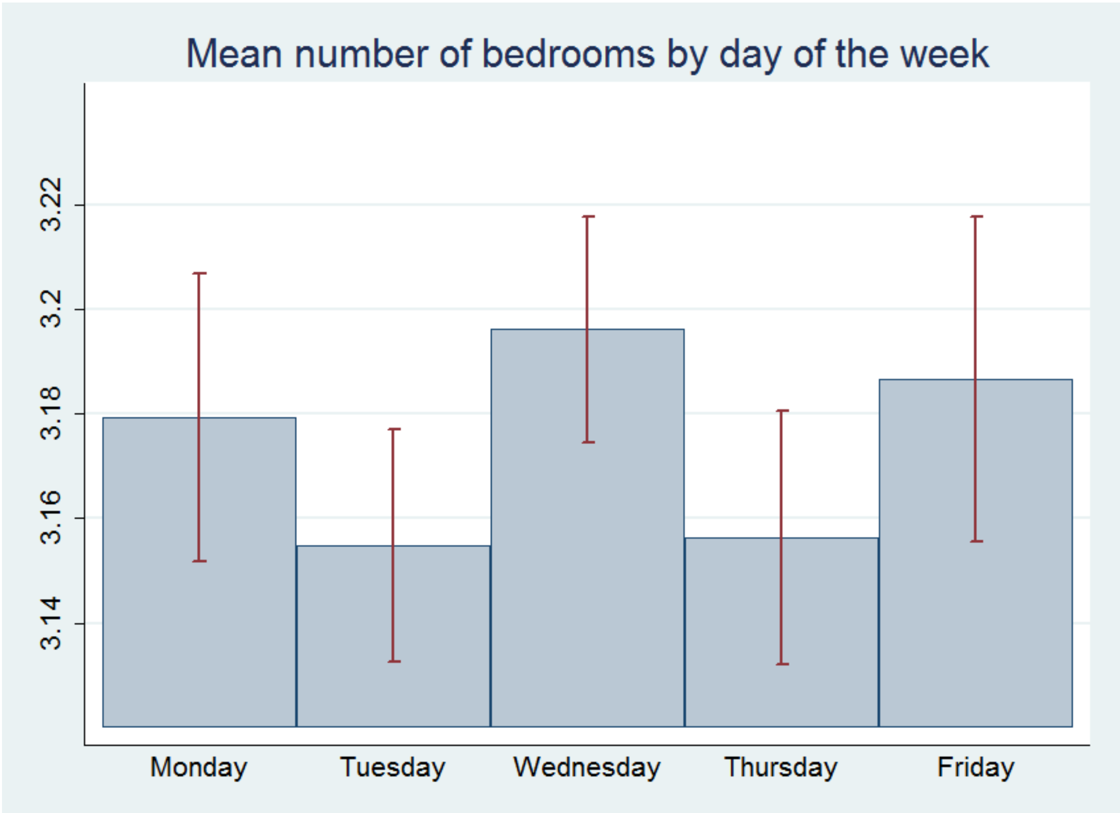


Table 5.1: Descriptive statistics by date (schooldays)

Earliest date	7th September 2005
Latest date	7th April 2011
Number of dates (school days)	1,116
Number of academic years (or parts of)	6
Fraction of dates for which have data for all four schools	**0.17**
School 1: fraction of days for which have data	0.28
School 2: fraction of days for which have data	0.64
School 3: fraction of days for which have data	0.34
School 4: fraction of days for which have data	0.93

Table 5.2: Schooldays by academic year

Academic Year	Number of days	Percent
2005-06	189	16.94
2006-07	200	17.92
2007-08	209	18.73
2008-09	192	17.2
2009-10	193	17.29
2010-11	133	11.92
Total	**1,116**	**100**

Table 5.3: Schooldays by day of the week

Day of the Week	N	Percent	N before a break of at least 3 days	N before a break of at least 4 days	N after a break of at least 3 days	N after a break of at least 4 days
Monday	215	19.27	0	0	27	25
Tuesday	225	20.16	1	1	17	12
Wednesday	227	20.34	4	4	5	5
Thursday	227	20.34	10	8	3	3
Friday	222	19.89	34	31	1	1
Total	**1,116**	**100**	**49**	**44**	**53**	**46**

Table 5.4: Descriptive statistics: pupils

	Mean	SD	Min	Max	N
Number of incidents per pupil (in dataset)	13.85	36.16	0	476	3,284
Number of school days per pupil in (in dataset)	451.81	221.73	9	954	3,284
Number of incidents per pupil per day	0.04	0.11	0	1.50	3,284
Assault or fighting: N incidents	0.45	1.26	0	24	3,284
Defiance: N incidents	2.42	5.97	0	62	3,284
Disruption: N incidents	3.80	15.11	0	263	3,284
Lateness or truancy: N incidents	1.75	9.91	0	198	3,284
Dangerous behaviour or property damage: N incidents	0.31	0.87	0	11	3,284
Verbal offence: N incidents	0.88	2.19	0	24	3,284
Other types of incident (minor): N incidents	1.66	4.48	0	57	3,284
Fraction male	0.51	0.50			3,284
Fraction with Special Educational Needs (SEN)	0.20	0.40			3,275
Fraction with emotional or behavioural disorder SEN	0.03	0.16			3,121
Age in September 2009	14.59	2.27	11	20.08	3,278
Born in autumn	0.25	0.43			3,278
Born in winter	0.23	0.42			3,278
Born in spring	0.26	0.44			3,278
Born in summer	0.26	0.44			3,278
Year of birth	1994.50	2.31	1989	1998	3,278
Fraction eligible for Free School Meals in past 6 years	0.30	0.46			3,276
IDACI score: neighbourhood deprivation measure	0.29	0.18	0	0.96	3,247
Mean Key Stage 2 score (combined English, maths and science)	4.51	0.67	1	5.74	2,710
Fraction white ethnicity	0.88	0.33			3,237
Fraction Bangladeshi or Pakistani ethnicity	0.05	0.22			3,237
Fraction black	0.01	0.12			3,237
Fraction other Asian ethnicity	0.02	0.15			3,237
Fraction other and mixed ethnicity	0.03	0.18			3,237
Depression symptoms score	9.65	6.53	0	44	1,303
Anxiety symptoms score	9.30	5.90	0	28	1,300
Self-reported behaviour score	12.10	5.50	0	34	1,304
Teacher-reported behaviour score	7.60	5.98	0	33	1,306

Table 5.5: Behaviour incidents by type and day of the week

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Total**
N (pupil x date observations)	285,673	301,667	304,095	302,741	289,935	1,484,111
Mean incidents per pupil per day	0.0360	0.0323	0.0278	0.0274	0.0273	0.0301
SD	0.2295	0.2170	0.1919	0.1919	0.1950	0.2054
Number of incidents	**10,270**	**9,737**	**8,469**	**8,286**	**7,906**	**44,668**
Broad type						
Missing categorisation	2,239	1,725	1,571	1,490	1,469	8,494
Disruption	2,947	2,799	2,113	2,245	2,382	12,486
Defiance	1,664	1,868	1,623	1,442	1,346	7,943
Late or truanting	1,231	1,209	1,077	1,100	1,064	5,681
Verbal offence	570	604	610	586	520	2,890
Dangerous behaviour or damage	209	224	203	187	186	1,009
Assault or fighting	226	323	296	309	318	1,472
Other (have got categorisation)	1,184	985	976	927	621	4,693

Notes: This table presents the broad category of behaviour incident by day of the week, for those incidents which can be assigned a specific date. Thus certain categories of incidents included in Table 4.2 have been excluded, e.g. no homework; missed detentions.

Table 5.6: Behaviour incidents per pupil by day of the week
Outcome: behaviour incidents per pupil per day

	(1) OLS: Incidents * 1000	(2) OLS: Incidents * 1000	(3) Negative binomial	(4) Negative binomial	(5) Negative binomial	(6) Zero-inflated negative binomial	(7) Zero-inflated negative binomial
Tuesday	-3.677***	-3.786***	-0.108***	-0.126***	-0.118***	-0.115***	-0.126***
SE	(1.380)	(1.251)	(0.038)	(0.040)	(0.016)	(0.038)	(0.040)
IRR			0.898	0.882	0.889	0.891	0.881
Wednesday	-8.101***	-7.974***	-0.255***	-0.248***	-0.268***	-0.279***	-0.248***
SE	(2.151)	(1.929)	(0.046)	(0.046)	(0.016)	(0.046)	(0.045)
IRR			0.775	0.781	0.765	0.757	0.781
Thursday	-8.594***	-8.139***	-0.273***	-0.268***	-0.276***	-0.300***	-0.268***
SE	(2.019)	(1.657)	(0.046)	(0.037)	(0.017)	(0.046)	(0.037)
IRR			0.761	0.765	0.758	0.741	0.765
Friday	-8.699***	-7.518***	-0.277***	-0.252***	-0.246***	-0.293***	-0.252***
SE	(1.716)	(1.379)	(0.041)	(0.038)	(0.017)	(0.042)	(0.038)
IRR			0.758	0.778	0.782	0.746	0.777
First day after 4+ days off		-5.278*		-0.183*	-0.152***		-0.183*
SE		(2.652)		(0.102)	(0.034)		(0.101)
IRR				0.832	0.859		0.833
Last day before 4+ days off		-7.644***		-0.392***	-0.362***		-0.392***
SE		(1.629)		(0.084)	(0.038)		(0.083)
IRR				0.676	0.697		0.676
Week 0		-5.968**		-0.372*	-0.327***		-0.367*
SE		(2.383)		(0.198)	(0.046)		(0.199)
IRR				0.689	0.721		0.693
Week 2		3.698**		0.124*	0.116***		0.122*
SE		(1.850)		(0.073)	(0.019)		(0.072)
IRR				1.132	1.123		1.130
Week 3		3.706**		0.137*	0.129***		0.134*
SE		(1.831)		(0.071)	(0.019)		(0.070)
IRR				1.147	1.137		1.143
Week 4		4.184**		0.160**	0.137***		0.157**
SE		(2.070)		(0.078)	(0.019)		(0.076)
IRR				1.174	1.147		1.170
Week 5		0.853		0.041	0.040**		0.038
SE		(2.078)		(0.086)	(0.020)		(0.085)
IRR				1.042	1.041		1.039
Week 6		0.627		0.009	0.023		0.009
SE		(2.300)		(0.087)	(0.021)		(0.086)
IRR				1.009	1.024		1.009
Week 7		-5.206*		-0.172	-0.160***		-0.172
SE		(2.823)		(0.121)	(0.028)		(0.120)
IRR				0.842	0.852		0.842
Week 8		-15.683***		-0.648***	-0.641***		-0.650***

SE		(2.246)		(0.164)	(0.086)		(0.164)
IRR				0.523	0.527		0.522
Inflate							
Girl						0.539***	-0.394**
						(0.068)	(0.162)
Special educational needs						-2.271*	-17.830***
						(1.218)	(0.259)
Free school meals eligible						-0.491***	0.475***
						(0.144)	(0.126)
Mean Key Stage 2 score						0.435***	1.441***
						(0.144)	(0.103)
N pupils	3,284	3,284	3,284	2,683	2,683	2,686	2,683
N	1,484,111	1,484,111	1,484,111	1,296,961	1,296,961	1,297,982	1,296,961
Clusters	78	78	78	78		78	78
Controls							
Half term, School, Half term*School	no	yes	no	yes	yes	no	yes
Pupil characteristics	no	no	no	yes	yes	no	yes
Pupil FE	no	yes	no	no	no	no	no
Random effects	no	no	no	no	yes	no	no

Notes: Table shows results of regressions of the number of behaviour incidents per pupil per day, relative to Mondays. Columns 1 and 2 use ordinary least squares regressions, and the outcome variable is behaviour incidents per pupil per day multiplied by 1000. Columns 3-7 have the number of behaviour incidents per pupil per day as the outcome variable. Columns 3-5 use negative binomial regressions, and columns 6 and 7 use zero-inflated negative binomial regressions. Each column represents a separate specification. Coefficients and incidence rate ratios (IRRs) are reported, along with standard errors clustered at the level of school*half term.

Table 5.7: Behaviour incidents over the week by pupil characteristics
Outcome: behaviour incidents per pupil per day

	Boys	Girls	SEN	not SEN	FSM	not FSM
Tuesday	-0.126***	-0.130**	-0.089***	-0.149***	-0.108***	-0.138***
SE	(0.040)	(0.057)	(0.032)	(0.054)	(0.040)	(0.047)
IRR	0.881	0.878	0.915	0.862	0.898	0.871
Wednesday	-0.270***	-0.219***	-0.243***	-0.254***	-0.191***	-0.289***
SE	(0.039)	(0.067)	(0.039)	(0.061)	(0.051)	(0.048)
IRR	0.763	0.803	0.784	0.776	0.826	0.749
Thursday	-0.287***	-0.241***	-0.252***	-0.275***	-0.231***	-0.293***
SE	(0.038)	(0.053)	(0.033)	(0.045)	(0.042)	(0.041)
IRR	0.751	0.786	0.777	0.760	0.794	0.746
Friday	-0.266***	-0.228***	-0.221***	-0.269***	-0.240***	-0.260***
SE	(0.039)	(0.051)	(0.035)	(0.054)	(0.040)	(0.047)
IRR	0.766	0.796	0.802	0.764	0.787	0.771
N pupils	1,369	1,314	560	2,123	760	1,923
N	647,913	649,048	257,810	1,039,151	390,512	906,449
Clusters	78	78	78	78	78	78
Controls						
Half term, School, Half term*School	yes	yes	yes	yes	yes	yes
Week of half term	yes	yes	yes	yes	yes	yes
Pupil characteristics	yes	yes	yes	yes	yes	yes
Pupil FE	no	no	no	no	no	no
Random effects	no	no	no	no	no	no

Notes: Table shows results of zero-inflated negative binomial regressions of the number of behaviour incidents per pupil per day, relative to Mondays, using the same specification as column 7 of Table 5.6. Each column represents a separate regression. Coefficients and incidence rate ratios (IRRs) are reported, along with standard errors clustered at the level of school*half term. Variables included in the logit part of the model are: gender; SEN; FSM eligibility; and mean Key Stage 2 score.

Table 5.8: Behaviour incidents over the week by yeargroup
Outcome: behaviour incidents per pupil per day

	Year 7	Year 8	Year 9	Year 10	Year 11
Tuesday	-0.100	-0.099	-0.138**	-0.173***	-0.020
SE	(0.084)	(0.083)	(0.055)	(0.051)	(0.072)
IRR	0.905	0.905	0.871	0.841	0.981
Wednesday	-0.153**	-0.058	-0.324***	-0.283***	-0.275**
SE	(0.078)	(0.078)	(0.050)	(0.050)	(0.110)
IRR	0.858	0.943	0.724	0.754	0.760
Thursday	-0.151	-0.126	-0.257***	-0.402***	-0.216**
SE	(0.092)	(0.080)	(0.052)	(0.061)	(0.096)
IRR	0.860	0.882	0.773	0.669	0.806
Friday	-0.268***	-0.279***	-0.227***	-0.185***	-0.488***
SE	(0.092)	(0.074)	(0.048)	(0.056)	(0.088)
IRR	0.765	0.756	0.797	0.831	0.614
N pupils	1,553	1,778	1,842	1,376	908
N	284,599	309,356	321,564	234,707	146,906
Clusters	97	105	116	102	69
Controls					
Half term, School, Half term*School	yes	yes	yes	yes	yes
Week of half term	yes	yes	yes	yes	yes
Pupil characteristics	yes	yes	yes	yes	yes
Pupil FE	no	no	no	no	no
Random effects	no	no	no	no	no

Notes: Table shows results of zero-inflated negative binomial regressions of the number of behaviour incidents per pupil per day, relative to Mondays, using the same specification as column 7 of Table 5.6. Each column represents a separate regression. Coefficients and incidence rate ratios (IRRs) are reported, along with standard errors clustered at the level of school*half term. Variables included in the logit part of the model are: gender; SEN; FSM eligibility; and mean Key Stage 2 score. The number of pupils will sum to more than 3,284 because most pupils appear in the dataset over more than one academic year, so they will be included in at least two columns.

Table 5.9: Behaviour incidents over the week by school
Outcome: behaviour incidents per pupil per day

	School 1	School 2	School 3	School 4
Tuesday	-0.050	-0.278**	0.060	-0.227***
SE	(0.040)	(0.124)	(0.065)	(0.047)
IRR	0.951	0.758	1.062	0.797
Wednesday	-0.379***	-0.360***	-0.053	-0.222***
SE	(0.043)	(0.135)	(0.071)	(0.046)
IRR	0.685	0.698	0.948	0.801
Thursday	-0.263***	-0.187	-0.260***	-0.274***
SE	(0.056)	(0.155)	(0.079)	(0.046)
IRR	0.768	0.829	0.771	0.760
Friday	-0.188***	-0.296**	-0.145*	-0.316***
SE	(0.053)	(0.140)	(0.077)	(0.046)
IRR	0.829	0.744	0.865	0.729
N pupils	170	243	864	1,409
N	52,128	121,195	277,560	846,078
Clusters	38	64	99	311
Controls				
Half term, School, Half term*School	yes	yes	yes	yes
Week of half term	yes	yes	yes	yes
Pupil characteristics	yes	yes	yes	yes
Pupil FE	no	no	no	no
Random effects	no	no	no	no

Notes: Table shows results of zero-inflated negative binomial regressions of the number of behaviour incidents per pupil per day, relative to Mondays, using the same specification as column 7 of Table 5.6. Each column represents a separate regression. Coefficients and incidence rate ratios (IRRs) are reported, along with standard errors clustered at the level of school*month*cohort. Variables included in the logit part of the model are: gender; SEN; FSM eligibility; and mean Key Stage 2 score.

Table 5.10: Behaviour incidents by type and time of day

	Lesson 1	Lesson 2	Lesson 3	Lesson 4	Lesson 5	Other time	Total
N (pupil x time period observations)	346,359	346,359	346,359	346,359	346,359	346,359	2,078,154
Mean incidents per pupil per lesson	0.0042	0.0058	0.0066	0.0062	0.0074	0.0120	0.0070
SD	0.0665	0.0777	0.0825	0.0804	0.0877	0.1190	0.0872
Number of incidents	**1,455**	**2,010**	**2,285**	**2,158**	**2,568**	**4,158**	**14,634**
Broad type							
Missing categorisation	198	224	282	236	261	514	1,715
Disruption	535	950	936	1,051	1,096	282	4,850
Defiance	281	416	511	438	614	604	2,864
Late or truanting	125	127	195	147	198	1,544	2,336
Verbal offence	86	97	97	81	118	234	713
Dangerous behaviour or damage	48	58	62	65	69	173	475
Assault or fighting	40	32	42	40	59	225	438
Other (have got categorisation)	142	106	160	100	153	582	1,243

Notes: time of day information is only available for two schools (14,634 of 44,668 incidents). The ‘other times’ category includes morning and afternoon registrations, break and lunch times, the transitions between lessons, and before and after school. It also includes a very small number of observations which could not be assigned to a single time of day, because they were recorded as happening ‘all day’, ‘generally’ or in multiple lessons. This category therefore covers at least two hours of time, but I cannot be sure of the exact duration because of the imprecision of ‘before’ and ‘after’ school and of incidents happening around or between lessons.

Table 5.11: Behaviour incidents by day of the week and time of day (two schools)

		Time of day						
		Lesson 1	Lesson 2	Lesson 3	Lesson 4	Lesson 5	Other times	All times of day
Monday	Number of incidents	409	436	713	594	507	948	3,607
	Mean incidents per pupil	0.0060	0.0064	0.0105	0.0088	0.0075	0.0140	0.0089
	SD	0.0795	0.0820	0.1049	0.0959	0.0885	0.1321	0.0988
	N (pupil x time obs)	67,739	67,739	67,739	67,739	67,739	67,739	406,434
Tuesday	Number of incidents	312	457	499	518	657	945	3,388
	Mean incidents per pupil	0.0044	0.0064	0.0070	0.0073	0.0092	0.0133	0.0079
	SD	0.0674	0.0811	0.0847	0.0869	0.0976	0.1246	0.0921
	N (pupil x time obs)	71,199	71,199	71,199	71,199	71,199	71,199	427,194
Wednesday	Number of incidents	216	372	388	340	479	764	2,559
	Mean incidents per pupil	0.0030	0.0052	0.0054	0.0048	0.0067	0.0107	0.0060
	SD	0.0566	0.0738	0.0747	0.0698	0.0833	0.1121	0.0802
	N (pupil x time obs)	71,202	71,202	71,202	71,202	71,202	71,202	427,212
Thursday	Number of incidents	261	270	382	366	520	772	2,571
	Mean incidents per pupil	0.0037	0.0039	0.0054	0.0052	0.0074	0.0110	0.0061
	SD	0.0632	0.0627	0.0751	0.0728	0.0885	0.1133	0.0812
	N (pupil x time obs)	70,116	70,116	70,116	70,116	70,116	70,116	420,696
Friday	Number of incidents	257	475	303	340	405	729	2,509
	Mean incidents per pupil	0.0039	0.0072	0.0046	0.0051	0.0061	0.0110	0.0063
	SD	0.0638	0.0872	0.0684	0.0740	0.0788	0.1117	0.0822
	N (pupil x time obs)	66,103	66,103	66,103	66,103	66,103	66,103	396,618
Total	Number of incidents	1,455	2,010	2,285	2,158	2,568	4,158	14,634
	Mean incidents per pupil	0.0042	0.0058	0.0066	0.0062	0.0074	0.0120	0.0070
	SD	0.0665	0.0777	0.0825	0.0804	0.0877	0.1190	0.0872
	N (pupil x time obs)	346,359	346,359	346,359	346,359	346,359	346,359	2,078,154

Table 5.12: Behaviour incident regressions by time of day and day of the week

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	OLS	OLS	Negative binomial	Negative binomial	Negative binomial	Zero-inflated negative binomial	Zero-inflated negative binomial
Lesson 2	1.588***	1.588***	0.322***	0.322***	0.324***	0.322***	0.322***
SE	(0.463)	(0.463)	(0.076)	(0.077)	(0.035)	(0.078)	(0.077)
IRR			1.379	1.380	1.383	1.379	1.380
Lesson 3	2.367***	2.367***	0.448***	0.444***	0.447***	0.442***	0.444***
SE	(0.428)	(0.428)	(0.062)	(0.067)	(0.034)	(0.065)	(0.066)
IRR			1.566	1.558	1.564	1.556	1.559
Lesson 4	2.015***	2.015***	0.393***	0.396***	0.402***	0.398***	0.397***
SE	(0.419)	(0.419)	(0.056)	(0.058)	(0.035)	(0.058)	(0.058)
IRR			1.481	1.487	1.495	1.488	1.487
Lesson 5	3.190***	3.190***	0.566***	0.570***	0.573***	0.568***	0.572***
SE	(0.544)	(0.544)	(0.051)	(0.053)	(0.034)	(0.053)	(0.052)
IRR			1.762	1.769	1.774	1.765	1.772
Other time	7.807***	7.807***	1.053***	1.051***	1.033***	1.045***	1.047***
SE	(1.443)	(1.444)	(0.083)	(0.083)	(0.031)	(0.083)	(0.082)
IRR			2.865	2.860	2.811	2.844	2.849
Tuesday		-0.816		-0.084	-0.085***		-0.086
SE		(0.518)		(0.054)	(0.025)		(0.054)
IRR				0.919	0.918		0.918
Wednesday		-2.662***		-0.346***	-0.354***		-0.347***
SE		(0.761)		(0.063)	(0.027)		(0.063)
IRR				0.708	0.702		0.707
Thursday		-2.489***		-0.327***	-0.329***		-0.327***
SE		(0.616)		(0.057)	(0.028)		(0.056)
IRR				0.721	0.720		0.721
Friday		-2.171***		-0.266***	-0.274***		-0.269***
SE		(0.590)		(0.046)	(0.028)		(0.046)
IRR				0.766	0.760		0.764
Inflate							
Girl						0.338***	-0.997***
						(0.041)	(0.122)
Special educational needs						-1.246***	-16.948***
						(0.078)	(0.935)
Free school meals eligible						-0.888***	-0.066
						(0.048)	(0.135)
Mean Key Stage 2 score						0.175***	1.270***
						(0.051)	(0.093)
N pupils	1,096	1,096	1,096	1,034	1,034	1,034	1,034
N	2,078,154	2,078,154	2,078,154	1,978,128	1,978,128	1,978,128	1,978,128
Clusters	41	41	41	41		41	41

Controls							
Half term, School, Half term*School	no	yes	no	yes	yes	no	yes
Week of term	no	yes	no	yes	yes	no	yes
Pupil characteristics	no	no	no	yes	yes	no	yes
Pupil FE	no	yes	no	no	no	no	no
Random effects	no	no	no	no	yes	no	no

Notes: Table shows results of regressions of the number of behaviour incidents per pupil per lesson, relative to lesson 1. Columns 1 and 2 use ordinary least squares regressions, and the outcome variable is behaviour incidents per pupil per lesson multiplied by 1000. Columns 3-7 use the number of behaviour incidents per pupil per lesson as the outcome variable. Columns 3-5 use negative binomial regressions, and columns 6 and 7 use zero-inflated negative binomial regressions. Each column represents a separate specification. Coefficients and incidence rate ratios (IRRs) are reported, along with standard errors clustered at the level of school*month*year. Week of term controls include controls for days before and after holidays.

Table 5.13: Number of behaviour incidents by time of day and pupil characteristics
Outcome: behaviour incidents per pupil per lesson

	Boys	Girls	SEN	not SEN	FSM	not FSM
Lesson 2	0.423***	0.132	0.340***	0.307***	0.272***	0.345***
SE	(0.080)	(0.097)	(0.098)	(0.085)	(0.088)	(0.088)
IRR	1.526	1.141	1.406	1.360	1.313	1.412
Lesson 3	0.514***	0.318***	0.464***	0.421***	0.349***	0.488***
SE	(0.070)	(0.082)	(0.060)	(0.094)	(0.072)	(0.069)
IRR	1.671	1.375	1.591	1.523	1.417	1.629
Lesson 4	0.456***	0.291***	0.443***	0.354***	0.397***	0.395***
SE	(0.068)	(0.082)	(0.071)	(0.075)	(0.067)	(0.066)
IRR	1.578	1.338	1.558	1.425	1.488	1.485
Lesson 5	0.592***	0.531***	0.643***	0.502***	0.519***	0.592***
SE	(0.059)	(0.074)	(0.062)	(0.068)	(0.060)	(0.060)
IRR	1.808	1.700	1.902	1.652	1.681	1.808
Other time	1.040***	1.060***	1.120***	0.984***	1.032***	1.051***
SE	(0.080)	(0.105)	(0.069)	(0.110)	(0.084)	(0.089)
IRR	2.830	2.885	3.063	2.676	2.807	2.860
N pupils	555	479	215	819	159	875
N	1,062,300	915,828	410,712	1,567,416	298,968	1,679,160
Clusters	41	41	41	41	41	41
Controls						
Half term, School, Half term*School	yes	yes	yes	yes	yes	yes
Week of half term	yes	yes	yes	yes	yes	yes
Pupil characteristics	yes	yes	yes	yes	yes	yes
Pupil FE	no	no	no	no	no	no
Random effects	no	no	no	no	no	no

Notes: Table shows results of zero-inflated negative binomial regressions of the number of behaviour incidents per pupil per lesson, relative to lesson 1, using the same specification as column 7 of Table 5.12. Each column represents a separate regression. Coefficients and incidence rate ratios (IRRs) are reported, along with standard errors clustered at the level of school*month*year. Week of term controls include controls for days before and after holidays. Variables included in the logit part of the model are: gender; SEN; FSM eligibility; and mean Key Stage 2 score.

Table 5.14: Behaviour incidents by time of day and yeargroup
Outcome: behaviour incidents per pupil per lesson

	Year 7	**Year 8**	**Year 9**	**Year 10**	**Year 11**
Lesson 2	0.585***	0.287*	0.440***	0.367***	-0.493***
SE	(0.156)	(0.161)	(0.119)	(0.096)	(0.179)
IRR	1.794	1.332	1.552	1.443	0.611
Lesson 3	0.324**	1.075***	0.414***	0.256***	0.526***
SE	(0.158)	(0.171)	(0.098)	(0.092)	(0.172)
IRR	1.382	2.929	1.513	1.291	1.693
Lesson 4	0.586***	0.556***	0.436***	0.422***	-0.081
SE	(0.150)	(0.144)	(0.109)	(0.090)	(0.156)
IRR	1.797	1.743	1.547	1.525	0.922
Lesson 5	0.663***	0.935***	0.710***	0.312**	0.460***
SE	(0.187)	(0.127)	(0.094)	(0.121)	(0.178)
IRR	1.940	2.548	2.034	1.367	1.584
Other time	1.072***	1.035***	0.294**	1.338***	1.542***
SE	(0.189)	(0.155)	(0.135)	(0.146)	(0.139)
IRR	2.921	2.815	1.341	3.811	4.674
N pupils	393	402	462	418	188
N	441,750	444,300	504,360	442,830	176,118
Clusters	22	22	33	41	30
Controls					
Half term, School, Half term*School	yes	yes	yes	yes	yes
Week of half term	yes	yes	yes	yes	yes
Pupil characteristics	yes	yes	yes	yes	yes
Pupil FE	no	no	no	no	no
Random effects	no	no	no	no	no

Notes: Table shows results of zero-inflated negative binomial regressions of the number of behaviour incidents per pupil per lesson, relative to lesson 1, using the same specification as column 7 of Table 5.12. Each column represents a separate regression. Coefficients and incidence rate ratios (IRRs) are reported, along with standard errors clustered at the level of school*month*year. Week of term controls include controls for days before and after holidays. Variables included in the logit part of the model are: gender; SEN; FSM eligibility; and mean Key Stage 2 score. The number of pupils will sum to more than 1,096 because most pupils appear in the dataset over more than one academic year, so they will be included in at least two columns.

Table 5.15: Behaviour incidents by time of day, separately by school
Outcome: behaviour incidents per pupil per lesson

	School 1	**School 3**
Lesson 2	0.320***	0.325***
SE	(0.115)	(0.090)
IRR	1.377	1.384
Lesson 3	0.391***	0.515***
SE	(0.075)	(0.115)
IRR	1.478	1.673
Lesson 4	0.400***	0.388***
SE	(0.080)	(0.084)
IRR	1.492	1.474
Lesson 5	0.536***	0.616***
SE	(0.075)	(0.070)
IRR	1.709	1.851
Other time	1.119***	0.966***
SE	(0.112)	(0.118)
IRR	3.062	2.627
N pupils	170	864
N	312,768	1,665,360
Clusters	19	22
Controls		
Half term, School, Half term*School	yes	yes
Week of half term	yes	yes
Pupil characteristics	yes	yes
Pupil FE	no	no
Random effects	no	no

Notes: Table shows results of zero-inflated negative binomial regressions of the number of behaviour incidents per pupil per lesson, relative to lesson 1, using the same specification as column 7 of Table 5.12. Each column represents a separate regression. Coefficients and incidence rate ratios (IRRs) are reported, along with standard errors clustered at the level of school*month*year. Week of term controls include controls for days before and after holidays. Variables included in the logit part of the model are: gender; SEN; FSM eligibility; and mean Key Stage 2 score.

Table 5.16: Psychological variables by day of the week, restricted sample

		Self-reported behaviour score	Depression score	Anxiety score	Number of different days of week pupils were surveyed on
Monday	Mean	11.55	8.87	8.65	2.67
	SD	6.42	7.34	6.77	
	N pupils	2,308	2,313	2,297	
	N	2,933	2,952	2,911	
Tuesday	Mean	11.02	8.49	8.17	2.58
	SD	6.28	7.11	6.65	
	N pupils	2,937	2,947	2,928	
	N	3,889	3,915	3,886	
Wednesday	Mean	10.6	7.88	7.72	2.58
	SD	6.11	6.78	6.49	
	N pupils	2,705	2,721	2,705	
	N	4,488	4,531	4,493	
Thursday	Mean	10.34	7.70	7.93	2.57
	SD	6.18	6.85	6.59	
	N pupils	2,621	2,621	2,614	
	N	3,675	3,680	3,660	
Friday	Mean	10.73	8.18	8.11	2.73
	SD	6.33	7.32	6.73	
	N pupils	1,920	1,927	1,918	
	N	2,191	2,197	2,189	

Notes: Data are from a separate dataset of measures of psychological health, collected through questionnaires repeatedly administered to pupils over a period of three years. Depression score is measured using the Children's Depression Inventory (minus one item); anxiety score is measured using the Revised Children's Manifest Anxiety Scale; self-reported behaviour score is measured using the child-report Goodman Strengths and Difficulties Questionnaire. For all three of these measures a higher score indicates more or more severe symptoms.

Table 5.17: Self-reported behaviour score by day of the week
Outcome: standardised pupil-reported Goodman SDQ score

	(1)	(2)	(3)	(4)	(5)
Tuesday	-0.087***	-0.070**	-0.030	-0.031	-0.062**
	(0.030)	(0.028)	(0.026)	(0.026)	(0.027)
Wednesday	-0.146***	-0.106***	-0.043	-0.052**	-0.056**
	(0.035)	(0.032)	(0.028)	(0.025)	(0.028)
Thursday	-0.193***	-0.179***	-0.088***	-0.103***	-0.101***
	(0.036)	(0.034)	(0.030)	(0.030)	(0.033)
Friday	-0.130***	-0.105***	-0.084**	-0.101***	-0.090**
	(0.038)	(0.037)	(0.035)	(0.034)	(0.036)
N pupils	4,858	4,858	4,858	4,858	4,858
N	16,520	16,520	16,520	16,520	16,520
Clusters (classes)	159	159	159	159	159
R squared	0.004	0.056	0.081	0.068	0.725
Adjusted R squared				0.058	0.611
Pupil characteristics	no	yes	yes	no	no
School FE	no	no	yes	no	no
Month-year dummies	no	yes	yes	yes	yes
Class FE	no	no	no	yes	yes
Pupil FE	no	no	no	no	yes

Notes: Tables shows results of ordinary least squares regressions of pupil-reported Goodman SDQ scores standardised to have a mean of 0 and a standard deviation of 1. Regressions are clustered by class grouping, which was a major factor in determining which day pupils completed the survey, intraclass correlation coefficient on day of week by class = 0.21. A higher score indicates worse behaviour.

Table 5.18: Depression symptom scores by day of the week
Outcome: standardised Children's Depression Inventory score (pupil-reported)

	(1)	(2)	(3)	(4)	(5)
Tuesday	-0.058**	-0.047*	-0.015	-0.023	-0.046*
	(0.026)	(0.026)	(0.026)	(0.026)	(0.027)
Wednesday	-0.139***	-0.094***	-0.029	-0.046*	-0.047
	(0.033)	(0.031)	(0.026)	(0.025)	(0.029)
Thursday	-0.169***	-0.154***	-0.073***	-0.093***	-0.093***
	(0.034)	(0.030)	(0.028)	(0.029)	(0.033)
Friday	-0.100***	-0.076**	-0.043	-0.067**	-0.059
	(0.036)	(0.035)	(0.034)	(0.033)	(0.036)
N pupils	4858	4858	4858	4858	4858
N	16,605	16,605	16,605	16,605	16,605
Clusters (classes)	159	159	159	159	159
R squared	0.004	0.049	0.071	0.059	0.705
Adjusted R squared				0.050	0.582
Pupil characteristics	no	yes	yes	no	no
School FE	no	no	yes	no	no
Month-year dummies	no	yes	yes	yes	yes
Class FE	no	no	no	yes	yes
Pupil FE	no	no	no	no	yes

Notes: Tables shows results of ordinary least squares regressions of Children's Depression Inventory scores standardised to have a mean of 0 and a standard deviation of 1. Regressions are clustered by class grouping, which was a major factor in determining which day pupils completed the survey, intraclass correlation coefficient on day of week by class = 0.21. A higher score indicates more or more severe symptoms.

Table 5.19: Anxiety symptom scores by day of the week
Outcome: standardised Revised Manifest Children's Anxiety Scale score (pupil-reported)

	(1)	(2)	(3)	(4)	(5)
Tuesday	-0.077***	-0.056**	-0.035	-0.050**	-0.067***
	(0.028)	(0.027)	(0.027)	(0.025)	(0.023)
Wednesday	-0.141***	-0.083**	-0.024	-0.032	-0.030
	(0.035)	(0.033)	(0.029)	(0.027)	(0.028)
Thursday	-0.114***	-0.118***	-0.044	-0.070**	-0.077**
	(0.038)	(0.034)	(0.030)	(0.030)	(0.030)
Friday	-0.082**	-0.044	-0.035	-0.062*	-0.057*
	(0.037)	(0.036)	(0.034)	(0.033)	(0.031)
N pupils	4,854	4,854	4,854	4,854	4,854
N	16,485	16,485	16,485	16,485	16,485
Clusters (classes)	159	159	159	159	159
R squared	0.002	0.046	0.071	0.068	0.744
Adjusted R squared				0.058	0.637
Pupil characteristics	no	yes	yes	no	no
School FE	no	no	yes	no	no
Month-year dummies	no	yes	yes	yes	yes
Class FE	no	no	no	yes	yes
Pupil FE	no	no	no	no	yes

Notes: Tables shows results of ordinary least squares regressions of Revised Children's Manifest Anxiety Scale scores standardised to have a mean of 0 and a standard deviation of 1. Regressions are clustered by class grouping, which was a major factor in determining which day pupils completed the survey, intraclass correlation coefficient on day of week by class = 0.21. A higher score indicates more or more severe symptoms.

Table 5.20: Reported number of bedrooms at home by day of the week
Outcome: standardised pupil-reported number of bedrooms at home

	(1)	(2)	(3)	(4)	(5)
Tuesday	-0.036	-0.055**	-0.035	-0.024	-0.015
	(0.027)	(0.023)	(0.024)	(0.022)	(0.024)
Wednesday	0.034	0.000	-0.011	0.011	0.019
	(0.033)	(0.030)	(0.026)	(0.027)	(0.028)
Thursday	-0.025	-0.042	-0.038	-0.030	-0.022
	(0.033)	(0.030)	(0.025)	(0.026)	(0.029)
Friday	0.025	0.002	0.028	0.044	0.018
	(0.035)	(0.031)	(0.029)	(0.027)	(0.028)
N pupils	4,681	4,681	4,681	4,681	4,681
N	13,865	13,865	13,865	13,865	13,865
Clusters (classes)	159	159	159	159	159
R squared	0.001	0.037	0.047	0.053	0.760
Adjusted R squared				0.042	0.638
Pupil characteristics	no	yes	yes	no	no
School FE	no	no	yes	no	no
Month-year dummies	no	yes	yes	yes	yes
Class FE	no	no	no	yes	yes
Pupil FE	no	no	no	no	yes

Notes: Tables shows results of ordinary least squares regressions of pupils' reports of the number of bedrooms they have at home, standardised to have a mean of 0 and a standard deviation of 1 for comparison with the other measures. Regressions are clustered by class grouping.

Table 5.21: Teacher-reported pupil behaviour by day of the week
Outcome: standardised teacher-reported Goodman SDQ score

	(1)	(2)	(3)	(4)	(5)
Tuesday	-0.030	-0.023	0.015	0.017	0.009
	(0.056)	(0.050)	(0.047)	(0.044)	(0.052)
Wednesday	-0.034	-0.034	-0.012	0.019	-0.004
	(0.059)	(0.055)	(0.049)	(0.040)	(0.046)
Thursday	0.020	0.024	0.060	0.040	0.045
	(0.058)	(0.054)	(0.049)	(0.044)	(0.052)
Friday	0.016	0.015	0.056	0.048	0.038
	(0.055)	(0.051)	(0.049)	(0.043)	(0.048)
Saturday	-0.132	-0.175	-0.231	-0.112	-0.056
	(0.134)	(0.147)	(0.174)	(0.102)	(0.177)
Sunday	0.170	0.203*	0.244**	0.176*	0.192
	(0.128)	(0.120)	(0.099)	(0.095)	(0.123)
N pupils	4,282	4,282	4,282	4,282	4,282
N	13,591	13,591	13,591	13,591	13,591
Clusters (classes)	159	159	159	159	159
R squared	0.002	0.164	0.195	0.165	0.743
Adjusted R squared				0.155	0.624
Pupil characteristics	no	yes	yes	no	no
School FE	no	no	yes	no	no
Month-year dummies	no	yes	yes	yes	yes
Class FE	no	no	no	yes	yes
Pupil FE	no	no	no	no	yes

Notes: Tables shows results of ordinary least squares regressions of teacher-reported Goodman SDQ scores, standardised to have a mean of 0 and a standard deviation of 1. Regressions are clustered by class grouping. A higher score indicates worse behaviour.

Table 5.22: Absences, lateness and exclusions by time and day

		Absence		Authorised absence		Unauthorised absence		Lateness		Excluded		Excluded: 4 schools
		Morning	Afternoon	Morning	Afternoon	Morning	Afternoon	Morning	Afternoon	Morning	Afternoon	
Monday	Mean	0.0764	0.0887	0.0583	0.0697	0.0212	0.0190	0.076	0.0134	0.0010	0.0011	0.0029
	SD	0.2657	0.2844	0.2344	0.2546	0.1442	0.1366	0.265	0.1149	0.0323	0.0327	0.0536
	N	4,801	4,678	4,801	4,678	4,801	4,678	4,434	4,263	4,801	4,678	44,773
Tuesday	Mean	0.0738	0.0861	0.0558	0.0674	0.0204	0.0187	0.0791	0.0078	0.0012	0.0012	0.0029
	SD	0.2615	0.2805	0.2296	0.2508	0.1413	0.1354	0.2699	0.0878	0.0344	0.0349	0.0541
	N	5,052	4,925	5,052	4,925	5,052	4,925	4,679	4,501	5,052	4,925	47,713
Wednesday	Mean	0.0732	0.0838	0.057	0.0673	0.019	0.0164	0.0755	0.0126	0.0018	0.0018	0.0028
	SD	0.2604	0.2771	0.2318	0.2506	0.1365	0.1271	0.2643	0.1116	0.0422	0.0427	0.0528
	N	5,057	4,930	5,057	4,930	5,057	4,930	4,687	4,517	5,057	4,930	47,960
Thursday	Mean	0.0745	0.0855	0.0546	0.0653	0.0223	0.0203	0.0699	0.0040	0.0015	0.0015	0.0027
	SD	0.2625	0.2797	0.2273	0.247	0.1477	0.1409	0.2549	0.0629	0.0382	0.0386	0.0517
	N	4,795	4,688	4,795	4,688	4,795	4,688	4,438	4,287	4,795	4,688	47,723
Friday	Mean	0.0976	0.1112	0.0682	0.0810	0.0311	0.0302	0.0704	0.0091	0.0009	0.0009	0.0028
	SD	0.2968	0.3145	0.2522	0.2729	0.1736	0.1712	0.2558	0.0951	0.0293	0.0296	0.0525
	N	4,661	4,567	4,661	4,567	4,661	4,567	4,206	4,059	4,661	4,567	46,255

Notes: The first five columns give the fraction of pupils absent, late, or excluded by time of day and day of the week. The data comes from a fifth school involved in the UKRP evaluation, with a similar timetable and similar demographic characteristics to the four which provided behaviour incident data, N pupils=135, pupils aged 12 at start of academic year. The data are for the academic year 2007-08. The last column gives the fraction of pupils excluded at the four schools which provided behaviour data, again for the academic year 2007-08, N pupils=1,286. Daily exclusions data obtained from the National Pupil Database.

Table 5.23: Absence and lateness by day of the week and time

	Absence		Authorised absence		Unauthorised absence		Lateness		Excluded		Excluded: 4 schools	
	Probit	OLS FE	Probit	OLS FE	Probit	OLS FE	Probit	OLS FE	Probit	OLS FE	Probit	OLS FE
Morning	-0.083***	-0.015***	-0.096***	-0.013***	0.035	0.000	0.993***	0.064***	-0.034	-0.000		
	(0.017)	(0.002)	(0.019)	(0.002)	(0.028)	(0.001)	(0.032)	(0.002)	(0.093)	(0.000)		
Tuesday	0.005	-0.002	0.005	-0.002	-0.013	-0.000	-0.028	-0.001	0.056	0.000	0.012	0.000
	(0.028)	(0.004)	(0.030)	(0.003)	(0.046)	(0.002)	(0.037)	(0.003)	(0.152)	(0.000)	(0.043)	(0.000)
Wednesday	-0.016	-0.004	-0.002	-0.002	-0.061	-0.002	-0.032	-0.001	0.225	0.001	-0.007	-0.000
	(0.028)	(0.004)	(0.030)	(0.003)	(0.046)	(0.002)	(0.037)	(0.003)	(0.140)	(0.001)	(0.044)	(0.000)
Thursday	-0.001	-0.003	-0.015	-0.004	0.028	0.001	-0.133***	-0.007**	0.128	0.000	-0.027	-0.000
	(0.028)	(0.004)	(0.030)	(0.003)	(0.046)	(0.002)	(0.039)	(0.003)	(0.145)	(0.001)	(0.044)	(0.000)
Friday	0.150***	0.022***	0.097***	0.011***	0.196***	0.010***	-0.078**	-0.004	-0.070	-0.000	-0.011	-0.000
	(0.027)	(0.004)	(0.030)	(0.004)	(0.043)	(0.002)	(0.039)	(0.003)	(0.162)	(0.000)	(0.044)	(0.000)
N Pupils	123	135	123	135	123	135	123	135	123	135	1,202	1,199
N	43,813	48,154	43,437	48,154	43,813	48,154	39,993	44,071	17,897	48,154	222,046	218,445

Notes: Data sources as in Table 5.22: the first five column panels refer to pupils at the fifth school; the final column on exclusions is for pupils in the behaviour dataset. The outcome in each case is a dummy variable for each session for each pupil (there are two sessions per day, morning and afternoon). Probit regressions include controls for pupil characteristics; OLS regressions include pupil fixed effects and month dummies. The coefficient on each day shows the rate relative to Mondays, and the coefficient on 'morning' is relative to the afternoon session.

Table 5.24: Day of the week and time of day regressions excluding maths and science

	(1)	(2)
Lesson 2		0.528***
SE		(0.102)
IRR		1.696
Lesson 3		0.495***
SE		(0.071)
IRR		1.640
Lesson 4		0.499***
SE		(0.088)
IRR		1.648
Lesson 5		0.582***
SE		(0.090)
IRR		1.790
Tuesday	-0.097	-0.106
SE	(0.067)	(0.080)
IRR	0.908	0.900
Wednesday	-0.247***	-0.468***
SE	(0.090)	(0.085)
IRR	0.781	0.626
Thursday	-0.261***	-0.333***
SE	(0.060)	(0.071)
IRR	0.770	0.717
Friday	-0.292***	-0.524***
SE	(0.047)	(0.064)
IRR	0.747	0.592
N pupils	2,683	1,034
N	1,296,961	1,978,128
Clusters	78	41
Controls		
Half term, School, Half term*School	yes	yes
Week of term	yes	yes
Pupil characteristics	yes	yes

Notes: Table shows results of zero-inflated negative binomial regressions, taking out incidents which occurred in maths or science lessons or which did not have a lesson specified. Column 1 uses the same sample and specification as column 7 of Table 5.6, with the number of incidents per pupil per day as the outcome. Column 2 uses the same specification and sample as column 7 of Table 5.12, with the number of incidents per pupil per lesson as the outcome. In both regressions incidents taking place in maths or science lessons were excluded by setting the incident count to zero. The outcome variables therefore reflect the incident counts in all other subjects when the subject is known.

Table 5.25: School timetables

	School 1		School 2		School 3		School 4
8.30	Morning registration and assembly	8.35	Morning registration	8.40	Morning registration and assembly	8.30	Morning registration
8.50	Lesson 1	8.40	Lesson 1	9.00	Lesson 1	8.40	Lesson 1
9.55	Lesson 2	9.40	Lesson 2	10.00	Lesson 2	9.40	Lesson 2
11.00	Break	10.40	Break	11.00	Break	10.40	Break
11.20	Lesson 3	11.00	Lesson 3	11.20	Lesson 3	11.00	Lesson 3
12.20	Lesson 4	12.00	Lesson 4	12.20	Lesson 4	12.00	Lesson 4
13.20	Lunch	13.00	Lunch	13.20	Lunch	13.00	Lunch
14.10	Afternoon registration	13.45	Afternoon registration and assembly	14.05	Afternoon registration	13.40	Tutorial/assembly
14.15	Lesson 5	14.00	Lesson 5	14.15	Lesson 5	14.00	Lesson 5
15.15	End of formal schoolday	15.00	End of formal schoolday	15.15	End of formal schoolday	15.00	End of formal schoolday

Notes: This reflects the school timetables as they were in the academic year 2009-10, obtained from the schools' websites.

Conclusions

There are many aspects of secondary schools of importance to pupils other than academic attainment, and I have explored some of them here. My two overarching aims were to estimate the impact of the UK Resilience Programme, and to understand (bad) behaviour in secondary schools. I outline my main findings below, and draw out the policy implications. I then present the main research contributions of this thesis, and the limitations of the work I have done. I suggest some extensions for further work.

Main findings

Chapters 1-3 covered the design, implementation and findings of the UK Resilience Programme evaluation, a pragmatic controlled trial of a wellbeing intervention for 11-year-old pupils at secondary schools. The main research questions for these chapters concerned the quality of implementation, and the impact of the intervention on a range of outcomes. I found that intervention dosage was generally high – most pupils assigned to the intervention actually received most of the lessons. The intervention had small but statistically significant impacts on depressive symptom scores, absence from school, and popularity at postintervention, but not at later follow-up periods. It also had a small impact on academic attainment which lasted until the two-year follow-up period. The intervention had no impact on anxiety symptoms or behaviour.

I found some evidence of heterogeneity in intervention impact: higher ‘quality’ workshops – those carried out in smaller groups and for more hours – were associated with larger impacts on depressive symptoms, absence, academic attainment and popularity. There may also have been some heterogeneity in impact by pupil characteristics. The four significant intervention impacts appeared to be independent from one another: they were not mediated through the impact on any other outcome variable (for example, the impact on academic attainment was not mediated by the impact on the depression symptoms score). Scaling up interventions often results in a diminution of intervention quality, which can lead to smaller estimates of programme impact. Importantly, I found small effect sizes of the intervention despite relatively high programme dosage, suggesting that intervention quality was not a major problem and that the small effect sizes may be typical of programme impact in a universal population.

In Chapter 4 I used behaviour incident data from school databases to investigate pupil behaviour in secondary schools, specifically: who misbehaved, and whether behaviour was stable through time and in different contexts. I found that most pupils misbehave at least once, but that incidents are highly concentrated amongst a few pupils. Demographic characteristics were strong predictors of the number of incidents per pupil, but did not explain much of the overall variance in behaviour. I found that the incidence of poor behaviour varied substantially by context: by time of day, day of the week, and the subject being studied. However, pupils' behaviour ranking was very stable across different contexts and over time.

In Chapter 5 I looked at the variation in behaviour incident rates by time of day and day of the week in more detail, finding that they were highly robust and appeared to reflect genuine differences in behaviour rather than mere reporting biases. I investigated mechanisms, and suggested that these patterns may be related to pupils' subjective wellbeing, particularly tiredness and boredom, but that pupil absence and lateness; teacher wellbeing; and teacher absence might also contribute.

Policy implications

There are several factors which have a bearing on whether programmes such as UKRP offer value for money. One is whether pupils actually need the programme, or whether there is room for substantial improvement in their outcomes. For psychological wellbeing, this may not be the case for the majority of pupils, who do not experience serious mental distress. Another factor is whether the intervention tested is actually effective. It is possible that the pupils in this trial could have seen a large improvement in their wellbeing, but that this particular intervention was not able to produce this. Based on the results of this and previous PRP trials it would seem that most pupils do not gain much from PRP, but that particular pupils might gain more. One further issue is whether schools are the appropriate places to be providing this sort of intervention. Schools are certainly well placed to access pupils in need of support, but whether they should be responsible for delivering interventions remains controversial.

Given the cost of the UK Resilience Programme, and its small average impacts on four different outcomes (three of which impacts do not persist beyond postintervention), it is not clear that offering this or similar programmes universally offers value for money. However, if the duration of the intervention impact could be extended, a policymaker interested in all four outcomes considered together might find the programme more worthwhile. Universally provided programmes may also avoid stigma, and can act as screening tools to identify pupils in need of greater help. If screening practices are currently not very effective, and if the harm prevented through early identification of pupils with serious issues is sufficiently great, then universal programmes may be cost effective because of this.

The cost of the intervention is substantially affected by the number of workshop groups each trained facilitator can teach; thus if a school can keep the cost of the intervention down through efficient use of trained staff then providing UKRP universally could be worthwhile. It may also be worth considering whether there are implementation factors which could increase intervention impact even when programmes such as these are scaled up. Some of these are likely to relate to simple dosage – providing enough time to complete the curriculum – but careful selection and support of staff are likely to be very important.

My findings with respect to the behaviour incident data might also support investment in targeted interventions: less than 10% of pupils were responsible for over 50% of behaviour incidents, and the tendency to misbehave was stable through time. This suggests that interventions capable of improving the behaviour of the worst-behaved pupils could be cost-effective even if these interventions were expensive, because of the very large reduction in incidents that could result. Ameliorating behaviour in this way could substantially decrease everyday disruption, and make schools more pleasant places to work and learn.

My finding that context has a sizeable impact on behaviour incident rates would also argue for whole-school approaches to behaviour management. Specifically, if demanding subjects could be scheduled at low-incident times of day, this could reduce the overall number of incidents as well as promoting pupil achievement. Since this intervention would be almost costless, even small impacts on behaviour and academic attainment resulting from it could be highly cost effective. However, such a policy should be tested by means of an RCT before being implemented. Consideration should also be given to other ways in which schools can make the context more conducive to good behaviour, particularly for pupils with a tendency to misbehave. For example, this could involve changes in teaching style to promote engagement.

My descriptive work on behaviour also suggested that teacher- and pupil-reports were not particularly good measures of behaviour. I would suggest that schools, LAs and researchers make more use of the behaviour incident data which many schools collect in order to understand where there are problems and whether interventions and strategies are having an effect. In particular, linking data on exclusions with incident data could clarify which pupils are at risk and why.

The pragmatic controlled trial around which this thesis is based was funded by a government department (DCSF, now DfE). However, the detailed data collected for the trial has not been made more widely available for research use, and may well be deleted. This is not a good use of public funding and does not provide value for money: datasets are a major output of any commissioned research, and should be anonymised and made available to researchers at other institutions after the end of a project.

Contributions to the literature

The UKRP evaluation is one of a very few scaled up interventions of this kind, and my use of a robust pragmatic trial design, a relatively long follow-up period, and a range of outcome measures allows a more general assessment of intervention impact than can be seen for most wellbeing or mental health interventions. In addition, I provide clear cost information and present basic analyses of cost-effectiveness, which can be used to compare my results with those of other interventions aiming to impact depressive symptoms, absence from school, academic attainment or popularity. In addition, my use of a large number of friendship nominations to estimate the intervention impact on popularity is novel, and my results suggest that paying more attention to the development and maintenance of social capital and social relationships could be valuable, as these can be lost as well as gained.

The behaviour data I used in Chapters 4 and 5 represents a substantial improvement over other attempts to measure and describe behaviour, contributing to both research and policy in this area. I am also able to observe behaviour in different contexts and through time, leading to inferences about the malleability and rank order stability of the tendency to misbehave. These are important to understanding how to improve behaviour in schools, as well as being important to an understanding of behaviour more generally.

The nature of scheduling in schools also allows me to make a substantial contribution to the literature on time of day and day of the week effects, as there is less endogeneity in activities and reporting in the behaviour data I use than in most of the literature. My results relating to behaviour in schools could generalise to other areas of life, suggesting a way of promoting efficiency and wellbeing in the workplace.

Limitations and topics for future research

The most important weakness of the UKRP evaluation was the lack of randomisation to condition assignment. In addition, it is not clear that all of the outcome measures were appropriate for use in a universal sample, or that they were able to identify the core concept of resilience. I was also lacking baseline data in the evaluation of the intervention impact on popularity. It is increasingly feasible to carry out RCTs in English schools, and I would recommend that any future intervention research uses randomised assignment of pupils, classes or yeargroups in order to generate a plausible estimate of intervention impact.

The (temporary) lack of attainment data from national exams is also a limitation for both the UKRP evaluation and the work using the behaviour data. Noncognitive aspects of secondary school are important and are the subject of this thesis, but academic attainment is very valuable and is a major factor in expanding life chances. I would have preferred to be able to relate the noncognitive aspects of school to pupils' exam results at 16, and given the quality of the data I have collected this could be an important future piece of research.

Although behaviour is an important outcome in its own right, we would need to understand the impact of scheduling on academic attainment before recommending scheduling changes to schools. Linking GCSE data into the behaviour work I presented above would allow a better understanding of this process. However, the ideal solution would be to design and implement a randomised controlled trial, exogenously varying the scheduling of lessons within schools and measuring the impact on academic attainment and behaviour incidents. This would provide good evidence as to whether such a policy ultimately has an impact on academic attainment.

Evaluations of school-based interventions should provide better information on costs under different implementation scenarios, in order to promote comparisons between different interventions aiming to impact similar outcomes. In addition, a range of outcomes should be assessed in order to understand the global impact of an intervention. Using data from school and national databases can be a relatively costless way of doing this. Trials should be pragmatic where possible, in order to provide relevant and accessible information on intervention impact to policymakers. Moreover, further research is needed into factors associated with the loss of intervention impact when programmes are implemented widely, in order to build a better understanding of what a successful intervention implementation involves.

References

Abbott, J., Klein, B., Hamilton, C., & Rosenthal, A. (2009). The impact of online resilience training for sales managers on wellbeing and work performance. *eJournal of Applied Psychology, 5*(1), 89-95.

Ahern, N. R., Kiehl, E. M., Sole, M. L., & Byers, J. (2006). A review of instruments measuring resilience. *Issues in Comprehensive Pediatric Nursing, 29*(2), 103-125.

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). Personality Psychology and Economics. *NBER Working Paper 16822*.

Aluja-Fabregat, A., Balleste-Almacellas, J., & Torrubia-Beltri, R. (1999). Self-reported personality and school achievement as predictors of teachers' perceptions of their students. *Personality and Individual Differences, 27*(4), 743-753.

Ammermueller, A., & Pischke, J.-S. (2009). Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics, 27*(3), 315-348.

Angrist, J. D. (2003). Randomized Trials and Quasi-Experiments in Education Research. *NBER Reporter: Research Summary*

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press.

Anson, J., & Anson, O. (2000). Thank God It's Friday: The Weekly Cycle of Mortality in Israel. *Population Research and Policy Review, 19*(2), 143-154.

APHO. (2011). IDACI 2010: Income Domain Affecting Children Index Score from the Indices of Deprivation 2010 applied to mid-2009 population estimates: Association of Public Health Observatories.

Areni, C. S., & Burger, M. (2008). Memories of “Bad” Days Are More Biased Than Memories of “Good” Days: Past Saturdays Vary, but Past Mondays Are Always Blue. *Journal of Applied Social Psychology 38*(6), 1395–1415.

Atkinson, M., Lamont, E., & Downing, D. (2007). CAMHS funding and priorities. Slough: National Foundation for Educational Research.

ATL. (2012). A third of education staff have dealt with physical violence from pupils in this school year, with parents failing to back schools http://www.atl.org.uk/media-office/media-archive/A-third-of-education-staff-have-dealt-with-physical-violence-from-pupils-in-this-school-year.asp. *ATL Press Release*

Babcock, P. (2008). From Ties to Gains? Evidence on Connectedness and Human Capital Acquisition. *Journal of Human Capital*, *2*(4), 379-409.

Bandiera, O., Barankay, I., & Rasul, I. (2009). Social Connections and Incentives in the Workplace: Evidence from Personnel Data. *Econometrica*, *77*(4), 1047-1094.

Bardone, A. M., Moffitt, T. E., Caspi, A., Dickson, N., Stanton, W. R., & Silva, P. A. (1998). Adult physical health outcomes of adolescent girls with conduct disorder, depression, and anxiety. *Journal of the American Academy of Child and Adolescent Psychiatry, 37*(6), 594-601.

Baron, R. M., & Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182.

Barrett, P. M., Webster, H. M., & Wallis, J. R. (1999). Adolescent Self-Esteem and Cognitive Skills Training: A School-Based Intervention. *Journal of Child and Family Studies, 8*(2), 217-227.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497-529.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspectives on Psychological Science, 2*(4), 396-403.

BBC (2008). Tests scrapped for 14-year-olds Retrieved 03/08/13, from http://news.bbc.co.uk/1/hi/education/7669254.stm

BBC (2011). School leavers 'need work skills and knowledge', 11 March 2011 Retrieved 03/08/13, from http://www.bbc.co.uk/news/education-12701594

Beck, A. T. (1967). *The diagnosis and management of depression*. Philadelphia, PA: University of Pennsylvania Press.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*(6), 561–571.

Berndt, T. J. (2004). Children's friendships: Shifts over a half century in perspectives on their development and their effects. *Merrill-Palmer Quarterly, 50*(3), 206–223.

Berndt, T. J., Hawkins, J. A., & Jiao, Z. (1999). Influences of Friends and Friendships on Adjustment to Junior High School. *Merrill-Palmer Quarterly, 45*(1), 13-41.

Bertrand, M., Luttmer, E. F. P., & Mullainathan, S. (2000). Network Effects and Welfare Cultures. *Quarterly Journal of Economics, 115*(3), 1019-1055.

Biederman, J., Monuteaux, M. C., Doyle, A. E., Seidman, L. J., Wilens, T. E., Ferrero, F., Morgan, C. L., Faraone, S. V. (2004). Impact of executive function deficits and attention-deficit/hyperactivity disorder (ADHD) on academic outcomes in children. *Journal of Consulting and Clinical Psychology, 72*(5), 757-766.

Bierman, K. L. (2003). *Peer Rejection: Developmental Processes and Intervention Strategies*. New York, NY: The Guilford Press.

Bierman, K. L., & Wargo, J. B. (1995). Predicting the longitudinal course associated with aggressive rejected, aggressive (nonrejected), and rejected (nonaggressive) status. *Development and Psychopathology, 7*(4), 669-682.

Bland, J.M. (2009). The tyranny of power: is there a better way to calculate sample size? *British Medical Journal, 339,* b3985.

Bloom, N., & Van Reenen, J. (2007). Measuring and Explaining Management Practices Across Firms and Nations. *Quarterly Journal of Economics, 122*(4), 1351-1408.

Bondonio, D. (1998). Predictors of accuracy in perceiving informal social networks. *Social Networks, 20*(4), 301-330.

Borghans, L., Duckworth, A. L., Heckman, J. J., & 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. (2006). Time Discounting and the Body Mass Index: Evidence from the Netherlands. *Economics and Human Biology, 4*(1), 39-61.

Botvin, G. J., Baker, E., Dusenbury, L., Botvin, E. M., & Diaz, T. (1995). Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. *Journal of the American Medical Association, 273*(14), 1106-1112.

Bouguen, A., & Gurgand, M. (2012). Randomized Controlled Experiments in Education *EENEE Analytical Report No. 11*: European Commission European Expert Network on Economics of Education.

Bowles, S., Gintis, H., & Osborne, M. (2001). The Determinants of Earnings: A Behavioral Approach. *Journal of Economic Literature, 39*(4), 1137-1176.

Bramoullé, Y., Djebbari, H., & Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics, 150*(1), 41-55.

Brown v. Board of Education of Topeka, Opinion; May 17, 1954; Records of the Supreme Court of the United States; Record Group 267; National Archives.

Brunwasser, S. M., Gillham, J. E., & Kim, E. S. (2009). A Meta-Analytic Review of the Penn Resiliency Program's Effect on Depressive Symptoms. *Journal of Consulting and Clinical Psychology, 77*(6), 1042-1054.

Bryson, A., & Forth, J. (2007). Are There Day of the Week Productivity Effects? *Manpower Human Resources Lab Discussion Paper, No. 4, London School of Economics*.

Bryson, A., & MacKerron, G. (2012). *Are You Happy While You Work?*

Buhrmester, D. (1990). Intimacy of friendship, interpersonal competence, and adjustment during preadolescence and adolescence. *Child Development, 61*(4), 1101-1111.

Buhs, E. S., & Ladd, G. W. (2001). Peer rejection as an antecedent of young children's school adjustment: an examination of mediating processes. *Developmental Psychology, 37*(4), 550-560.

Burgess, S., & Umaña-Aponte, M. (2011). *Raising your sights: the impact of friendship networks on educational aspirations*. CMPO Working Papers. University of Bristol. Retrieved from http://www.bristol.ac.uk/cmpo/publications/papers/2011/wp271.pdf

Calear, A. L., & Christensen, H. (2010). Systematic review of school-based prevention and early intervention programs for depression. *Journal of Adolescence, 33*(3), 429-438.

Calear, A. L., Christensen, H., Mackinnon, A., Griffiths, K. M., & O'Kearney, R. (2009). The YouthMood Project: a cluster randomized controlled trial of an online cognitive behavioral program with adolescents. *Journal of Consulting and Clinical Psychology, 77*(6), 1021-1032.

Calvó-Armengol, A., Patacchini, E., & Zenou, Y. (2009). Peer Effects and Social Networks in Education. *Review of Economic Studies, 76*(4), 1239-1267.

Card, D. (1999). The Causal Effect of Education on Earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 3A). Amsterdam: Elsevier.

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

Card, D., & McCall, B. P. (1996). Is Workers' Compensation Covering Uninsured Medical Costs? Evidence from the “Monday Effect”. *Industrial and Labor Relations Review, 49*(4), 690-706.

Carrell, S. E., Maghakian, T., & West, J. E. (2011). A's from Zzzz's? The Causal Effect of School Start Time on the Academic Achievement of Adolescents. *American Economic Journal: Economic Policy, 3*(3), 1-22.

Cauce, A. M. (1986). Social networks and social competence: exploring the effects of early adolescent friendships. *American Journal of Community Psychology, 14*(6), 607-628.

CBI (2012). Business and education leaders back CBI calls for school improvement, 19 November 2012, from http://www.cbi.org.uk/media-centre/news-articles/2012/11/business-back-cbi-calls-for-school-improvement/

CEM (2013). MidYIS - Introduction Retrieved 03/08/13, from http://www.cem.org/midyis

Challen, A. R., & Bailey, L. (2012). The UK Penn Resilience Programme: A summary of research and implementation. *Psychology of Education Review, 36*(2), 32-39.

Challen, A. R., Machin, S. J., & Gillham, J. E. (forthcoming). The UK Resilience Programme: a school-based universal non-randomised pragmatic controlled trial. *Journal of Consulting and Clinical Psychology.*

Challen, A. R., Machin, S. J., Noden, P., & West, A. (2009). UK Resilience Programme Evaluation: Interim Report: Department for Children, Schools and Families.

Challen, A. R., Machin, S. J., Noden, P., & West, A. (2010). UK Resilience Programme Evaluation: Second Interim Report: Department for Education.

Challen, A. R., Machin, S. J., Noden, P., & West, A. (2011). UK Resilience Programme Evaluation: Final Report: Department for Education.

Chamorro-Premuzic, T., & Furnham, A. (2003). Personality Predicts Academic Performance: Evidence from Two Longitudinal University Samples. *Journal of Research in Personality, 37*(4), 319-338.

Chan, A.W., Hrobjartsson, A., Haahr, M.T., Gøtzsche, P.C., Altman, D.G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials—comparison of protocols to published articles. *Journal of the American Medical Association, 291*(20), 2457-65.

Chan, A.W., Krleza-Jerić, K., Schmid, I., Altman, D.G. (2004). Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal, 171*(7), 735-40.

Chenet, L., McKee, M., Leon, D., Shkolnikov, V., & Vassin, S. (1998). Alcohol and cardiovascular mortality in Moscow; new evidence of a causal association. *Journal of Epidemiology and Community Health, 52*(12), 772-774.

Chowdry, H., & Sibieta, L. (2011). Trends in Education and Schools Spending *IFS Briefing Note*. London: IFS.

Coe, R. (2002). *It's the Effect Size, Stupid: What effect size is and why it is important*. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter. http://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Coie, J. D. (2004). The Impact of Negative Social Experiences on the Development of Antisocial Behavior. In J. B. Kupersmidt & K. A. Dodge (Eds.), *Children's peer relations: From development to intervention* (pp. 243-267). Washington, DC: American Psychological Association.

Coie, J. D., Lochman, J. E., Terry, R., & Hyman, C. (1992). Predicting early adolescent disorders from childhood aggression and peer rejection. *Journal of Consulting and Clinical Psychology, 60*(5), 783-792.

Coleman, J., & Hagell, A. (2007). The nature of risk and resilience in Adolescence. In J. Coleman & A. Hagell (Eds.), *Adolescence, risk and resilience: against the odds*. Chichester: John Wiley & Sons.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology, 94*, S95-S120.

Collishaw, S., Maughan, B., Goodman, R., & Pickles, A. (2004). Time trends in adolescent mental health. *Journal of Child Psychology and Psychiatry, 45*(8), 1350-1362.

Conduct Problems Prevention Research Group (1999). Initial impact of the Fast Track prevention trial for conduct problems: II. Classroom effects. *Journal of Consulting and Clinical Psychology, 67*(5), 648-657.

Conduct Problems Prevention Research Group (2002). An end of third-grade evaluation of the impact of the Fast Track prevention trial with children of high risk for adolescent conduct problems. *Journal of Abnormal Child*

Psychology, 30(1), 19-35.

Conti, G., Galeotti, A., Mueller, G., & Pudney, S. (2012). *Popularity*. NBER Working Paper No. 18475. Retrieved from http://www.nber.org/papers/w18475.pdf?new_window=1

Coolidge, F. L., DenBoer, J. W., & Segal, D. L. (2004). Personality and neuropsychological correlates of bullying behavior. *Personality and Individual Differences, 36*(7), 1559-1569.

Cortes, K. E., Moussa, W. S., & Weinstein, J. M. (2012). *Making the Grade: The Impacts of Classroom Disruption and Class Size on Academic Achievement*.

Costa-Font, J., & Jofre-Bonet., M. (2013). Anorexia, Body Image and Peer Effects: Evidence from a Sample of European Women. *Economica, 80*(317), 44–64.

Costello, D. M., Swendsen, J., Rose, J. S., & Dierker, L. C. (2008). Risk and Protective Factors Associated with Trajectories of Depressed Mood from Adolescence to Early Adulthood. *Journal of Consulting and Clinical Psychology, 76*(2), 173-183.

Covey, L. S., Glassman, A. H., & Stetner, F. (1998). Cigarette smoking and major depression. *Journal of Addictive Diseases, 17*, 35-46.

Crawford, C., Dearden, L., & Greaves, E. (2011). Does when you are born matter? The impact of month of birth on children's cognitive and non-cognitive skills in England. London: Institute for Fiscal Studies.

Csikszentmihalyi, M. (1999). If We Are So Rich, Why Aren't We Happy? *American Psychologist, 54*(10), 821-827.

Csikszentmihalyi, M., & Hunter, J. (2003). Happiness in Everyday Life: The Uses of Experience Sampling. *Journal of Happiness Studies, 4*(2), 185-199.

Currarini, S., Jackson, M. O., & Pin, P. (2009). An Economic Model of Friendship: Homophily, Minorities, and Segregation. *Econometrica, 77*(4), 1003-1045.

DCLG (2008). The English Indices of Deprivation 2007 *Statistics*. London: Department for Communities and Local Government.

DCLG (2011). The English Indices of Deprivation 2010 *Statistics*. London: Department for Communities and Local Government.

DCSF (2006). National Curriculum Assessments at Key Stage 2 in England, 2006 (Provisional): Department for Children, Schools and Families.

DCSF (2007a). Schools and Pupils in England: January 2007 (Final). London: Department for Children, Schools and Families.

DCSF (2007b). DCSF: National Curriculum Assessments at Key Stage 2 in England, 2007 (Revised): Department for Children, Schools and Families.

DCSF (2008a). Youth Cohort Study and Longitudinal Study of Young People in England: The activities and experiences of 16 year olds: England 2007 *DCSF Statistical Bulletin*. London: DCSF.

DCSF (2008b). Pupil Characteristics and Class Sizes in Maintained Schools in England: January 2008 (Provisional). London: Department for Children, Schools and Families.

DCSF (2009a). Children with Special Educational Needs 2009: an analysis. London: Department for Children, Schools and Families.

DCSF (2009b). Schools, Pupils and Their Characteristics: January 2009. London: Department for Children, Schools and Families.

DCSF (2009c). National Curriculum Assessments at Key Stage 2 in England 2009 London: Department for Children, Schools and Families.

DCSF (2009d). Measuring Progress at Pupil, School and National levels: Department for Children, Schools and Families.

De Giorgi, G., Pellizzari, M., & Redaelli, S. (2010). Identification of Social Interactions through Partially Overlapping Peer Groups. *American Economic Journal: Applied Economics*, *2*(2), 241-275.

De Neve, J.-E., & Oswald, A. J. (2012). Estimating the influence of life satisfaction and positive affect on later income using sibling fixed effects. *Proceedings of the National Academy of Sciences*, *109*(49), 19953-19958.

Deaton, A. (2009). *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development*. The Keynes Lecture, 9th October 2008, British Academy, London.

DfE (2010a). What matters most in education - your views. London: Department for

Education.

DfE (2010b). The Importance of Teaching: Schools White Paper. London: Department for Education.

DfE (2010c). Consultation on school funding 2011-12: Introducing a pupil premium: Department for Education.

DfE (2011). New powers for teachers to improve discipline in schools, 04 April 2011. *Department for Education press notice* Retrieved 29/11/12, from http://www.education.gov.uk/inthenews/inthenews/a0076417/new-powers-for-teachers-to-improve-discipline-in-schools

DfE (2012a). Pupil behaviour in schools in England *DfE Research Reports*. London: Department for Education.

DfE (2012b). Behaviour and discipline in schools, A guide for head teachers and school staff Retrieved 29/11/12, from http://www.education.gov.uk/aboutdfe/advice/f0076803/behaviour-and-discipline-in-schools-a-guide-for-headteachers-and-school-staff

DfE (2012c). Schools Funding Settlement 2012-13 including Pupil Premium. London: Department for Education.

DfE (2012d). Pupil Absence in Schools in England, Including Pupil Characteristics: 2010/11. London: Department for Education.

DfE (2012e). Length of school day/year Retrieved 29/11/12, from http://www.education.gov.uk/schools/adminandfinance/schooladmin/schoolyear/a0064221/length-of-school-dayyear

DfE (2012f). Special educational needs in England: January 2012, from https://www.gov.uk/government/publications/special-educational-needs-in-england-january-2012

DfE (2013a). Personal, Social, Health and Economic (PSHE) education 22 March 2013. Retrieved 03/08/13, from http://www.education.gov.uk/schools/teachingandlearning/curriculum/b00223087/pshe

DfE (2013b). Statutory Guidance: The duty to secure independent and impartial careers guidance for young people in schools. London: Department for Education.

DfE (2013c). The school curriculum Retrieved 04/08/13, from http://www.education.gov.uk/schools/teachingandlearning/curriculum

DfE (2013d). Funding for postgraduate teacher training Retrieved 04/08/13, from http://www.education.gov.uk/get-into-teaching/funding/postgraduate-funding

DfE (2013e). GCSE and Equivalent Attainment by Pupil Characteristics in England, 2011/12 Retrieved 08/03/13, from http://www.education.gov.uk/rsgateway/DB/SFR/s001111/index.shtml

DfE (2013f). School and Local Statistics: St Clement Danes School Retrieved 04/08/13, from http://www.education.gov.uk/cgi-bin/schools/performance/school.pl?urn=136901

DfE (2013g). School funding reform Retrieved 06/08/13, from http://www.education.gov.uk/schools/adminandfinance/financialmanagement/schoolsrevenuefunding/a00215225/school-funding-reform

Diener, E., & Seligman, M. E. P. (2002). Very happy people. *Psychological Science 13*(1), 81-84.

Dobeš, M., Fedáková, D., Lehotská, V., & Koscelníková, K. (2010). *Effects of Long-term Social-Psychological Intervention Programme in Primary Schools*. Retrieved from http://www.academia.edu/551376/Effects_of_Long-term_Social-Psychological_Intervention_Programme_in_Primary_Schools

Dodge, K. A. (1993). *Social information-processing and peer rejection factors in the development of behavior problems in children.* Paper presented at the Biennial meeting of the Society for Research in Child Development, New Orleans, LA.

Dodge, K. A. (2006). Translational science in action: Hostile attributional style and the development of aggressive behavior problems. *Development and Psychopathology, 18*(3), 791-814.

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-Discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science,* *16*(12), 939-944.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed Self-Assessment: Implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest, 5*(3), 69-106.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405-432.

EEF (2013). Education Endowment Foundation - Projects Retrieved 03/08/13, from http://educationendowmentfoundation.org.uk/projects

Egloff, B., Tausch, A., Kohlmann, C.-W., & Krohne, H. W. (1995). Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. *Motivation and Emotion, 19*(2), 99-110.

Eisenberg, D., & Neighbors, K. (2009). Benefits and Costs of Prevention. In M. E. O'Connell, T. Boat & K. E. Warner (Eds.), *Preventing Mental, Emotional, and Behavioral Disorders Among Young People*. Washington, DC: National Academies Press.

Eisner, M. (2009). No effects in independent prevention trials: can we reject the cynical view? *Journal of Experimental Criminology, 5*(2), 163-183.

Ellis, A. (1962). *Reason and emotion in psychotherapy*. New York, NY: Lyle Stewart.

Ellis, A. (1977). The Basic Clinical Theory of Rational Emotive Therapy In A. Ellis & R. Grieger (Eds.), *Handbook of rational-emotive therapy* (pp. 3-34). New York: Springer Pub. Co.

Ennett, S. T., & Bauman, K. E. (2000). Adolescent Social Networks: Friendship Cliques, Social Isolates, and Drug Use Risk. In W. B. Hansen, S. M. Giles & M. Fearnow-Kenney (Eds.), *Improving Prevention Effectiveness*: Tanglewood Research.

Ercolani, M. G. (2006). *UK Employees' Sickness Absence: 1984-2005*. Working paper. Department of Economics, University of Birmingham. Retrieved from ftp://ftp.bham.ac.uk/pub/RePEc/pdf/Ercolani0602.pdf

Evans, C., Chalmers, J., Capewell, S., Redpath, A., Finlayson, A., Boyd, J., Pell, J., McMurray, J., Macintyre, K., & Graham, L. (2000). “I don't like Mondays”— day of the week of coronary heart disease deaths in Scotland: study of routinely collected data. *BMJ, 320*(2189).

Fergusson, D. M., Horwood, L. J., & Ridder, E. M. (2005). Show me the child at seven: the consequences of conduct problems in childhood for psychosocial functioning in adulthood. *Journal of Child Psychology and Psychiatry, 46*(8), 837-849.

Fergusson, D. M., & Woodward, L. J. (2000). Educational, psychosocial, and sexual outcomes of girls with conduct problems in early adolescence. *Journal of Child Psychology and Psychiatry, 41*(6), 779-792.

Fergusson, D. M., & Woodward, L. J. (2002). Mental health, educational, and social role outcomes of adolescents with depression. *Archives of General Psychiatry, 59*(3), 225-231.

Finn, J. D. (1989). Withdrawing from School. *Review of Educational Research, 59*(2), 117-142.

Folkard, S., & Lombardi, D. A. (2006). Modelling the Impact of the Components of Long Work Hours on Injuries and “Accidents”. *American Journal of Industrial Medicine, 49*(11), 953-963.

Formby, E., Coldwell, M., Stiell, B., Demack, S., Stevens, A., Shipton, L., Wolstenholme, C., Willis, B. (2011). Personal, Social, Health and Economic (PSHE) Education: A mapping study of the prevalent models of delivery and their effectiveness *Research Reports* London: Department for Education.

Försterling, F., & Binser, M. J. (2002). Depression, school performance and the veridicality of perceived grades and causal attributions. *Personality and Social Psychology Bulletin, 28*(10), 1441–1449.

Fortson, K. N. (2004). The Diurnal Pattern of On-the-Job Injuries. *Monthly Labor Review, 127*(9), 18-25.

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European*

Journal of Personality, 23(5), 369-401.

Galton, M., Gray, J., & Ruddock, J. (1999). The Impact of School Transitions and Transfers on Pupil Progress and Attainment. Norwich: DfEE publications.

Gapin, J. I., Labban, J. D., & Etnier, J. L. (2011). The effects of physical activity on attention deficit hyperactivity disorder symptoms: The evidence. *Preventive Medicine*, *52*(Supplement), S70-S74.

Garrison, C. Z., Jackson, K. L., Addy, C. L., McKeown, R. E., & Waller, J. L. (1991). Suicidal behaviors in young adolescents. *American Journal of Epidemiology, 133*(10), 1005-1014.

Gest, S. D. (2006). Teacher Reports of Children's Friendships and Social Groups: Agreement with Peer Reports and Implications for Studying Peer Similarity. *Social Development, 15*(2), 248–259.

Gibbons, S., & Chevalier, A. (2008). Assessment and Age 16+ Education Participation. *Research Papers in Education, 23*(2), 113-123.

Gibbons, S., & Silva, O. (2011). Faith schools: better schools or better pupils. *Journal of Labor Economics, 29*(3), 589-635.

Gillham, J. E., Brunwasser, S. M., & Freres, D. R. (2008). Preventing depression in early adolescence: The Penn Resiliency Program. In J. R. Z. Abela & B. L. Hankin (Eds.), *Handbook of Depression in Children and Adolescents* (pp. 309-332). New York: Guilford Press.

Gillham, J. E., & Reivich, K. (2007). Resilience Research in Children Retrieved 24/07/2013, from http://www.ppc.sas.upenn.edu/prpsum.htm

Gillham, J. E., Reivich, K., & Jaycox, L. H. (2008a). *Penn Resiliency Curriculum: Leader's Manual*. Philadelphia, PA: University of Pennsylvania.

Gillham, J. E., Reivich, K., & Jaycox, L. H. (2008b). *Penn Resiliency Curriculum: Student's Notebook.* Philadelphia, PA: University of Pennsylvania.

Gillham, J. E., Reivich, K. J., Freres, D. R., Chaplin, T. M., Shatté, A. J., Samuels, B., Elkon, A.G.L., Litzinger, S., Lascher, M., Gallop, R., & Seligman, M. E. P. (2007). School-based prevention of depressive symptoms: A randomized controlled study of the effectiveness and specificity of the Penn Resiliency Program. *Journal of Consulting and Clinical Psychology, 75*, 9-19.

Gillham, J. E., Reivich, K. J., Freres, D. R., Lascher, M., Litzinger, S., Shatté, A. J., & Seligman, M. E. P. (2006). School-based prevention of depression and anxiety symptoms in early adolescence: A pilot of a parent intervention component. *School Psychology Quarterly, 21*, 323-348.

GL assessment (2012). CAT3 Cognitive Abilities Test Third Edition Retrieved 03/08/13, from http://www.gl-assessment.co.uk/products/cat3-cognitive-abilities-test-third-edition

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry, 38*(5), 581-586.

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(11), 1337-1345.

Gotlib, I. H., Lewinsohn, P. M., & Seeley, J. R. (1995). Symptoms versus diagnosis of depression: Differences in psychosocial functioning. *Journal of Consulting and Clinical Psychology, 63*(1), 90-100.

Graham, C., & Hill, M. (2002). The Transition to Secondary School: Glasgow Centre for the Child & Society: University of Glasgow.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380.

Granovetter, M. (1995). *Getting a Job: Study of Contacts and Careers* (2nd ed.). Chicago: University of Chicago Press.

Green, F., Machin, S., Murphy, R., & Zhu, Y. (2008). Competition for private and state school teachers. *Journal of Education and Work, 21*(5), 383-404.

Green, H., McGinnity, A., Meltzer, H., Ford, T., & Goodman, R. (2005). Mental health of children and young people in Great Britain, 2004. London: Office for National Statistics.

Greenberg, M. T., Kusche, C. A., & Cook, E. T. (1995). Promoting emotional competence in school-aged children: The effects of the PATHS curriculum. *Development & Psychopathology, 7*(1), 117-136.

Greene, W. H. (2003). *Econometric Analysis* (5th ed.): Pearson Education International.

Grucza, R. A., & Goldberg, L. R. (2007). The Comparative Validity of 11 Modern Personality Inventories: Predictions of Behavioral Acts, Informant Reports, and Clinical Indicators. *Journal of Personality Assessment, 89*(2), 167-187.

Gutman, L. M., & Vorhaus, J. (2012). The Impact of Pupil Behaviour and Wellbeing on Educational Outcomes. London: Department for Education.

Hagell, A., Peck, S. C., Zarrett, N., Giménez-Nadal, J. I., & Symonds, J. (2012). Trends in adolescent time use in the United Kingdom. In A. Hagell (Ed.), *Changing adolescence: Social trends and mental health* (pp. 47-73). Bristol: The Policy Press.

Hallam, S., Rhamie, J., & Shaw, J. (2006). The Evaluation of the Primary Behaviour and Attendance Pilot. In D. f. E. a. Skills (Ed.): Institute of Education.

Hall-Lande, J. A., Eisenberg, M. E., Christenson, S. L., & Neumark-Sztainer, D. (2007). Social isolation, psychological health, and protective factors in adolescence. *Adolescence, 42*(166), 265-286.

Hanushek, E. A. (2003). The Failure of Input-based Schooling Policies. *Economic Journal, 113*(485), F64–F98.

Hargreaves, D. H. (1996). *Teaching as a Research-based Profession: possibilities and prospects*. Paper presented at the Teacher Training Agency Annual Lecture.

Harrington, R., Fudge, H., Rutter, M., Pickles, A., & Hill, J. (1990). Adult outcomes of childhood and adolescent depression. *Archives of General Psychiatry, 47*(5), 465-473.

Hartup, W. W. (1996). The company they keep: Friendships and their developmental significance. *Child Development, 67*(1), 1-13.

Hattie, J. A. C. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses*

Relating to Achievement. Abingdon, Oxon: Routledge.

Haynes, L., Service, O., Goldacre, B., & Torgerson, D. (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. London: Cabinet Office Behavioural Insights Team.

Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics, 54*(1), 3-56.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market. *Journal of Labor Economics, 24*(3), 411-482.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128.

Helson, R., Kwan, V. S. Y., John, O. P., & Jones, C. (2002). The Growing Evidence for Personality Change in Adulthood: Findings from Research with Personality Inventories. *Journal of Research in Personality, 36*(4), 287-306.

Henry, D. B., & Metropolitan Area Child Study Research Group (2006). Associations between peer nominations, teacher ratings, self-reports, and observations of malicious and disruptive behavior. *Assessment, 13*(3), 241-252.

Higgins, S., Kokotsaki, D., & Coe, R. (2012). The Teaching and Learning Toolkit. London: Education Endowment Foundation.

Humphrey, N., Lendrum, A., & Wigelsworth, M. (2010). Social and emotional aspects of learning (SEAL) programme in secondary schools: national evaluation Department for Education.

Hymel, S., Rubin, K. H., Rowden, L., & LeMare, L. (1990). Children's Peer Relationships: Longitudinal Prediction of Internalizing and Externalizing Problems from Middle to Late Childhood. *Child Development, 61*(6), 2004–2021.

Ialongo, N., Edelsohn, G., Werthamer-Larsson, L., Crockett, L., & Kellam, S. (1996). Social and cognitive impairment in first grade children with anxious and depressive symptoms. *Journal of Clinical Child Psychology, 25*(1), 15-24.

Ichino, A., & Moretti, E. (2009). Biological Gender Differences, Absenteeism, and the Earnings Gap. *American Economic Journal: Applied Economics, 1*(1), 183-218.

IRIS (2013). Welcome to IRIS - Schools Data Services Ltd. Retrieved 07/08/13, from http://www.iris.ac/home

Jessor, R., Turbin, M. S., & Costa, F. M. (1998). Risk and protection in successful outcomes among disadvantaged adolescents. *Applied Developmental Science, 2*(4), 194-208.

Jonas, S. (2011). Who's paying for mental health services for young people? Retrieved 19/12/2012, from http://www.kingsfund.org.uk/blog/2011/09/whos-paying-mental-health-services-young-people

J-PAL. (2013). History [of J-PAL] Retrieved 03/08/13, from http://www.povertyactionlab.org/History

Juvonen, J. (2006). Sense of Belonging, Social Bonds, and School Functioning. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 655-674). Mahwah, NJ: Lawrence Erlbaum Associates.

Keslair, F., Maurin, E., & McNally, S. (2011). *An Evaluation of "Special Educational Needs" Programmes in England*. Working paper. Centre for the Economics of Education. London.

Kim-Cohen, J., Caspi, A., Moffitt, T. E., Harrington, H., Milne, B. J., & Poulton, R. (2003). Prior juvenile diagnoses in adults with mental disorder. *Archives of General Psychiatry, 60*(7), 709-717.

Kirby, E. G., & Kirby, S. L. (2006). Improving task performance: The relationship between morningness and proactive thinking. *Journal of Applied Social Psychology, 36*(11), 2715-2729.

Klima, T., & Repetti, R. L. (2008). Children's Peer Relations and Their Psychological Adjustment Differences between Close Friendships and the Larger Peer Group *Merrill-Palmer Quarterly, 54*(2), 151-178.

Knapp, M., McDaid, D., & Parsonage (Eds.), M. (2011). Mental Health Promotion and Prevention: The Economic Case. London: Department of Health.

Kovacs, M. (2003). *Manual of the Children's Depression Inventory*. Toronto: Multi-Heath Systems.

Kozma, A., Stone, S., & Stones, M. J. (2000). Stability in components and predictors of subjective well-being (SWB): Implications for SWB structure. In E. Diener & D. R. Rahtz (Eds.), *Advances in Quality of Life Theory and Research*. Dordrecht, NL: Kluwer.

Kuhn, P., & Weinberger, C. (2005). Leadership skills and wages. *Journal of Labor Economics, 23*(3), 395-436.

Ladd, G. W. (1990). Having friends, keeping friends, making friends, and being liked by peers in the classroom: predictors of children's early school adjustment? *Child Development, 61*(4), 1081-1100.

Ladd, G. W. (2006). Peer rejection, aggressive or withdrawn behavior, and psychological maladjustment from ages 5 to 12: an examination of four predictive models. *Child Development, 77*(4), 822-846.

Laugeson, E. A., Frankel, F., Gantman, A., Dillon, A. R., & Mogil, C. (2012). Evidence-Based Social Skills Training for Adolescents with Autism Spectrum Disorders: The UCLA PEERS Program. *Journal of Autism and Developmental Disorders, 42*(6), 1025-1036.

Laugeson, E. A., Frankel, F., Mogil, C., & Dillon, A. R. (2009). Parent-Assisted Social Skills Training to Improve Friendships in Teens with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders, 39*(4), 596–606.

Lavy, V., Passerman, D., & Schlosser, A. (2012). Inside the Black Box of Ability Peer Effects: Evidence from Variation in Low Achievers in the Classroom. *Economic Journal, 122*(559), 208-237.

Lavy, V., & Sand, E. (2012). *The Friends Factor: How Students' Social Networks Affect Their Academic Achievement and Well-Being?* NBER Working Paper No. 18430. Retrieved from http://www.nber.org/papers/w18430

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

Layard, R., Clark, A., & Senik, C. (2012). The Causes of Happiness and Misery. In J. Helliwell, R. Layard & J. Sachs (Eds.), *World Happiness Report* (pp. Appendix B, p89). New York: Earth Institute.

Layard, R., & CEP Mental Health Policy Group, (2006). The Depression Report: A New Deal for Depression and Anxiety Disorders. London: Centre for Economic Performance, London School of Economics.

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

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review, 53*(1), 1-18.

Lindahl, M. (2001). *Summer Learning and the Effect of Schooling: Evidence from Sweden*. IZA Discussion Paper. Institute for the Study of Labor (IZA). Bonn.

Lindqvist, E., & Vestman, R. (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics, 3*(1), 101-128.

Little, P. M. D., Wimer, C., & Weiss, H. B. (2008). After School Programs in the 21st Century: Their Potential and What it Takes to Achieve It. In H. F. R. Project (Ed.), *Issues and opportunities in out-of-school time evaluation*. Cambridge, MA: Harvard Graduate School of Education.

Lochman, J. E., Boxmeyer, C., Powell, N., Qu, L., Wells, K., & Windle, M. (2009). Dissemination of the Coping Power Program: Importance of intensity of counselor training. *Journal of Consulting and Clinical Psychology, 77*, 397-409.

Lucas, R. E., & Donnellan, M. B. (2011). Personality Development Across the Life Span: Longitudinal Analyses With a National Sample From Germany. *Journal of Personality and Social Psychology, 101*(4), 847–861.

Luthar, S. S. (Ed.). (2003). *Resilience and vulnerability*. Cambridge: Cambridge University Press.

Machin, S. J., Telhaj, S., & Wilson, J. (2006). The Mobility of English School

Children. *Fiscal Studies, 27*(3), 253-280.

MacKerron, G., & Mourato, S. (2010). Mappiness blog Retrieved 29/11/12, from http://blog.mappiness.org.uk/

Mahar, M. T. (2011). Impact of short bouts of physical activity on attention-to-task in elementary school children. *Preventive Medicine, 52*(Supplement), S60-S64.

Malcolm, H., Wilson, V., Davidson, J., & Kirk, S. (2003). Absence from School: A study of its causes and effects in seven LEAs *Research Report 424*. London: DfES.

Maliphant, R., Hume, F., & Furnham, A. (1990). Autonomic nervous system (ANS) activity, personality characteristics and disruptive behaviour in girls. *Journal of Child Psychology and Psychiatry, 31*(4), 619-628.

Malti, T., Ribeaud, D., & Eisner, M. (2011). The effectiveness of two universal preventive interventions in reducing children's externalizing behavior: a cluster randomized controlled trial. *Journal of Clinical Child & Adolescent Psychology, 40*(5), 677-692.

Manski, C.F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies, 60*(3), 531-542.

Masia-Warner, C., Nangle, D. W., & Hansen, D. J. (2006). Bringing evidence-based child mental health services to the schools: General issues and specific populations. *Education & Treatment of Children, 29*(2), 165-172.

McGee, C., Ward, R., Gibbons, J., & Harlow, A. (2004). Transition to Secondary School: A Literature Review. Hamilton: The University of Waikato.

Merry, S. N., Hetrick, S. E., Cox, G. R., Brudevold-Iversen, T., Bir, J. J., & McDowell, H. (2011). Psychological and educational interventions for preventing depression in children and adolescents. *Cochrane Database of Systematic Reviews, 12 Art. No.: CD003380*. doi: 10.1002/14651858.

Middleton, S., Perren, K., Maguire, S., Rennison, J., Battistin, E., Emmerson, C., & Fitzsimons, E. (2005). Evaluation of Education Maintenance Allowance Pilots: Young People Aged 16 to 19 Years - Final Report of the Quantitative

Evaluation *DfES Research Report 678*. London: Department for Education and Skills.

Mihalopoulos, C., & Vos, T. (2013). Cost-effectiveness of Preventive Interventions for Depressive Disorders: An Overview. *Expert Review of Pharmacoeconomics & Outcomes Research, 13*(2), 237-242.

Miller, D. C., Sen, A., Malley, A. B., & Burns, S. D. (2009). Comparative Indicators of Education in the United States and other G-8 Countries: 2009 Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S Department of Education.

Miller, R. T. (2008). Tales of Teacher Absence. Washington, DC: Center for American Progress.

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P.J., Elbourne, D., Egger, M., Altman, D.G. & Consolidated Standards of Reporting Trials Group (2010). CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. Journal of Clinical Epidemiology, 63(8), 1-37.

Moody, J. (2001). Race, School Integration, and Friendship Segregation in America. *American Journal of Sociology, 107*(3), 679–716.

Mortimer, D., & Segal, L. (2008). Comparing the Incomparable? A Systematic Review of Competing Techniques for Converting Descriptive Measures of Health Status into QALY-Weights. *Medical Decision Making, 28*(1), 66-89.

Mueller, G., & Plug, E. (2006). Estimating the Effect of Personality on Male and Female Earnings. *Industrial and Labor Relations Review, 60*(1), 3-22.

Muris, P., Meesters, C., Eijkelenboom, A., & Vincken, M. (2004). The self-report version of the Strengths and Difficulties Questionnaire: its psychometric properties in 8- to 13-year-old non-clinical children. *British Journal of Clinical Psychology, 34*(Pt 4), 437-448.

Murphy, R. (2011). *The Impact of UK Penn Resilience Programme training on Staff* Unpublished manuscript. Centre for Economic Performance: London School of Economics.

NFER (2012). NFER Teacher Voice Omnibus February 2012 Survey: Pupil Behaviour *DFE Research Report*. London: Department for Education.

NICE (2006). Methods for development of NICE public health guidance. London: National Institute for Health and Clinical Excellence.

Offord, D. R., Kraemer, H. C., Kazdin, A. E., Jensen, P. S., & Harrington, R. (1998). Lowering the burden of suffering from child psychiatric disorder: trade-offs among clinical, targeted, and universal interventions. *Journal of the American Academy of Child and Adolescent Psychiatry, 37*(7), 686-694.

Ofsted (2005). Managing challenging behaviour HMI 2363, from http://www.ofsted.gov.uk/Ofsted-home/Publications-and-research/Browse-all-by/Care/Childcare/Managing-challenging-behaviour

Ofsted (2012a). The framework for school inspection from January 2012. London: Ofsted.

Ofsted (2012b). Official statistics: Maintained school inspections and outcomes, from http://www.ofsted.gov.uk/resources/official-statistics-maintained-school-inspections-and-outcomes

Ofsted (2013). Not yet good enough: personal, social, health and economic education in schools. London: Ofsted.

Oliver, S., Bagnall, A. M., Thomas, J., Shepherd, J., Sowden, A., White, I., Dinnes, J., Rees, R., Colquitt, J., Oliver, K., & Garrett, Z. (2010). Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technology Assessment, 14*(16), 1-165.

ONS (2003). *Census 2001: Key Statistics for local authorities in England and Wales*. London: Office for National Statistics: TSO.

ONS (2012). Annual Survey of Hours and Earnings, 2012 Provisional Results: Office for National Statistics.

Oswald, A. J., Proto, E., & Sgroi, D. (2009). *Happiness and Productivity*. IZA Discussion Papers. Institute for the Study of Labor (IZA). Bonn.

Parker, J. G., & Asher, S. R. (1987). Peer relations and later personal adjustment: are low-accepted children at risk? *Psychological Bulletin, 102*(3), 357-389.

Parkhurst, J. T., & Hopmeyer, A. (1998). Sociometric Popularity and Peer-Perceived Popularity: Two Distinct Dimensions of Peer Status. *Journal of Early Adolescence, 18*(2), 125-144.

Parsons, T. (1959). The school class as a social system. *Harvard Educational Review, 29*(4), 297-318.

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience*, *13*(2), 217-224.

Payton, J. W., Weissberg, R. P., Durlak, J. A., Dymnicki, A. B., Taylor, R. D., Schellinger, K. B., & Pachan, M. (2008). The positive impact of social and emotional learning for kindergarten to eighth-grade students: Findings from three scientific reviews. Chicago, Illinois: CASEL (Collaborative for Academic, Social and Emotional Learning).

Petrides, K. V., Frederickson, N., & Furnham, A. (2004). The role of trait emotional intelligence in academic performance and deviant behavior at school. *Personality and Individual Differences, 36*(2), 277-293.

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology, 1*(4), 435-450.

Pettengill, G. N. (2003). A Survey of the Monday Effect Literature. *Quarterly Journal of Business and Economics, 42*(3-4), 3-27.

Phillips, B., Ball, C., Sackett, D., Badenoch, D., Straus, S., Haynes, B., Dawes, M., & Howick, J. (2009). Oxford Centre for Evidence-Based Medicine - Levels of Evidence Retrieved 03/08/13, from http://www.cebm.net/?o=1025

Postlewaite, A., & Silverman, D. (2005). Social Isolation and Inequality. *Journal of Economic Inequality, 3*(3), 243-262.

Preckel, F., Lipnevich, A. A., Boehme, K., Brandner, L., Georgi, K., Könen, T., Mursin, K., & Roberts, R. D. (2012). Morningness-eveningness and educational outcomes: The lark has an advantage over the owl at high school *British Journal of Educational Psychology*.

Rees, G., Goswami, H., Pople, L., Bradshaw, J., Keung, A., & Main, G. (2013). The Good Childhood Report 2013. London: The Children's Society.

Reinherz, H. Z., Paradis, A. D., Giaconia, R. M., Stashwick, C. K., & Fitzmaurice, G. (2003). Childhood and adolescent predictors of major depression in the transition to adulthood. *American Journal of Psychiatry, 160*(12), 2141-2147.

Reis, H. T., & Collins, W. A. (2004). Relationships, Human Behavior, and Psychological Science. *Current Directions in Psychological Science*, *13*(6), 233-237.

Reivich, K., & Gillham, J. E. (2008). *Penn Resiliency Programme Core Skills Manual*. Philadelphia, PA.

Resing, W. C. M., Bleichrodt, N., & Dekker, P. H. (1999). Measuring Personality Traits in the Classroom. *European Journal of Personality, 13*(12), 493-509.

Reynolds, C. R., & Richmond, B. O. (1985). *Revised Children's Manifest Anxiety Scale Manual*. Los Angeles: Western Psychological Services.

Reynolds, W. M. (1992). Depression in children and adolescents. In W. M. Reynolds (Ed.), *Internalizing disorders in children and adolescents*. New York: Wiley.

Roberts, B. W. (2007). Contextualizing Personality Psychology. *Journal of Personality, 75*(6), 1071-1082.

Roberts, B. W., Harms, P. D., Caspi, A., & Moffitt, T. E. (2007). Predicting the Counterproductive Employee in a Child-to-Adult Prospective Study. *Journal of Applied Psychology, 92*(5), 1427–1436.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science*, *2*(4), 313-345.

Roland, M., & Torgerson, D. J. (1998). What are pragmatic trials? *British Medical Journal, 316*(7127), 285.

Roseth, C. J., Johnson, D. W., & Johnson, R. T. (2008). Promoting early adolescents' achievement and peer relationships: the effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin, 134*(2), 223-246.

Rothon, C., Head, J., Clark, C., Klineberg, E., Cattell, V., & Stansfeld, S. (2009). The impact of psychological distress on the educational achievement of adolescents at the end of compulsory education. *Social Psychiatry and Psychiatric Epidemiology, 44*(5), 421-427.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet, 365*(9453), 82-93.

Roza, S. J., Hofstra, M. B., van der Ende, J., & Verhulst, F. C. (2003). Stable prediction of mood and anxiety disorders based on behavioral and emotional problems in childhood: a 14-year follow-up during childhood, adolescence, and young adulthood. *American Journal of Psychiatry*, *160*(12), 2116-2121.

Rubin, K. H., Bukowski, W. M., & Parker, J. G. (2006). Peer interactions, relationships and groups. In W. Damon, R. M. Lerner & N. Eisenberg (Eds.), *Handbook of child psychology* (6th ed., Vol. 3: Social, emotional, and personality development, pp. 571–645). New York: Wiley.

Rucklidge, J. J., & Tannock, R. (2001). Psychiatric, psychosocial, and cognitive functioning of female adolescents with ADHD. *Journal of the American Academy of Child Adolescent Psychiatry*, *40*(5), 530-540.

Rutter, M. (2006). Implications of Resilience Concepts for Scientific Understanding. In B. M. Lester, A. Masten & B. McEwen (Eds.), *Resilience in Children* (pp. 1-12). Malden, Mass.: Blackwell Publishing.

Rutter, M., & Smith, D. J. (Eds.). (1995). *Psychosocial Disorders in Young People. Time Trends and Their Causes*. Chichester: John Wiley & Sons.

Salmivalli, C., Huttunen, A., & Lagerspetz, K. M. J. (1997). Peer networks and bullying in schools. *Scandinavian Journal of Psychology, 38*(4), 305–312.

Sawyer, M. G., Harchak, T. F., Spence, S. H., Bond, L., Graetz, B., Kay, D., Patton, G., & Sheffield, J. (2010). School-based prevention of depression: A 2-year follow-up of a randomized controlled trial of the beyondblue schools research initiative. *Journal of Adolescent Health*, *47*(3), 297-304.

Sawyer, M. G., Pfeiffer, S., Spence, S. H., Bond, L., Graetz, B., Kay, D., Patton, G., & Sheffield, J. (2010). School-based prevention of depression: a randomised controlled study of the beyondblue schools research initiative. *Journal of*

Child Psychology and Psychiatry, 51(2), 199-209.

Schwartz, D., & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases, 20*(8), 637-648.

Scott, S., Knapp, M. R. J., Henderson, J., & Maughan, B. (2001). Financial cost of social exclusion: follow up study of antisocial children into adulthood. *British Medical Journal, 323*, 191-195.

Segal, C. (2008). Classroom Behavior. *Journal of Human Resources, 43*(4), 783–814.

Segal, C. (forthcoming). Misbehavior, Education, and Labor Market Outcomes. *Journal of the European Economic Association*.

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist, 55*(1), 5-14.

Shahar, G., Henrich, G., Winokur, A., Blatt, S., Kuperminc, G., & Leadbeater, B. (2006). Self-criticism and depressive symptomatology interact to predict middle school academic achievement. *Journal of Clinical Psychology, 62*(1), 147–155.

Simo, S., & Perez, J. (1991). Sensation seeking and antisocial behaviour in a junior student sample. *Personality and Individual Differences*, *12*(9), 965-966.

Smith, S., Windmeijer, F., & Wright, E. (2013). *Peer effects in charitable giving: Evidence from the (running) field*. CMPO Working Papers. University of Bristol. Retrieved from http://www.bristol.ac.uk/cmpo/publications/papers/2013/wp302.pdf

Specht, J., Egloff, B., & Schmukle, S. C. (2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology, 101*(4), 862-882.

Stallard, P., Sayal, K., Phillips, R., Taylor, J. A., Spears, M., Anderson, R., Araya, R., Lewis, G., Millings, A., & Montgomery, A. A. (2012). Effect of classroom based cognitive behavioural therapy on symptoms of depression in high risk adolescents: pragmatic cluster randomised controlled trial. *British Medical Journal, 345*(e6058).

Stansfeld, S. A., Clark, C., Rodgers, B., Caldwell, T., & Power, C. (2011). Repeated exposure to socioeconomic disadvantage and health selection as life course pathways to mid-life depressive and anxiety disorders. *Social Psychiatry and Psychiatric Epidemiology, 46*(7), 549-558.

StataCorp. (2011). Stata 12. College Station, TX: StataCorp.

Steinberg, L. (2007). Risk Taking in Adolescence: New Perspectives from Brain and Behavioral Science. *Current Directions in Psychological Science, 16*(2), 55-59.

Stice, E., Shaw, H., Bohon, C., Marti, C. N., & Rohde, P. (2009). A meta-analytic review of depression prevention programs for children and adolescents: factors that predict magnitude of intervention effects. *Journal of Consulting and Clinical Psychology, 77*(3), 486-503.

Stolberg, H. O., Norman, G., & Trop, I. (2004). Randomized controlled trials. *American Journal of Roentgenology, 183*(6), 1539-1544.

Stone, A. A., Schkade, D., Schwarz, J. E., Schwarz, N., Krueger, A. B., & Kahneman, D. (2006). A population approach to the study of emotion: diurnal rhythms of a working day examined with the Day Reconstruction Method. *Emotion, 6*, 139-149.

Strauss, R. S., & Pollack, H. A. (2003). Social marginalization of overweight children. *Archives of Pediatrics & Adolescent Medicine, 157*(8), 746-752.

Taylor, C. (2013). Behaviour and the tail. In P. Marshall (Ed.), *The Tail: How England's schools fail one child in five – and what can be done* (pp. 175-186). London: Profile Books Ltd.

Taylor, M. P. (2006). Tell Me Why I Don't Like Mondays: Investigating Day of the Week Effects on Job Satisfaction and Psychological Well-Being. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(1), 127-142.

Terracciano, A., McCrae, R. R., & Costa, P. T. (2010). Intra-individual Change in Personality Stability and Age. *Journal of Research in Personality, 44*(1), 31-37.

Thomas, M. S. C. (2012). Brain plasticity and education. *British Journal of*

Educational Psychology - Monograph Series II: Educational Neuroscience, 8, 142-156.

Thorpe, K. E., Zwarenstein, M., Oxman, A. D., Treweek, S., Furberg, C. D., Altman, D. G., Tunis, S., Bergel, E., Harvey, I., Magid, D.J., & Chalkidou, K. (2009). A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Canadian Medical Association Journal, 180*(10), E47-57.

Tooley, J., & Darby, D. (1998). Educational Research: a Critique: a Survey of Published Educational Research. London: Office for Standards in Education.

Torges, C. M., Stewart, A. J., & Nolen-Hoeksema, S. (2008). Regret Resolution, Aging, and Adapting to Loss. *Psychology and Aging, 23*(1), 169-180.

Treweek, S., & Zwarenstein, M. (2009). Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials, 10*(37).

Uher, R., & McGuffin, P. (2010). The moderation by the serotonin transporter gene of environmental adversity in the etiology of depression: 2009 update. *Molecular Psychiatry, 15*(1), 18-22.

UNC Carolina Population Center: The National Longitudinal Study of Adolescent Health (Add Health). (2013) Retrieved 18/06/13, from http://www.cpc.unc.edu/projects/addhealth/

Vaquera, E., & Kao, G. (2008). Do You Like Me as Much as I Like You? Friendship Reciprocity and Its Effects on School Outcomes among Adolescents. *Social Science Research, 37*(1), 55–72.

Vernon, H. M. (1921). *Industrial Fatigue and Efficiency*. London: George Routledge & Sons, Ltd.

Waasdorp, T. E., Bradshaw, C. P., & Leaf, P. J. (2012). The impact of schoolwide positive behavioral interventions and supports on bullying and peer rejection: a randomized controlled effectiveness trial. *Archives of Pediatrics & Adolescent Medicine, 166*(2), 149-156.

Weinberg, B. (2007). *Social Interactions and Endogenous Association*. NBER Working Paper No. 13038. Retrieved from http://www.nber.org/papers/w13038

Weisz, J. R., & Jensen, A. L. (2001). Child and adolescent psychotherapy in research and practice contexts: review of the evidence and suggestions for improving the field. *European Child and Adolescent Psychiatry, 10*(Suppl 1), 12-18.

Weisz, J. R., Jensen-Doss, A. L., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis. *American Psychologist, 61*(7), 671–689.

Wentzel, K. R., Barry, C. M., & Caldwell, K. A. (2004). Friendships in Middle School: Influences on Motivation and School Adjustment. *Journal of Educational Psychology, 96*(2), 195-203.

Wetz, J. (2009). Urban village schools: putting relationships at the heart of secondary school organisation and design London: Calouste Gulbenkian Foundation.

WHO (2004). Prevention of Mental Disorders: Effective interventions and policy options. Geneva: World Health Organization.

Willich, S. N., Löwel, H., Lewis, M., Hörmann, A., Arntz, H. R., & Keil, U. (1994). Weekly variation of acute myocardial infarction. Increased Monday risk in the working population. *Circulation, 90*(1), 87-93.

Wilson, J. M., & Marcotte, A. C. (1996). Psychosocial adjustment and educational outcome in adolescents with a childhood diagnosis of attention deficit disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*(5), 579-587.

Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American Journal of Preventive Medicine, 33*(Suppl. 2S), 130–143.

Witvliet, M., van Lier, P. A. C., Cuijpers, P., & Koot, H. M. (2009). Testing links between childhood positive peer relations and externalizing outcomes through a randomized controlled intervention study. *Journal of Consulting and Clinical Psychology, 77*(5), 905-915.

Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement, 55*(2), 177–185.

Wolfson, A. R., & Carskadon, M. A. (1998). Sleep schedules and daytime functioning in adolescents. *Child Development, 69*(4), 875-887.

Woodward, L. J., & Fergusson, D. M. (2000). Childhood peer relationship problems and later risks of educational under-achievement and unemployment. *Journal of Child Psychology and Psychiatry, 41*(2), 191-201.

Wössmann, L. (2003). Schooling Resources, Educational Institutions, and Student Performance: The International Evidence. *Oxford Bulletin of Economics and Statistics, 65*(2), 117-170.

Yakubovich, V. (2005). Weak Ties, Information, and Influence: How Workers Find Jobs in a Local Russian Labor Market. *American Sociological Review, 70*(3), 408-421.

Yang, C.-M., & Spielman, A. J. (2001). The effect of a delayed weekend sleep pattern on sleep and morning functioning. *Psychology and Health, 16*(6), 715-725.

Yao, Y., Dresner, M., & Zhu, K. (2010). *Improving Order Fulfillment by Beating the "Monday Blues"*. Working paper.

Youniss, J., & Haynie, D. L. (1992). Friendship in adolescence. *Journal of developmental and behavioral pediatrics, 13*(1), 59-66.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: evidence from a natural experiment. *Review of Economics and Statistics, 85*(1), 9-23.

Zwarenstein, M., Oxman, A. D., & PRACTIHC (2006). Why are so few randomized trials useful, and what can we do about it? *Journal of Clinical Epidemiology, 59*(11), 1125-1126.

Appendix to Chapter 2

Figure A2.1: Distribution of academic attainment at two-year follow-up by cohort

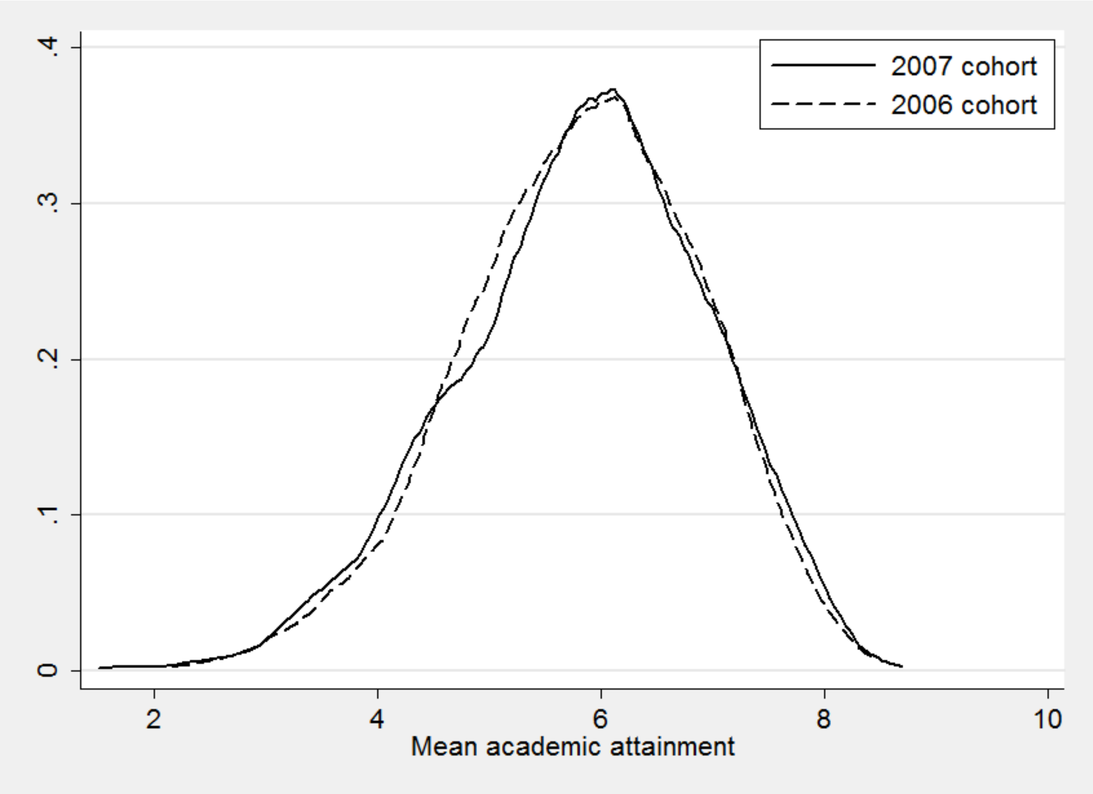


Table A2.1: Raw outcomes at baseline and each follow-up point

	Baseline		Post		1-year follow-up		2-year follow-up	
	T	C	T	C	T	C	T	C
Depression score	8.58	8.07	8.62	8.35	8.14	7.27	8.35	7.34
SD	6.95	6.24	7.32	7.23	7.12	6.68	7.60	6.98
N	1,767	1,833	1,728	1,779	1,624	1,623	1,547	1,505
Anxiety score	9.32	9.00	8.57	7.95	7.77	6.93	7.51	6.60
SD	6.81	6.35	6.77	6.53	6.66	6.33	6.81	6.14
N	1,764	1,835	1,711	1,775	1,607	1,614	1,530	1,485
Absence rate	0.06	0.06	0.07	0.07	0.08	0.09	0.08	0.09
SD	0.06	0.07	0.08	0.08	0.09	0.10	0.11	0.11
N	1,227	1,219	1,227	1,218	1,221	1,207	1,207	1,191
Self-reported behaviour score	11.09	10.93	11.41	10.96	10.86	10.31	10.58	9.79
SD	6.38	6.18	6.37	6.46	6.18	6.21	6.30	5.91
N	1,766	1,828	1,717	1,762	1,624	1,616	1,533	1,470
Self-reported prosocial score	7.47	7.32	7.26	7.07	6.99	6.89	6.79	6.76
SD	1.91	2.05	1.94	2.08	1.94	2.02	2.03	2.06
N	1,766	1,828	1,717	1,761	1,624	1,616	1,533	1,470
Teacher-reported behaviour score	6.00	6.55	6.72	6.86	7.10	6.98	7.08	6.71
SD	5.96	6.15	6.61	6.30	6.56	6.44	6.38	6.31
N	1,740	1,773	1,719	1,704	1,581	1,529	1,515	1,305
Teacher-reported prosocial score	7.17	7.11	7.09	7.10	6.85	6.78	6.85	6.83
SD	2.38	2.58	2.59	2.59	2.60	2.62	2.71	2.67
N	1,738	1,773	1,719	1,704	1,580	1,529	1,513	1,305
Combined academic attainment	4.54	4.45	5.08	4.93	5.41	5.22	5.91	5.74
SD	0.70	0.74	0.87	0.87	0.89	0.93	1.09	1.09
N	1,594	4,423	1,531	4,084	1,538	4,172	1,576	4,263

Notes: This table presents the raw scores of the outcome variables at each time point. Included are all pupils for whom I have a baseline and at least one follow-up measure for the relevant outcome variable. I have not excluded pupils without demographic data, so the sample sizes may be larger than shown in the regression tables. T=treatment group; C=control group.

Table A2.2: Programme impact on self- and teacher-reported Goodman SDQ prosocial scores

	Outcome: Self-reported prosocial score				Outcome: Teacher-reported prosocial score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated*Postintervention	0.03	-0.004	-0.038	-0.046	-0.039	-0.050	0.123	0.046
Standard error	-0.04	-0.041	-0.049	-0.053	(0.079)	(0.076)	(0.106)	(0.102)
Treated*1-year follow-up	-0.02	-0.042	-0.069	-0.101	0.041	0.020	0.152	0.071
Standard error	-0.042	-0.043	-0.057	-0.063	(0.077)	(0.077)	(0.106)	(0.104)
Treated*2-year follow-up	-0.059	-0.073	-0.141***	-0.118*	-0.015	-0.042	0.003	-0.085
Standard error	-0.045	-0.046	-0.054	-0.064	(0.098)	(0.096)	(0.123)	(0.129)
Treated	0.067	0.079**	0.122**		0.020	-0.021	-0.035	
Standard error	-0.042	-0.04	-0.053		(0.085)	(0.082)	(0.093)	
Number of Pupils	3,457	3,457	3,457	3,457	3,376	3,376	3,376	3,376
Sample size	12,836	12,836	12,836	12,836	12,382	12,382	12,382	12,382
Number of clusters	147	147	147	147	144	144	144	144
Month & day of the week	x	x	x	x	x	x	x	x
Pupil characteristics		x	x			x	x	
School, School*Time			x	x			x	x
Pupil Fixed Effects				x				x

Notes: Each column represents a separate regression. Standard errors (in parentheses) are clustered at the level of treatment assignment (class). All specifications include controls for the time period (postintervention, 1-year or 2-year follow-up). Control variables for student characteristics are dummies for gender; SEN; FSM; Key Stage 2 maths and English attainment levels; and ethnicity (5 categories). The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1 based on the pooled (intervention and control group) distribution at baseline.

Table A2.3: Programme impact on depression and anxiety symptom scores and absence, using sample with all outcome variables

	Outcome: Depression symptoms score				Outcome: Anxiety symptoms score				Outcome: Absence from school			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treated*Postintervention	-0.091	-0.108	-0.162**	-0.146**	0.013	-0.015	-0.083	-0.074	-0.102**	-0.102**	-0.134**	-0.134*
Standard error	(0.066)	(0.067)	(0.078)	(0.072)	(0.060)	(0.055)	(0.062)	(0.060)	(0.049)	(0.049)	(0.061)	(0.070)
Treated*1-year follow-up	-0.062	-0.059	-0.003	0.008	-0.050	-0.043	-0.004	0.003	-0.074	-0.074	-0.032	-0.032
Standard error	(0.065)	(0.065)	(0.084)	(0.084)	(0.055)	(0.056)	(0.066)	(0.074)	(0.061)	(0.061)	(0.085)	(0.098)
Treated*2-year follow-up	0.041	0.016	0.064	0.054	0.037	0.014	-0.011	-0.031	-0.105	-0.104	-0.130*	-0.127
Standard error	(0.079)	(0.082)	(0.098)	(0.101)	(0.062)	(0.063)	(0.080)	(0.093)	(0.063)	(0.064)	(0.076)	(0.088)
Treated	0.129*	0.169***	0.062		0.094	0.130**	0.041		-0.033	0.011	0.031	
Standard error	(0.072)	(0.057)	(0.064)		(0.061)	(0.053)	(0.061)		(0.049)	(0.044)	(0.042)	
Number of Pupils	1,879	1,879	1,879	1,879	1,859	1,859	1,859	1,859	1,879	1,879	1,879	1,879
Sample size	7,068	7,068	7,068	7,068	6,950	6,950	6,950	6,950	7,511	7,511	7,511	7,511
Number of clusters	86	86	86	86	86	86	86	86	91	91	91	91
Month & day of the week	x	x	x	x	x	x	x	x				
Pupil characteristics		x	x			x	x			x	x	
School, School*Time			x	x			x	x			x	x
Pupil Fixed Effects				x				x				x

Notes: Each column represents a separate regression. Standard errors (in parentheses) are clustered at the level of treatment assignment (class). All specifications include controls for the time period (postintervention, 1-year or 2-year follow-up). Control variables for student characteristics are dummies for gender; SEN; FSM; Key Stage 2 maths and English attainment levels; and ethnicity (5 categories). The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1.

Table A2.4: Programme impact on behaviour scores and academic attainment, using sample with all outcome variables

	Outcome: Self-reported behaviour score				Outcome: Teacher-reported behaviour score				Outcome: Academic attainment			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treated*Postintervention	0.093	0.072	0.034	0.038	0.064	0.016	-0.142	-0.044	0.221***	0.219***	0.055	0.030
Standard error	(0.061)	(0.056)	(0.059)	(0.064)	(0.098)	(0.084)	(0.086)	(0.080)	(0.074)	(0.073)	(0.066)	(0.075)
Treated*1-year follow-up	0.023	0.016	0.058	0.079	0.071	0.041	-0.005	0.099	0.204***	0.187***	0.066	0.049
Standard error	(0.055)	(0.056)	(0.071)	(0.072)	(0.087)	(0.076)	(0.105)	(0.097)	(0.064)	(0.063)	(0.054)	(0.067)
Treated*2-year follow-up	0.141**	0.104	0.096	0.046	0.074	0.055	-0.041	0.050	0.154*	0.144	0.148*	0.155
Standard error	(0.068)	(0.065)	(0.064)	(0.071)	(0.122)	(0.114)	(0.139)	(0.170)	(0.093)	(0.092)	(0.078)	(0.096)
Treated	-0.024	0.023	-0.052		-0.113	0.006	0.043		0.237*	0.075	0.186*	
Standard error	(0.070)	(0.052)	(0.058)		(0.106)	(0.075)	(0.104)		(0.130)	(0.086)	(0.102)	
Number of Pupils	1,865	1,865	1,865	1,865	1,879	1,879	1,879	1,879	1,879	1,879	1,879	1,879
Sample size	6,969	6,969	6,969	6,969	7,055	7,055	7,055	7,055	7,372	7,372	7,372	7,372
Number of clusters	86	86	86	86	86	86	86	86	331	331	331	331
Month & day of the week	x	x	x	x	x	x	x	x				
Pupil characteristics		x	x			x	x			x	x	
School, School*Time			x	x			x	x			x	x
Pupil Fixed Effects				x				x				x

Notes: Each column represents a separate regression. Standard errors (in parentheses) are clustered at the level of treatment assignment (class). All specifications include controls for the time period (postintervention, 1-year or 2-year follow-up). Control variables for student characteristics are dummies for gender; SEN; FSM; Key Stage 2 maths and English attainment levels; and ethnicity (5 categories), except when academic attainment is the outcome when Key Stage 2 attainment is omitted. The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1.

Table A2.5: Programme impact on depression and anxiety symptom scores and absence, using sample of schools with within-year control group

	Outcome: Depression symptoms score				Outcome: Anxiety symptoms score				Outcome: Absence from school			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treated*Postintervention	-0.131**	-0.139**	-0.176***	-0.140**	-0.033	-0.040	-0.063	-0.055	-0.081	-0.081	-0.108*	-0.107
Standard error	(0.054)	(0.056)	(0.058)	(0.063)	(0.048)	(0.049)	(0.048)	(0.052)	(0.060)	(0.060)	(0.059)	(0.068)
Treated*1-year follow-up	-0.057	-0.040	-0.039	-0.008	-0.014	0.001	-0.018	0.011	-0.003	-0.003	0.024	0.021
Standard error	(0.058)	(0.058)	(0.058)	(0.063)	(0.044)	(0.044)	(0.042)	(0.051)	(0.074)	(0.074)	(0.073)	(0.085)
Treated*2-year follow-up	0.002	0.026	0.027	0.041	-0.026	-0.011	-0.034	-0.017	0.020	0.019	-0.007	-0.017
Standard error	(0.073)	(0.073)	(0.068)	(0.075)	(0.058)	(0.056)	(0.054)	(0.062)	(0.093)	(0.094)	(0.086)	(0.098)
Treated	0.078	0.108**	0.104**		0.067	0.083*	0.069		-0.060	-0.019	0.024	
Standard error	(0.061)	(0.050)	(0.047)		(0.055)	(0.049)	(0.044)		(0.061)	(0.048)	(0.042)	
Number of Pupils	2,562	2,562	2,562	2,562	2,566	2,566	2,566	2,566	1,895	1,895	1,895	1,895
Sample size	9,483	9,483	9,483	9,483	9,441	9,441	9,441	9,441	7,519	7,519	7,519	7,519
Number of clusters	109	109	109	109	109	109	109	109	76	76	76	76
Month & day of the week	x	x	x	x	x	x	x	x				
Pupil characteristics		x	x			x	x			x	x	
School, School*Time			x	x			x	x			x	x
Pupil Fixed Effects				x				x				x

Notes: Each column represents a separate regression. Standard errors (in parentheses) are clustered at the level of treatment assignment (class). All specifications include controls for the time period (postintervention, 1-year or 2-year follow-up). Control variables for student characteristics are dummies for gender; SEN; FSM; Key Stage 2 maths and English attainment levels; and ethnicity (5 categories). The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1.

Table A2.6: Programme impact on behaviour scores and academic attainment, using sample of schools with within-year control group

	Outcome: Self-reported behaviour score				Outcome: Teacher-reported behaviour score				Outcome: Academic attainment			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treated*Postintervention	0.032	0.034	0.011	0.023	0.034	0.003	-0.072	-0.003	0.086	0.075	0.073	0.062
Standard error	(0.050)	(0.054)	(0.052)	(0.056)	(0.079)	(0.078)	(0.080)	(0.076)	(0.091)	(0.089)	(0.048)	(0.055)
Treated*1-year follow-up	-0.004	0.011	-0.001	0.025	0.043	0.036	0.045	0.150*	0.081	0.064	0.066	0.056
Standard error	(0.051)	(0.051)	(0.052)	(0.058)	(0.087)	(0.087)	(0.093)	(0.087)	(0.073)	(0.071)	(0.040)	(0.048)
Treated*2-year follow-up	0.048	0.070	0.041	0.033	0.062	0.063	0.056	0.135	0.095	0.090	0.132**	0.131**
Standard error	(0.059)	(0.058)	(0.050)	(0.056)	(0.102)	(0.105)	(0.109)	(0.125)	(0.084)	(0.084)	(0.054)	(0.065)
Treated	-0.006	0.030	0.016		-0.051	0.079	0.042		0.153	0.125**	0.187***	
Standard error	(0.067)	(0.055)	(0.048)		(0.102)	(0.078)	(0.075)		(0.097)	(0.057)	(0.060)	
Number of Pupils	2,562	2,562	2,562	2,562	2,440	2,440	2,440	2,440	3,459	3,459	3,459	3,459
Sample size	9,413	9,413	9,413	9,413	8,832	8,832	8,832	8,832	13,499	13,499	13,499	13,499
Number of clusters	109	109	109	109	107	107	107	107	466	466	466	466
Month & day of the week	x	x	x	x	x	x	x	x				
Pupil characteristics		x	x			x	x			x	x	
School, School*Time			x	x			x	x			x	x
Pupil Fixed Effects				x				x				x

Notes: Each column represents a separate regression. Standard errors (in parentheses) are clustered at the level of treatment assignment (class). All specifications include controls for the time period (postintervention, 1-year or 2-year follow-up). Control variables for student characteristics are dummies for gender; SEN; FSM; Key Stage 2 maths and English attainment levels; and ethnicity (5 categories), except when academic attainment is the outcome when Key Stage 2 attainment is omitted. The outcome measure in each case is standardised to have a mean of 0 and a standard deviation of 1.

Table A2.7: Mediation of programme impact by depression symptoms scores

	Absence from school			Academic attainment		
	Step 1	Step 2	Step 3	Step 1	Step 2	Step 3
Treated*Postintervention	-0.051	-0.160**	-0.046	0.023	-0.147**	0.020
Standard error	(0.064)	(0.066)	(0.065)	(0.051)	(0.058)	(0.051)
Treated*1-year follow-up	0.026	-0.017	0.027	0.026	-0.002	0.025
Standard error	(0.076)	(0.072)	(0.077)	(0.049)	(0.064)	(0.049)
Treated*2-year follow-up	-0.032	0.035	-0.033	0.114	0.016	0.114
Standard error	(0.072)	(0.090)	(0.072)	(0.075)	(0.073)	(0.075)
Depression score			0.033*			-0.014
Standard error			(0.018)			(0.009)
Number of Pupils	2,216	2,216	2,216	2,999	2,999	2,999
Sample size	8,175	8,175	8,175	11,220	11,220	11,220
Number of clusters	88	88	88	136	136	136
Pupil characteristics						
School, School*Time	x	x	x	x	x	x
Pupil Fixed Effects	x	x	x	x	x	x

Notes: Table shows estimates of mediation of the dependent variable (absence or academic attainment) through the depression symptoms score, following Baron and Kenny (1986). Step 1 gives the regression results for the dependent variable on 'treated*time' for this sample; Step 2 uses the same specification but with the mediator (depression score) as the dependent variable; and Step 3 runs the same specification as Step 1 but adds in the mediator (depression score) on the right hand side.

Table A2.8: Treated pupils by workshop characteristics

		N in group 16 or fewer	N in group > 16	Total
At least 17 hours of workshops	N pupils	1,288	37	1,325
	N observations	4,864	141	5,005
	N schools	20	3	20
Less than 17 hours of workshops	N pupils	349	93	442
	N observations	1,314	347	1,661
	N schools	8	2	9
Total	N pupils	1,637	130	1,767
	N observations	6,178	488	6,666
	N schools	22	4	22

Notes: Table shows number of treated pupils and observations by workshop characteristics, based on the sample of pupils with data on symptoms of depression. In regressions I take high intensity workshops to be those with at least 17 hours of workshops in groups of 16 or fewer (top left cell of table), and all other workshops to be low intensity. This puts 27% of treated pupils in low intensity workshops, and 27% of treated pupil observations in these workshops. If thresholds of at least 18 hours and 15 or fewer pupils were used, 60% of pupils would be in low intensity workshops. Most schools (20 of 22) ran at least one high intensity workshop group; half (11 schools) ran at least one low intensity workshop group. Note that the measures only take into account the time scheduled and the number of pupils assigned to each class by the school, not whether pupils were actually present in lessons or how much of the course they actually received. The figures in this table may differ from those presented in a similar table in Chapter 3 because of the difference in pupils for whom I have data on each outcome measure.

Table A2.9: Robustness checks: impact on academic attainment

	Outcome: Academic attainment				Outcome: Academic attainment, excluding 4 schools			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated*Postintervention	0.093	0.094	0.087**	0.092**	0.150**	0.148**	0.107***	0.120***
Standard error	(0.070)	(0.070)	(0.034)	(0.040)	(0.068)	(0.068)	(0.039)	(0.045)
Treated*1-year follow-up	0.069	0.068	0.065**	0.067*	0.119**	0.110*	0.078**	0.084*
Standard error	(0.052)	(0.052)	(0.031)	(0.036)	(0.061)	(0.060)	(0.039)	(0.046)
Treated*2-year follow-up	0.090*	0.089*	0.092***	0.091**	0.116	0.112	0.107**	0.109*
Standard error	(0.050)	(0.050)	(0.033)	(0.038)	(0.071)	(0.070)	(0.048)	(0.057)
Treated	-0.023***	-0.022**	-0.017**		-0.017	-0.009	0.046	
Standard error	(0.008)	(0.010)	(0.008)		(0.092)	(0.054)	(0.052)	
Number of Pupils	5,192	5,192	5,192	5,192	4,179	4,179	4,179	4,179
Sample size	20,270	20,270	20,270	20,270	16,272	16,272	16,272	16,272
Number of clusters	506	506	506	506	470	470	470	470
Month & day of the week								
Pupil characteristics		x	x			x	x	
School, School*Time			x	x			x	x
Pupil Fixed Effects				x				x
Baseline academic level*Time	x	x	x	x				

Notes: Table displays results for the same basic specifications as for academic attainment in Table 2.5. However, columns 1-4 include dummies for Key Stage 2 level and interactions of these with time, in order to control for differential trends of pupils with different baseline attainment. Columns 5-8 exclude data from the four schools which assigned higher ability sets to intervention (see Chapter 1), and which therefore have treatment and control groups which are not matched on academic attainment at baseline.

Table A2.10: Impact on academic attainment by baseline attainment (Key Stage 2 levels)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treated*Postintervention	0.001	-0.002	-0.019	0.021	0.045	0.044	0.029	0.028	0.226**	0.229**	0.186***	0.188***
Standard error	(0.089)	(0.087)	(0.054)	(0.057)	(0.065)	(0.065)	(0.035)	(0.041)	(0.105)	(0.106)	(0.048)	(0.056)
Treated*1-year follow-up	0.052	0.044	0.001	0.025	0.046	0.045	0.052	0.046	0.115	0.116	0.091*	0.094*
Standard error	(0.060)	(0.058)	(0.050)	(0.051)	(0.056)	(0.056)	(0.037)	(0.043)	(0.081)	(0.081)	(0.048)	(0.056)
Treated*2-year follow-up	0.085	0.082	0.089*	0.084	0.062	0.061	0.076*	0.076	0.133	0.133	0.109**	0.112*
Standard error	(0.063)	(0.063)	(0.048)	(0.055)	(0.064)	(0.064)	(0.044)	(0.051)	(0.081)	(0.081)	(0.050)	(0.057)
Treated	-0.005	0.010	0.028		-0.003	-0.002	0.004		-0.017	-0.016	-0.008	
Standard error	(0.075)	(0.062)	(0.064)		(0.019)	(0.020)	(0.018)		(0.018)	(0.017)	(0.016)	
Number of Pupils	1,581	1,581	1,581	1,581	2,066	2,066	2,066	2,066	1,545	1,545	1,545	1,545
Sample size	6,099	6,099	6,099	6,099	8,065	8,065	8,065	8,065	6,106	6,106	6,106	6,106
Number of clusters	265	265	265	265	271	271	271	271	186	186	186	186
Month & day of the week												
Pupil characteristics		x	x			x	x			x	x	
School, School*Time			x	x			x	x			x	x
Pupil Fixed Effects				x				x				x

Notes: Table displays results for the same basic specifications as for academic attainment in Table 2.5. However, the sample in columns 1-4 consists of pupils who obtained a sum of Key Stage 2 levels of 11 or less, for example: a pupil who obtained level 3 in each of English, maths and science will have a level sum of 9. Columns 5-8 include pupils with level sums of 12 or 13, and columns 9-12 include pupils with level sums of 14 or 15.

Appendix to Chapter 3

Table A3.1: Questionnaire page asking about students' friends

This shows the wording and response space for the question on pupils' friendships. It also gives an example of a respondent filling in the names of friends which were difficult to positively identify. Original paper size was A4.

Section 5

We'd like to know who you are good friends with.

Please write down the first name, surname and form group (if you know it!) of everyone who is a good friend of yours. If they're not at the same school as you, please write down their school as well.

My good friends are:

First name and surname	Form or class group (if you know it!)	School (if they're not at your school)
aves ?		
Loss ?		
Jown ?		
Lock ?		
cresc ?		
harry ?		

12

Table A3.2: Questionnaire response by timing and cohorts

		Year 7	Year 8	Year 9	Total
	Treated	1,769	0	0	1,769
June 2008	Control	1,702	2,012	0	3,714
	All	**3,471**	**2,012**	**0**	**5,483**
	Treated	0	1,677	0	1,677
June 2009	Control	0	1,750	1,933	3,683
	All	**0**	**3,427**	**1,933**	**5,360**
	Treated	0	0	1,501	1,501
June 2010	Control	0	0	1,501	1,501
	All	**0**	**0**	**3,002**	**3,002**
	Treated	1,769	1,677	1,501	4,947
Total	Control	1,702	3,762	3,434	8,898
	All	**3,471**	**5,439**	**4,935**	**13,845**

Notes: Table shows the number of pupils in each cohort and yeargroup at each point in time who completed the friends section of the questionnaire. Note that this is not the effective sample size, which is shown in Table 3.1: pupils can be included in regressions even if they did not respond to the friends question as the in-degree of friendship relies on data obtained from other pupils and the out-degree of friendship is zero when pupils do not respond to the questionnaire. This table merely gives information on questionnaire response at each time period; this will be very similar to the rate of questionnaire response for the depression symptoms score as presented in Chapter 1.

Table A3.3: Treated pupils by workshop characteristics

		N in group 16 or fewer	N in group > 16	Total
At least 17 hours of workshops	N pupils	1,403	38	1,441
	N observations	4,114	112	4,226
	N schools	20	3	20
Less than 17 hours of workshops	N pupils	378	117	495
	N observations	1,109	351	1,460
	N schools	8	4	10
Total	N pupils	1,781	155	1,936
	N observations	5,223	463	5,686
	N schools	22	6	22

Notes: Table shows number of treated pupils and observations by workshop characteristics. In regressions I take high intensity workshops to be those with at least 17 hours of workshops in groups of 16 or fewer (top left cell of table), and all other workshops to be low intensity. This puts 28% of treated pupils in low intensity workshops, and 28% of treated pupil observations in these workshops. If thresholds of at least 18 hours and 15 or fewer pupils were used, 60% of pupils would be in low intensity workshops. Most schools (20 of 22) ran at least one high intensity workshop group; half (11 schools) ran at least one low intensity workshop group. Note that the measures only take into account the time scheduled and the number of pupils assigned to each class by the school, not whether pupils were actually present in lessons or how much of the course they actually received. The definition used here is the same as that used to define high/low quality workshops in Chapter 2, however, because the sample size is different for each outcome the number of observations in each category will differ (and indeed, will vary for each outcome measure used in Chapter 2).

Table A3.4: Probability of outcome variable being zero

Outcome	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	In-degree friends			**Out-degree friends**			**In-degree friends 1-10**			**In-degree friends 11+**		
	Dummy=1 if friend listing=0			Dummy=1 if friends listed=0			Dummy=1 if friend listing 1-10=0			Dummy=1 if friend listing 11+=0		
Year 7*Treated	-0.401	-0.246	-0.167	-0.477**	-0.463**	-0.139	-0.419**	-0.320	-0.113	-0.380***	-0.329**	-0.267*
SE	(0.266)	(0.309)	(0.327)	(0.195)	(0.204)	(0.195)	(0.209)	(0.228)	(0.224)	(0.139)	(0.145)	(0.140)
Odds ratio	0.670	0.782	0.846	0.621	0.630	0.870	0.658	0.726	0.893	0.684	0.720	0.766
Year 8*Treated	-0.192	0.124	0.074	-0.078	-0.015	0.067	-0.113	0.109	0.143	-0.220	-0.095	-0.149
SE	(0.233)	(0.239)	(0.236)	(0.130)	(0.133)	(0.127)	(0.185)	(0.187)	(0.175)	(0.138)	(0.139)	(0.131)
Odds ratio	0.826	1.132	1.077	0.925	0.985	1.069	0.893	1.115	1.153	0.802	0.910	0.862
Year 9*Treated	-0.458*	-0.119	-0.178	-0.160	-0.094	-0.024	-0.248	-0.003	0.031	-0.059	0.079	0.035
SE	(0.259)	(0.263)	(0.243)	(0.146)	(0.146)	(0.117)	(0.202)	(0.215)	(0.195)	(0.139)	(0.134)	(0.102)
Odds ratio	0.633	0.888	0.837	0.852	0.910	0.976	0.780	0.997	1.031	0.942	1.082	1.035
N pupils	6,510	6,079	6,079	6,510	6,079	6,079	6,510	6,079	6,079	6,510	6,079	6,079
N	16,317	15,537	15,537	16,317	15,537	15,537	16,317	15,537	15,537	16,317	15,537	15,537
Clusters	178	169	169	178	169	169	178	169	169	178	169	169
Demographic controls	no	no	yes	no	no	yes	no	no	yes	no	no	yes

Notes: Table shows results of logistic regressions on dummies for year group (timing) and treatment status*year group. The outcome variable in each case is a dummy which is equal to 1 if the original variable (given by column headings) is equal to zero. The specifications follow the same pattern as in Table 3.4: the first column for each outcome includes all observations; the second uses the same raw specification as the first but includes only observations which have full demographic data; and the third includes demographic and school controls. Standard errors are heteroskedasticity robust and are clustered by class group, which is the unit of treatment assignment.

Table A3.5: Programme impact by year when outcome is greater than zero

Outcome	**In-degree friends**			**Out-degree friends**			**In-degree friends 1-10**			**In-degree friends 11+**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Year 7*Treated	0.115***	0.105***	0.048*	0.015	0.006	0.013	0.088***	0.079***	0.012	0.106**	0.103**	0.078*
SE	(0.036)	(0.036)	(0.029)	(0.031)	(0.031)	(0.033)	(0.028)	(0.027)	(0.023)	(0.047)	(0.047)	(0.040)
IRR	1.122	1.111	1.049	1.015	1.006	1.013	1.092	1.082	1.013	1.112	1.109	1.081
Year 8*Treated	0.052**	0.034	0.028	0.024	0.016	0.018	0.046**	0.030	0.016	0.039	0.031	0.040
SE	(0.025)	(0.025)	(0.024)	(0.022)	(0.022)	(0.021)	(0.022)	(0.021)	(0.020)	(0.033)	(0.034)	(0.031)
IRR	1.053	1.035	1.028	1.024	1.016	1.019	1.047	1.031	1.016	1.040	1.032	1.041
Year 9*Treated	0.002	-0.017	-0.025	0.012	0.003	0.007	0.018	0.001	-0.012	-0.025	-0.034	-0.030
SE	(0.038)	(0.037)	(0.028)	(0.026)	(0.026)	(0.025)	(0.035)	(0.034)	(0.024)	(0.042)	(0.043)	(0.039)
IRR	1.002	0.983	0.975	1.012	1.003	1.007	1.018	1.001	0.988	0.975	0.966	0.970
N pupils	6,410	6,026	6,026	6,183	5,842	5,842	6,371	5,999	5,999	6,152	5,844	5,844
N	15,974	15,284	15,284	13,845	13,268	13,268	15,720	15,051	15,051	14,132	13,637	13,637
Clusters	178	169	169	177	166	166	178	168	168	176	167	167
Demographic controls	no	no	yes	no	no	yes	no	no	yes	no	no	yes

Notes: Table shows results of negative binomial regressions of the four outcome variables (given by column headings) on dummies for year group (timing) and treatment status*year group. The regressions are the same as those reported in Table 3.4, but exclude observations where the outcome variable is equal to zero. The first column for each outcome includes all observations; the second uses the same raw specification as the first but includes only observations which have full demographic data; and the third includes demographic and school controls. Standard errors are heteroskedasticity robust and are clustered by class group, which is the unit of treatment assignment.

Table A3.6: Programme impact on friends by year group using ordinary least squares regressions

Standardised outcome	In-degree friends			Out-degree friends			In-degree friends 1-10			In-degree friends 11+		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Year 7*Treated	0.195***	0.176***	0.080*	0.099	0.083	0.010	0.170***	0.148***	0.033	0.180***	0.169**	0.119**
SE	(0.060)	(0.060)	(0.045)	(0.060)	(0.061)	(0.056)	(0.052)	(0.052)	(0.039)	(0.064)	(0.065)	(0.054)
Year 8*Treated	0.089**	0.054	0.047	0.050	0.028	0.016	0.082**	0.046	0.024	0.077	0.050	0.064
SE	(0.044)	(0.043)	(0.038)	(0.041)	(0.042)	(0.040)	(0.041)	(0.040)	(0.035)	(0.048)	(0.049)	(0.044)
Year 9*Treated	0.016	-0.022	-0.029	0.058	0.029	0.016	0.040	0.001	-0.022	-0.016	-0.044	-0.031
SE	(0.056)	(0.055)	(0.039)	(0.053)	(0.053)	(0.047)	(0.058)	(0.057)	(0.037)	(0.054)	(0.054)	(0.044)
N pupils	6,510	6,079	6,079	6,510	6,079	6,079	6,510	6,079	6,079	6,510	6,079	6,079
N	16,317	15,537	15,537	16,317	15,537	15,537	16,317	15,537	15,537	16,317	15,537	15,537
Clusters	178	169	169	178	169	169	178	169	169	178	169	169
Demographic controls	no	no	yes	no	no	yes	no	no	yes	no	no	yes

Notes: Table shows results of ordinary least squares regressions of the four outcome variables (given by column headings) each standardised to have a mean of zero and a standard deviation of 1, using the mean and standard deviation of the control group in Year 7. This allows interpretation of the coefficients in terms of effect sizes. The outcomes are regressed on dummies for year group (timing) and treatment status*year group. The specifications are the same as those used in Table 3.4, see notes to that table for details.

Table A3.7: Programme impact on friends by year group using zero-inflated negative binomial regressions

Outcome	In-degree friends		Out-degree friends		In-degree friends 1-10		In-degree friends 11+	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Year 7*Treated	0.106***	0.050*	0.006	0.013	0.082***	0.014	0.130**	0.112**
SE	(0.037)	(0.029)	(0.031)	(0.033)	(0.029)	(0.024)	(0.054)	(0.047)
IRR	1.112	1.052	1.006	1.014	1.085	1.014	1.139	1.119
Year 8*Treated	0.035	0.028	0.016	0.018	0.030	0.015	0.039	0.055
SE	(0.025)	(0.024)	(0.022)	(0.021)	(0.022)	(0.021)	(0.038)	(0.036)
IRR	1.035	1.029	1.016	1.018	1.030	1.015	1.040	1.056
Year 9*Treated	-0.016	-0.024	0.003	0.007	-0.001	-0.013	-0.051	-0.037
SE	(0.038)	(0.028)	(0.026)	(0.026)	(0.036)	(0.025)	(0.052)	(0.045)
IRR	0.984	0.976	1.003	1.007	0.999	0.987	0.951	0.964
N pupils	6,079	6,079	6,079	6,079	6,079	6,079	6,079	6,079
N	15,537	15,537	15,537	15,537	15,537	15,537	15,537	15,537
Clusters	169	169	169	169	169	169	169	169
Demographic controls	no	yes	no	yes	no	yes	no	yes

Notes: Table shows results of zero-inflated negative binomial regressions of the four outcome variables (given by column headings) on dummies for year group (timing) and treatment status*year group, using gender, SEN, FSM and Key Stage 2 score as the inflate variables. The specifications used are the same as the second and third columns for each outcome variable in Table 3.4.

Table A3.8: Programme impact on friends excluding pupils with many friends out of school

Outcome	In-degree friends			Out-degree friends			In-degree friends 1-10			In-degree friends 11+		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Year 7*Treated	0.123***	0.111***	0.050*	0.071*	0.060	0.018	0.104***	0.091***	0.015	0.157***	0.145**	0.113**
SE	(0.039)	(0.039)	(0.030)	(0.043)	(0.043)	(0.042)	(0.033)	(0.032)	(0.025)	(0.059)	(0.058)	(0.049)
IRR	1.131	1.118	1.051	1.074	1.061	1.018	1.109	1.095	1.015	1.170	1.156	1.120
Year 8*Treated	0.048*	0.027	0.021	0.032	0.017	0.012	0.045*	0.024	0.008	0.053	0.033	0.046
SE	(0.027)	(0.026)	(0.025)	(0.028)	(0.028)	(0.030)	(0.025)	(0.024)	(0.023)	(0.040)	(0.040)	(0.038)
IRR	1.049	1.028	1.021	1.033	1.017	1.012	1.046	1.024	1.008	1.054	1.033	1.047
Year 9*Treated	0.012	-0.015	-0.022	0.045	0.020	0.021	0.028	0.002	-0.012	-0.016	-0.045	-0.038
SE	(0.041)	(0.039)	(0.029)	(0.042)	(0.042)	(0.038)	(0.040)	(0.038)	(0.025)	(0.057)	(0.056)	(0.047)
IRR	1.013	0.985	0.978	1.046	1.020	1.021	1.029	1.002	0.988	0.984	0.956	0.963
N pupils	6,454	6,041	6,041	6,454	6,041	6,041	6,454	6,041	6,041	6,454	6,041	6,041
N	15,622	14,899	14,899	15,622	14,899	14,899	15,622	14,899	14,899	15,622	14,899	14,899
Clusters	178	169	169	178	169	169	178	169	169	178	169	169
Demographic controls	no	no	yes	no	no	yes	no	no	yes	no	no	yes

Notes: Table shows results of negative binomial regressions of the four outcome variables (given by column headings) on dummies for year group (timing) and treatment status*year group. The specifications are the same as those used in Table 3.4, see notes to that table for details of control variables used. The sample used excludes the 608 pupils (695 observations) whose out-degree friends were primarily out of school, specifically, if a pupil listed at least some friends and more than half of them were not at the same school, they are excluded from this sample.

Appendix to Chapter 4

Table A4.1: Mean absence rates by pupil characteristics

	Mean	Standard deviation	Min	Max	N
Overall absence rate	0.085	0.107	0	0.993	2,880
Boy	0.082	0.105	0	0.993	1,461
Girl	0.089	0.108	0	0.873	1,419
FSM	0.123	0.134	0	0.944	841
Not FSM	0.070	0.089	0	0.993	2,038
SEN	0.123	0.147	0	0.993	591
Not SEN	0.076	0.091	0	0.944	2,287
White ethnicity	0.089	0.110	0	0.993	2,487
Non-white ethnicity	0.065	0.079	0	0.725	351
IDACI score above median	0.097	0.121	0	0.993	1,353
IDACI score at or below median	0.075	0.091	0	0.985	1,500
Autumn born	0.084	0.092	0	0.806	710
Winter born	0.093	0.126	0	0.993	677
Spring born	0.080	0.094	0	0.825	735
Summer born	0.085	0.113	0	0.985	758
Mean KS2 score above median	0.065	0.075	0	0.932	1,335
Mean KS2 score at or below median	0.101	0.119	0	0.993	1,369
EBD SEN	0.135	0.138	0	0.748	86
Not EBD SEN	0.085	0.106	0	0.993	2,718
Depression score above median	0.098	0.107	0	0.993	648
Depression score at or below median	0.073	0.087	0	0.944	655
Anxiety score above median	0.096	0.109	0	0.993	647
Anxiety score at or below median	0.076	0.084	0	0.944	653
Self-reported behaviour score above median	0.100	0.110	0	0.993	637
Self-reported behaviour score at or below median	0.072	0.083	0	0.944	667
Teacher-reported behaviour score above median	0.116	0.128	0	0.993	646
Teacher-reported behaviour score at or below median	0.059	0.054	0	0.347	660
Does not live with either parent	0.088	0.074	0	0.347	89
Lives with one parent	0.102	0.110	0	0.993	600
Lives with both parents	0.067	0.076	0	0.625	607
Developed ability score at age 11 above median	0.061	0.062	0	0.459	581
Developed ability score at or below median	0.093	0.109	0	0.993	631

Notes: Table shows mean absence rates by pupil characteristics for the academic years for which I have behaviour data for each pupil. Absence data is not available for all pupils.

Table A4.2: Predicting the number of behaviour incidents per pupil controlling for the absence rate

	(1)	(2)	(3)	(4)	(5)
Absence rate	3.981***	3.972***	4.953***	0.919	3.831***
SE	(0.401)	(0.418)	(0.796)	(0.630)	(0.727)
IRR	53.572	53.114	141.668	2.508	46.126
Boy	0.597***	0.625***	0.822***	0.355***	0.943***
SE	(0.065)	(0.066)	(0.100)	(0.089)	(0.096)
IRR	1.818	1.868	2.276	1.427	2.568
FSM	0.386***	0.368***	0.279**	0.115	0.206*
SE	(0.075)	(0.076)	(0.124)	(0.103)	(0.116)
IRR	1.470	1.445	1.322	1.122	1.229
SEN	0.749***	0.427***	0.341***	-0.333***	
SE	(0.079)	(0.089)	(0.130)	(0.116)	
IRR	2.115	1.533	1.407	0.717	
White	0.200**	0.230**	0.583***	0.231	0.432***
SE	(0.099)	(0.108)	(0.170)	(0.144)	(0.163)
IRR	1.221	1.259	1.791	1.260	1.541
IDACI score of neighbourhood	0.522**	0.518**	0.803**	0.131	0.440
SE	(0.210)	(0.215)	(0.365)	(0.305)	(0.352)
IRR	1.686	1.679	2.231	1.140	1.552
Winter born	-0.069	-0.096	-0.320**	-0.360***	-0.384***
SE	(0.091)	(0.093)	(0.139)	(0.115)	(0.133)
IRR	0.933	0.908	0.726	0.698	0.681
Spring born	-0.238***	-0.301***	-0.266**	-0.251**	-0.238*
SE	(0.089)	(0.091)	(0.135)	(0.113)	(0.129)
IRR	0.788	0.740	0.767	0.778	0.788
Summer born	-0.226**	-0.300***	-0.381***	-0.456***	-0.490***
SE	(0.089)	(0.091)	(0.136)	(0.114)	(0.131)
IRR	0.797	0.741	0.683	0.634	0.612
Mean KS2 score		-0.402***	-0.335***	0.017	-0.161
SE		(0.058)	(0.085)	(0.106)	(0.120)
IRR		0.669	0.715	1.017	0.851
EBD SEN				0.568***	
SE				(0.187)	
IRR				1.764	
Depression score				0.039***	0.114***
SE				(0.013)	(0.013)
IRR				1.040	1.121
Anxiety score				-0.107***	-0.077***
SE				(0.013)	(0.014)
IRR				0.898	0.926
Self-reported behaviour score				0.092***	
SE				(0.017)	
IRR				1.097	
Teacher-reported behaviour score				0.150***	
SE				(0.010)	
IRR				1.161	
Does not live with parents				0.248	0.655***
SE				(0.167)	(0.193)
IRR				1.281	1.926
Lives with one parent				0.220**	0.462***
SE				(0.087)	(0.100)
IRR				1.246	1.588
IQ score at age 11				-0.010*	-0.016***
SE				(0.005)	(0.006)
IRR				0.990	0.984
N	2,834	2,677	1,041	1,041	1,041
N clusters	70	69	52	52	52

Notes: Table reports results of negative binomial regressions predicting the number of incidents per pupil, reporting coefficients, standard errors and incidence rate ratios (IRRs), with the exposure set to the number of days a pupil appears in the dataset (number of school days). Regressions include dummies for school; standard errors are clustered by class group. Each column represents a separate regression. Absence rates are for the academic years for which I have behaviour data for each pupil.