CENTRE FOR EDUCATION POLICY
& EQUALISING OPPORTUNITIES

Adolescent School Bullying Victimisation and Later Life Outcomes

Centre for Education Policy and Equalising Opportunities (CEPEO)

Emma Gorman, Colm Harmon, Silvia Mendolia, Anita Staneva and Ian Walker

Working Paper No. 20-05

February 2020

Highlights

- Bullying is widespread in schools, yet there is little evidence available on the short- and long-term consequences of different types and frequencies of bullying
- We analysed data on over 7,000 secondary school pupils in England to quantify the effects of being bullied in schools
- Experiencing bullying of any kind has negative consequences for high-stakes academic achievement at school, as well as economic outcomes and mental health in early adulthood
- While all types of bullying have negative consequences, these effects were more pronounced for persistent bullying and violent types of bullying
- Targeted policies to reduce more extreme forms of bullying may be warranted

Why does this matter?

- **Being bullied in school, especially persistent or violent bullying, has long-term negative consequences**
- **Bullying reduction policies focussing on more extreme types of bullying may be warranted**

Adolescent School Bullying Victimisation and Later Life Outcomes

Emma Gorman
University of Westminster and IZA, Bonn

Colm Harmon
University of Edinburgh and IZA, Bonn

Silvia Mendolia
University of Wollongong and IZA, Bonn

Anita Staneva
Griffith University

Ian Walker
Lancaster University Management School and IZA, Bonn

This version: January 16th, 2020

Keywords: bullying, education outcomes, long term outcomes
JEL codes: **I21, I24, I26, J24**

Abstract

We analyse the long-term effects of experiencing bullying victimisation in junior high school, using rich data on a large cohort of English adolescents. The data contain self-reports of five types of bullying and their frequency, for three waves, when the pupils were aged 13 to 16 years. We assess the effects of bullying victimisation on short- and long-term outcomes, including educational achievements, earnings, and mental ill-health at age 25 years using a variety of estimation strategies - least squares, matching, and inverse probability weighting. We also consider attenuation associated with relying on self-reports. The detailed longitudinal data, linked to administrative data, allows us to control for many of the determinants of child outcomes that have been explored in previous literature, together with comprehensive sensitivity analyses, to assess the potential role of unobserved variables. The pattern of results strongly suggests that there are quantitatively important long run effects on victims – stronger than correlation analysis would otherwise suggest. In particular, we find that both type of bullying and its intensity matters for long run outcomes such as obtaining a degree, income, and mental health.

Acknowledgements: We are grateful to our respective institutions for providing support to facilitate our collaboration. Gorman and Walker were supported by the UK Economic and Social Research Council funded project ES/R003629/1 on the effects of schooling on long-term outcomes. Harmon and Staneva's contribution was supported by the Australian Research Council through a Discovery Program Grant (DP140102491), and the Centre of Excellence for Children and Families over the Life Course (CE140100027). None of the above played any direct role in the research. We are grateful to comments from seminar and conference participants at the Society of Labour Economics 2019 conference, the Royal Economics Society 2019 conference, Temple U, CUNY, Sydney U, Kentucky, VU Wellington; and particularly to Alexander Mas at Princeton, Ken Troske at U Kentucky, and Miguel Sarzosa at Purdue for their comments, support and advice. The data was provided via the secure server at the UK DataService and is available to other researchers subject to registration and training. The authors are happy to assist other researchers by providing their code and advice to facilitate replication and extensions to the work.

Declarations of interest: none

Corresponding author: Professor Ian Walker, Department of Economics, Lancaster University Management School, Lancaster LA1 4YX, UK. Email: ian.walker@lancaster.ac.uk

1. Introduction

Bullying at school is thought to be a widespread phenomenon that harms many children.[1,2] Yet there is relatively little quantitative research into the wider and long-term effects of having been bullied as a child—most studies concentrate on short term educational outcomes alone. Little research has explored the implications of the intensity (frequency) of bullying, or the implications of different types of bullying.[3]

This paper aims to quantify the impacts of bullying on important life outcomes. This is relevant for policy, as it allows estimates of the potential benefits of bullying reduction to put alongside the costs of anti-bullying policies (see Tofi and Farrington, 2011, for a review of existing anti-bullying programs).[4] We also highlight the differences in the effects by bullying type - evidence that may support a redistribution of resources towards tackling particularly harmful types.

We regard our primary contribution to be the use of detailed data to explore heterogenous effects of different bullying types and patterns, and to provide more credible estimates of the long-run effects of bullying than were previously available. We use data from a cohort study of English children matched to detailed administrative education records. We largely use statistical methods which rely on a selection on observed variables assumption, paired with a comprehensive range of sensitivity analyses and falsification tests, and an assessment of the robustness of estimates to deviations from the conditional independence assumption. In particular, we use least squares to adjust for observable factors, as well as matching and weighting methods to reduce any effects of functional form assumptions - employing propensity score matching (PSM) where we consider a single discrete treatment, and inverse probability weighted regression analysis (IPWRA) where we consider multiple treatments. The IPWRA analysis of treatments also facilitates the estimation of the effects of

---

[1] Throughout we refer to victimisation through bullying at school simply as bullying. Bullying in this paper is wholly school based – we do not consider, for example, workforce bullying.

[2] The 2017 edition of the Annual Bullying Survey, a large on-line non-random 'snowball' survey of young people in secondary schools and colleges across the UK, records 54% of all respondents had been bullied at some point in their lives. According to this survey, one-third of all victims experience social anxiety, one-third experience depression, and a quarter of the victims had suicidal thoughts.

[3] A recent comprehensive review of the psychology literature can be found in Ren and Voelkel (2017), and a succinct review of the education literature that focusses on England can be found in Brown (2018).

[4] For example, the influential *Olweus Bullying Prevention Program* aims to provide structured classroom discussions to discourage bullying and to reward helpful behavior, and has been positively evaluated, See Olweus (2013) for England. For Norway, the USA, and elsewhere see http://www.violencepreventionworks.org/public/olweus_history.page.

different types and intensities. As well as conventional robustness testing we employ a new test, developed by Oster (2019) from earlier work, that formalises the common approach on exploring the sensitivity of the coefficients(s) of interest to changes in specification.

We analyse a range of education and labour market outcomes: Advanced (A) level educational qualifications (usually taken at the end of senior high school at the age of 18) and A-level points score[5]; GCSE qualifications (usually taken at 16 at the end of compulsory schooling); and having a university degree. In addition, our data allows us to explore the effects on income; unemployment; and a mental (ill) health index – all recorded at age 25. Moreover, the previous literature has typically relied on a simple binary treatment as a measure of bullying victimisation, and we use this measure as well as more detailed treatment variables. In addition to a binary treatment, we use factor analysis to create a summary variable capturing the richness of the variation in the type and frequency of bullying. Secondly, we construct a multi-valued categorical treatment, which allows the effects of bullying to differ by type and intensity.

There are possible weaknesses in this analysis: in particular, estimates could be affected by attenuation (downward) bias, because of the subjective nature of self-reported victimization; and by (upward) bias due to selection on unobservables. We explore the use of cross-reported bullying (by parents) as a possible instrument for own-reported bullying in the spirit of the literature on measurement error. Results from this analysis are consistent with the main findings and our priors. In our setting, we have data on many of the determinants of bullying identified in the previous literature, and we build a credible case for dependence on a selection on observed variables assumption. However, we recognise that bias from unobserved variable may remain a concern. We investigated various interactions between the measures of the potential supply of bullies, vulnerability to bullying, and the school environment, as sources of exogenous variation – but we failed, as in Eriksen et al. (2014) and other attempts, to find a convincingly strong first stage and/or a convincing narrative to support . However, we do implement tests of potential for selection on unobservables that suggest it would take implausible levels of selection bias to drive our results to zero (see, in particular, Oster (2019), but also Krauth (2016), Nanninci (2007); and Ichino *et. al.* (2008)). We offer this work in the spirit of shining a light, on an important but difficult issue, where currently there is little.

[5] Usually in three or four relatively narrow subjects that were selected at age 16 and studied over a two-year period in senior high school. The grade results from these examinations are used as the primary admission criterion by universities and are often collapsed to a single A-level points score for this purpose.

We report a mosaic of results reflecting the range of possible definitions of the treatments, estimation methods, and control variables. Together, the results suggest that there are important long-run effects of bullying victimisation – possibly stronger than simple correlations analysis would suggest.

The rest of the paper is organized as follows. Section 2 describes the education system and the treatment of bullying. Section 3 briefly focusses on the key papers in the subset of literature that also attempt to provide causal estimates. Section 4 describes the data and the construction of the bullying intensity measure. Section 5 discusses the estimation methodologies. Section 6 presents the main results and our evaluation of them. Finally, Section 7 offers some reservations, concluding comments, and suggestions for further work.

2. Educational Context

Compulsory schooling in England starts at age four to five: children are expected to be attending full-time schooling at the age of five and are admitted each September to a "Primary" school before they reach that age. At age 11 (Year Six) there is usually a transition to "secondary" school for a further five years of study, although in a few areas there is an intermediate stage of schooling provided by "middle schools" that cover 11 to 13. The end of compulsory schooling is now somewhat blurred with children being expected to continue in school (usually in secondary schools from 16 to 18 often in the same location/campus as earlier schooling occurred, but sometimes in a "Sixth Form" college that admits children from nearby secondary schools for further study). Further Education colleges offer an alternative route of mainly vocational training up to age 18; and all those in work from age 16 are expected to combine this with at least 20 hours of training per week that may be based in a FE college or in the workplace (Harmon, 2017). There is often an element of selection by ability, based on earlier attainment, in the admission to the academic track in post-compulsory schools and sixth form colleges.

There is a common curriculum across almost all English schools which is organized into 'Key Stages' with KS1 being up to age seven, KS2 being from ages eight to 10, KS3 being ages 11-13, KS4 being ages 14-16, and KS5 being 17-19. There are low stakes tests at the end of each of KS1-3. At the end of KS4 at the age 16, students take the high-stakes General Certificate of Secondary Education examinations (GCSEs). Students are usually examined in between five and ten subjects, and usually need to attain passing grades (A, B, C) in at least five of them, including Mathematics and English, in order to be tracked into further academic study in senior high school. After their GCSEs, students may decide to pursue further studies from age 16 to 18, typically in just three or four subjects for study at Advanced (A) level, depending on their academic preferences and intentions toward higher education. Higher Education (HE) admission is driven largely by A-level examinations, taken at the end of KS5, that are graded A to E and grades are often converted into a cardinal scale by assigning points to grades to facilitate aggregation to a single metric.

The overwhelming majority of children attend publicly-funded secondary schools which admit children based on parental preferences and ration places at the margin, usually according to school proximity to home, if capacity constraints bind. These include "community" schools that are managed by their Local Education Authorities, although funding is provided under complex arrangements that involve central and local governments—and these arrangements are currently evolving into a national funding formula which allows for high need schools. Approximately 6% of high school children attend private schools, which usually have charitable status and operate on a not-for-profit basis – a small minority of which are faith schools. The majority are non-residential, but some provide "boarding" places. They can admit by ability and can charge fees and may provide bursaries. Home schooling is rare in England (current estimates suggest less than 0.5%), whilst a small proportion of secondary school children (less than 5%) attend academically selective publicly-funded (Grammar) schooling (a reduction from around 20% of much earlier cohorts—see Harmon and Walker, 2000). There are also publicly-funded schools that use religious background as an admission requirement and have a degree of autonomy from local government control. Finally, there are an increasing number other publicly-funded schools, known as Free Schools and Academies, that are similar to US Charter Schools in having a (greater) degree of autonomy from local government, are funded largely from central government, but are different in that both types operate on a non-profit basis.[6] For a broad discussion of the issues of school choice and school type see Burgess *et al*. (2015).

In the UK, the policy approach to bullying has not been prescriptive. A range of resources are made available for the school leadership, allowing the school to choose the most appropriate as opposed to the more formal processes seen in other countries. Thompson and

[6] Free Schools and Academies often belong to chains of similarly branded schools and are effectively new entrants to the sector. Many former community schools have converted to Academy status. Many have a faith focus to them. Academy status was originally given to failing secondary schools, in an attempt to turn them around, but this status has been increasingly given to successful secondary schools who are often then required to assist in the management and operations of nearby weaker schools.

Smith (2010) provide a detailed overview with a selection of case study schools that showed good practice in their anti-bullying work. Among the successful practices in the UK, the authors listed: proactive peer support (peer listening and buddy schemes, peer mediators trained in restorative approaches, and peer "playgrounders"); Head Teacher/School Principal 'open door' policy for parents and children; positive play sessions, safe-haven designated spaces; home visits; reactive and restorative strategies (see Thompson and Smith (2010) and Smith and Thompson (2014) for an overview). The Department of Education (DfE) has, from 2014, required schools to implement an effective anti-bullying strategy by adopting anti-bullying policies with clear definitions and procedures that are communicated to the whole school community (see Department of Education, 2017).

In addition to investigating educational achievement at school, we also consider the effects on longer term outcomes – at HE and in work. HE is usually pursued from age 18 at over 150 Higher Education Institutions (HEIs), some very small and specialised, which are collectively referred to as universities. Higher education participation rates are over 40% of the cohort and this has grown dramatically in the last three decades. Course fees have been dramatically increased (and public funding almost eliminated) as austerity measures since 2010 but there is now a comprehensive, sophisticated, and highly subsidized, student loan program that supports access, especially for low parental income students. Take-up of these loans is high and repayments are income contingent with any balance after 30 years being written off. As a result, demand for university is relatively inelastic to the posted price, and there is little evidence that fees have resulted in any fall in participation - neither overall, nor for low SES students (see Murphy et al, 2017). Dropping-out is relatively scarce (around 8% across the sector). Although England is geographically small, and so proximity to a university is much easier than in most other countries, the majority of students move away from their parental homes to study HE, and most of those that do will form (or join) households elsewhere when they graduate and start work. Finally, with respect to HE attainment, HEIs in England, (and Wales and Northern Ireland) offer undergraduate courses that are typically 3 years duration, studied mostly on a full-time basis and mostly straight from senior high school. Courses are usually specialized where a single narrow major is often pursued exclusively. Unlike the US, UK undergraduate professional courses such as law, medicine, and management are available across many HEIs.

3. Existing Bullying Literature

There are many papers that address bullying[7], but we focus mostly on the small literature that aims to estimate causal effects. In Table 1 we endeavor to interpret the effect sizes from these studies in a comparable way. Reviews of the work on bullying in the education and psychological literature can be found for example, in Sharp (1995), Ladd et al. (2017), Bond *et al* (2001), Due *et al* (2005), Arseneault *et al* (2010), Ford et al., (2017), Woods and Wolke (2004). Victims of frequent bullying have reported a range of psychological, psychosomatic and behavior problems including anxiety and depression, low self-esteem, mental health problems, sleeping difficulties, sadness, and frequent pain.

There is a relative paucity of economics research on bullying. The most relevant study to this work is Eriksen *et al*. (2014) which uses large surveys of Danish parents and teachers that record bullying prevalence and severity and combines this with outcomes from Danish administrative data on 9th grade (at age 16 in Denmark) based on tests in language and mathematics skills. Some 27% of the estimation sample record being bullied (to any extent), with 20% of those bullied reporting severe bullying. They estimate the relationship between bullying and future outcomes through an identification strategy based on classroom peer effects, assuming that the proportion of children whose parents had criminal backgrounds affects other life outcomes only through their effect on bullying by other children.[8]

The authors report an OLS estimate of -0.14 of a standard deviation of the grade point average (GPA) from bullying but find that bullied children have very much lower academic achievement in 9th grade in their IV results, although these results are noisy. Their results are robust to exclusion of individuals with no classmate parents convicted of crimes (13%) but not robust to excluding individuals with more than half of classmate parents convicted of crimes (7%). However, the authors do not provide supporting tests for their identification assumption neither do they discuss the relevance and validity of their instrument. Instrument validity is key in this work - it seems likely that having children from extremely challenging backgrounds in

[7] Little comparative work across countries about school bullying and its effects exists. Apart from OECD (2017), Due et al (2005) applied the same survey instrument to 123,227 students (age 11, 13 and 15) from a nationally representative sample of schools in 28 European and North American countries in 1997–98. There was widespread agreement across all countries that the health effects were negative and serious. Ammermueller (2012) uses a dataset of all students from classes in particular grades in randomly selected schools in the TIMMS project from 11 European countries. The author studies the effect of personally experiencing being physically hurt or experiencing theft at school and so is focused on severe bullying.

[8] It seems unlikely that classroom peer effects operate solely through bullying. For example, Carrell and Hoekstra (2010) show that troubled children have a direct negative spillover effect and significantly decrease the reading and math test scores of their peers.

the classroom would have an impact on other children in a variety of ways, and not just through a bullying channel. Their negative effects of bullying are even larger when they use teacher reported bullying than with parent reported bullying. It is possible that these two variables are correlated with the severity of the (unrecorded) actual bullying experienced by the child in different ways. Parents are probably more likely to get to know bullying because of changes in the behavior of their child - for example, making them more reluctant to attend school. In contrast, teachers are more likely to observe minor forms of bullying, as well as major ones.

Ponzo (2013) uses Italian data from the 2007 Trends in International Mathematics and Science Study (TIMMS) and the 2006 Progress in International Reading Literacy Study (PIRLS) programs. They use both OLS estimation and a Propensity Score Matching (PSM) analysis to model the probability of being bullied. Being bullied is defined as having a positive response to any question about experiences of each type of bullying behavior—so this is a very low threshold. In the OLS analysis, the author finds that bullying has bigger adverse effects on numeracy at age 13 than at age seven, while there is a similarly large negative effect of bullying on literacy using the age seven PIRLS data. The author also explores the effects of a count of different forms of bullying as an intensity measure and finds larger negative effects on numeracy.

Oliveira et al (2018) also uses PSM estimation applied to a sample of almost 30,000 children around age 11 in the Brazilian city of Recife in 2013 to measure the effect of bullying on numeracy test performance. Two definitions of bullying are used: 'definitely bullied' or 'maybe bullied' – in their descriptive analysis, however the estimates make no distinction between the two definitions. Their results suggest that bullying has a negative impact on test scores of around 0.5 of a standard deviation. Black, younger and students with high BMI are more likely to report being bullied.

Brown and Taylor (2008) use the much earlier National Child Development Study (NCDS) cohort of children born in a particular week in March 1958. It has the advantage that it records long-term outcomes and some, relatively crude, information on the bullies. The strength of this early contribution to the economics literature on bullying is that it uses a high quality and large cohort study that follows children through school and long into the labour market. Being bullied (at 7 and 11) is defined only from a maternal cross-report, and only in quite broad classifications (none, sometimes, often). From these responses the authors construct two indices to measure the extent of bulling at ages seven and 11. However, the authors do not account for the downward bias due to measurement error or the (likely, upward) bias due to the possibility of the existence of reverse causality in their estimates. They find that being bullied at school increases the likelihood of failing high school exams by 1.7 percentage points, while a one-point change in their bullying index at age 7 (or 11) decreases earnings by approximately 3.1 (or 2.8) percentage points.

Vignoles and Meschi (2010) use LSYPE (but only up to age 16 outcomes) in their analysis of the effect of bullying educational attainment at 16. They use OLS estimation and rely on lagged, rather than current, bullying (a count of the number of types of bullying reported by the main parent) and control for lagged outcomes, and a rich set of other controls. However, some controls are likely to be "bad" controls (absences, for example). Moreover, the bullying measure is a count of the number of types cross- reported by the parent and so treats violence as equivalent to name-calling. Finally, lagged cross-reported bullying is likely to be a very poor measure of current actual bullying so attenuation due to measurement error is likely to be large.

Another recent study by Delprato et al., (2017) examines the impact of bullying on learning and non-cognitive outcomes for sixth grade students in 15 Latin America countries using 2013 survey data, applying OLS and PSM methods. The study uses an overall measure of bullying and also two types of bullying, i.e. physical and psychological, however no intensity effects are documented. The authors report considerable variation in the prevalence of bullying across countries: physical bullying varies from 11% in Costa Rica to 26% in Peru, and psychological bullying between 25% in Mexico to 40% in Argentina. For the whole sample of the 15 countries, matching estimates show that bullied students achieve lower scores in mathematics and reading (about 0.11 of a standard deviation in learning outcomes).

Sarzosa and Urzua (2015) use a longitudinal survey of 14-18 years olds with matched administrative education data from South Korea, much like our LSYPE data, to identify the determinants of being bullied at age 15 on subsequent mental and physical health, and risky behaviors measured at age of 18 and older. The authors estimate a structural model of endogenous bullying and counterfactual outcomes. To facilitate identification, they use the random allocation of students to classrooms, that is a feature of the Korean schooling system, to provide exogenous variation that affects the probability of being victimized. They introduce two additional variables into their system of structural equations - the proportion of peers that self-report as bullies in the class, and the proportion of peers in the classroom that come from a violent family. The bullying definition refers to events where students have been severely teased, threatened, collectively harassed, severely beaten, or robbed. Neither the bullying intensity nor the impact of different types of bullying is explored in their model but their

Table 1: Summary of selected studies in existing literature

Authors	Year	Country Sample	Estimation Method	Dependent Variable	Main control variables	Effect size*
Brown and Taylor (2008)	1958	Britain—data from 1958 National Child Development Study (NCDS)	Ordered probit; OLS; IV	Education: number of GCSEs at 16; degree/none degree at age 23; wages at age 23, 33 and 42	Quadratic in maths and reading test scores, birth weight, body mass index, controls for number of schools attended, child physical characteristics, indicator for financial problems/unemployed parent, whether child is in care or attends special classes, personality, index of how frequently child prefers to spend time alone, whether child fights, and is upset by new situations.	6% to 25% lower prob of degree, diploma, O-level, or no qual; 2.5% lower wages
Vignoles and Meschi (2010)	2004-2006	Engalnd - LSYPE	Value-Added; School FE, RE	KS4 point score; Attitude to school at 16; Bullying at 16	Gender, ethnicity, English is the first language, eligible for free school meals; Special Education Need identified; unauthorised absences; attitude and behaviours likely to affect school choice of and pupils' engagement.	Insignificant 0.02SD bullying effect on outcomes.
Ammermueller (2012)	1965, 1969 2003	11 EU countries, data from 1958 NCDS; 2003 TIMMS	School fixed effects model	Reading at 11/16, maths/science at Grade 4/8; Highest education at 33; Earnings at 33	Gender, parents born abroad, social class of father, parent's interest, free meal, older/younger siblings, wears glasses, height, attractive look, twitches, BMI,BSACG score, teacher's initiative to discuss child, pupil-teacher ratio, school type, dummies for streaming of school, school FE.	Insignificant 18% in math test scores
Ponzo (2013)	2006-2007	Italian data from 2006 PIRLS and 2007 TIMSS	OLS, PSM	Reading, maths and science scores (Grade 4 & 8)	Gender, age, native parent, parent's education, total school enrolment, number of books at home, computer possession, own room, study desk, economic situation of the family, residence & city size dummies, % of students from disadvantaged families.	Around 22-23% in reading, maths and science
Eriksen et al. (2014)	2001	Denmark-administrative data	OLS, IV—using % of troubled home peers	Grades in Match and Danish at age 16 (Grade 9).	*Child controls:* gender, birth weight, birth complications, # younger siblings, ethnicity, #moves, mental disorders, emergency ward visit, psychosocial factors, impaired hearing, wear glasses, cross-eyed; *Parent controls*: age at birth, smoking, education, income, managerial level, mental behaviour, antidepressant. heart medication; classroom FE.	Insignificant IV 21% on GPA grades
Oliveira et al. (2018)	2013	Brazil, city of Recife; 6th grade students in public schools	OLS, PSM	Math test performance (Grade 6)	*Child controls*: gender, age, race, BMI, non-cognitive skills, any reported disease; *Parental Controls*: family per capita income, higher education, high school dummies, presence of those responsible for the student; *Teacher Controls*: gender, age, experience; *School controls:* class size, drop-out levels; absence, and proportion of girls per class;	16%-17% in math test score
Delprato et al.(2017)	2013	15 Latin American countries	OLS. PSM	Maths and reading test scores	Age, gender, whether repeated a grade, study conditions, family, school (school type, infrastructure) and teacher characteristic	Around 10% in maths and 13% in reading
Sarzosa and Urzua (2015)	2003	S. Korea KYP-JHSP longitudinal survey.	LIML structural estimation	Sickness, mental health, stress, and smoking at 18.	Younger siblings, income per capita, both parents present, and father's education levels.	75% rise in sickness, 50% in mental (ill) health, 20% in stress.

Notes: Effect size is expressed as a % of the SD of the dependent variable.

definition seems stronger than the usual "any bullying" used elsewhere. Sarzosa and Urzua (2015) show that non-cognitive skills reduce significantly the likelihood of being a victim of bullying. Although their bullying is subjective and self-reported, and therefore likely to be subject to measurement error that might attenuate effects, the authors estimate that victims have significantly higher incidence of self-reported depression, sickness, mental health issues and stress: being bullied at 15 increases sickness and mental health issues by 0.75 and 0.5 of a standard deviation, respectively, at age 18.

Overall, there is little coherence in the existing literature both in the definition of bullying used, and in the interpretation of outcomes. Most studies use one or two definitions and there is little that we can learn about the importance of different types, frequencies, intensities, and persistence of bullying on life outcomes. The problem is confounded by the differences in the dependent variables used, which have mostly focused on educational, rather than on long run outcomes. Only Sarzosa and Urzua (2015) have a credible identification strategy and they show severe long-term mental health effects - for South Korean youths who are well-known to already face the stress of highly competitive very high stakes school tests at the age of 15.

4. Data and Specification

We use a large representative cohort study of English children, born in 1989/90, who have been followed from age 13/14 to age 25 years, at which point educational attainment has largely been completed and labour market outcomes are recorded. The data is known as *Next Steps* (and also as the Longitudinal Study of Young People in England, LSYPE.[9,10] LSYPE covers a wide range of topics, apart from bullying, including family relationships, and attitudes toward school. It includes family, education, and labour market variables, and covers sensitive or challenging issues, such as risky behaviours, and personal relationships. LSYPE selected observations to be representative of the English population, but specific groups were oversampled - in particular, youths from low socioeconomic backgrounds and minorities (see

[9] The Wave 8 survey sought consent from LSYPE participants to allow further administrative data matched to LSYPE. We intend to return to this issue if such a longer-term follow-up of the LSYPE cohort becomes successful.

[10] The data is similar in structure to the earlier, shorter, and smaller, US National Longitudinal Study of Youths (NLSY) dataset that has been extensively used in other longitudinal research studies in other contexts and, in this bullying context, by Lam (2016).

Department of Education, 2010). More details can be found in Centre for Longitudinal Studies (2016) and Anders (2012).

The survey started in 2004 when the young people were at the age of 13/14 (in school year 9). In the first wave of LSYPE, around 15,000 young people were interviewed across more than 700 high schools. The survey continuously followed these individuals for 7 years (age 14-21) and then re-interviewed them in Wave 8 at age 25. The non-response rate in the first wave was approximately 25%, and thereafter there was approximately 10% attrition in each subsequent annual wave. There was then a four-year break between Waves 7 (age 21) and 8 (age 25) – a period when a lot of new household formation occurs, which contributed to a further drop. There does not seem to have been any substantial attrition as children completed compulsory schooling or when the survey moved to mixed (a choice of either conventional survey home/school visits or new on-line completion) methods. The survey data are matched to an administrative register known as the National Pupil Database (NPD), which includes the LSYPE sample of that 1990 birth cohort and detailed histories of educational attainment.

4.1 Outcomes

We study the impact of bullying on the following outcomes—most of which we think of as being long-term ones, but we also include the most important proximate high-stakes educational outcomes:

- Having 5+ GCSE or GNVQ passes, including Maths and English, which is an important criterion for advancing, after 16, on an academic track ("*5+ GCSE*")
- Having an A-level qualification, or vocational "Level 3" qualification which attracts UCAS points to contribute to university entrance ("*Any A-levels*")
- Sum of UCAS points, based on the best three qualifications - most commonly A-levels, but can include other qualifications[11] (*"Best 3 A-level points")*
- Receiving a university degree (*"University degree")*
- Natural log of weekly income (*"Ln Income")*
- Not in employment, defined as not being an employee or self-employed, and so includes not in the labour force (*"Not employed")*
- General Health Questionnaire, measuring mental ill-health from 0 to 12, where 0 represents perfect health and 12 represents maximum distress (*"Mental health").*

[11] The total UCAS tariff points (which are assigned based on the grade achieved) from the best three A-level or equivalent qualifications are commonly used as the basis of admission by most UK HEIs. See: https://www.ucas.com/sites/default/files/2015-uk-qualifications.pdf .

4.2 *Bullying Data*

Our bullying data is unusually comprehensive because it consists of five types, seven frequencies (including none) and three waves of data, providing the flexibility of defining a large number of possible treatments. The data asks students (and the main parent) whether the child was a victim of bullying in the last year. In particular, in each of the first three waves young people were asked whether they had experienced any of five forms of bullying last year:

- Upset by name-calling, including text or email (N);
- Excluded from a group of friends (Social exclusion, S);
- Made to hand over money or possessions (E, for extortion);
- Threatened with violence (T);
- Experienced actual violence. (V)

In addition to type of bullying, the data contains information on frequency: "every day"; "a few times a week"; "once or twice a week"; "once every two weeks"; "once a month"; and "less often than this".[12] However, estimating close to one hundred treatment effects on a dataset with a relatively small sample is unlikely to yield precise estimates. We therefore examine appropriate ways of creating summary measures that seem acceptable to the data and we create three different types of bullying treatment variables. In preliminary OLS estimation, available on request, we use nested testing to aggregate types and intensities to achieve a statistically acceptable specification that would be sufficiently parsimonious to allow estimation using a number of methods. The first definition of a treatment is a binary variable equal to one if a child has experienced any bullying across the three waves, and zero otherwise. The overwhelming majority of the existing quantitative literature uses just one variable to define bullying, and this treatment provides a baseline specification that is comparable with previous studies. Second, we define a richer summary measure, using factor analysis, which combines information on type and frequency of bullying over the three waves. Finally, we define a multi-valued categorical variable to capture potential heterogeneity in treatment effects.

The rationale behind these variables is as follows. Rather than impose constraints on the raw data to generate more parsimonious specifications, we first take a data-driven approach

[12] A not insubstantial group indicate the response "it varies" (n=885/7,569), and we set their frequency to missing in the reported results. However, in further analysis that is available on request, we have also explored alternative imputations which do not materially change the results.

using exploratory factor analysis.[13] We conduct the factor analysis on the frequency of bullying variables, which are distinct by type and wave. We find evidence of just one common factor which we interpret as a measure of cumulative bullying intensity.[14] This score is standardized to have a mean of zero and a standard deviation of one, which allows us to interpret subsequent results in terms of a standard deviation of the bullying intensity. This approach extracts the variation available by type, frequency and wave in a data-driven, pragmatic way. The third variable we create aims to allow different effects by type and intensity of bullying. We first reduce the number of treatments by collapsing the number of types to two, by combining the three types that relate to violence (actual violence, threatened violence, and demanding money or belongings under duress) and collapsing the two non-violent types (name calling and social exclusion) into one. This is largely a practical matter to preserve cell sizes. We justify this aggregation on the grounds that that some types, e.g., extortion, have a very low incidence so the data would be unlikely to have the power to detect small effects on outcomes, and the variables in these grouping are naturally correlated: extortion usually occurs because of violence, or some implied threat.

To reveal heterogeneity in the treatment, by type and intensity, we begin by summing across waves for each of the two types separately, to produce a cumulative sum of bullying instances (this could also be achieved by imposing the same coefficient on each wave's frequency variable for each type separately). We start by imposing cardinal interpretations to the bullying frequency reports. That is, we define frequency not as an indicator for each level, but as a number corresponding to the level. [15] This restriction does not allow heterogeneous effects by the timing of bullying, but rather measures the *cumulative* effect of being bullied. After collapsing to two types, we create two continuous variables by summing the total instances of violent and non-violent bullying instances across the three waves. For example, because for each type there are a maximum of 200 instances in each wave, the maximum number of non-violent instances across the three waves would be 1200.

[13] Factor analysis is commonly used when using data sets with large numbers of observed variables that are thought to reflect a smaller number of underlying latent variables.

[14] These are found using standard procedures according to which only factors with eigenvalues greater than or equal to one should be retained. See Fiorini and Keane (2014) for a similar application. The first factor explains 73% of the variance. We tried oblique rotation techniques to allow the factors to be correlated but the rotation did not affect the estimates.

[15] Assuming 200 school days in a year, we make the following imputations: "every day" = 200 instances per annum; "a few times a week" = 100 ; "once or twice a week" = 60 instances; "once every two weeks" = 20 instances; "once a month" = 10, and "less often than this" = 2.

To capture the heterogeneity in the pattern of bullying, we create a multi-valued treatment variable summarising the violent and non-violent frequency variables. We create a variable which takes on nine values indicating each combination of: violent, non-violent, no or little bullying, moderate bullying and high bullying. No or little bullying is defined as a frequency of zero days, or the lowest frequency of two days. This means this lowest category is 0 to 4 days for non-violent bullying (2 days multiplied by 2 types) and 0-6 days for violent (maximum of 2 days multiplied by 3 types). High bullying is defined as being in the top quartile of the bullying frequency distribution: experiencing 100 days or more of bullying in a school year. Moderate bullying is the remaining group. Table 2 summarises the nature of this variable.

In summary, we have focused on three definitions of bullying – a *binary variable* indicating whether the pupil has been bullied, of any type or frequency, at any point over the three waves of data (and a corresponding variable based on the parent reports); a *continuous variable* constructed via a factor analysis of the frequency of each type of bullying in each wave (and a corresponding variable based on the parent reports); and a *multi-valued discrete treatment* for each combination of violent or non-violent bullying type, and none, moderate or high cumulative frequency of occurrence over three waves.

Table 2: Nine categories of the multi-valued treatment

		Non-violent		
		None	*Moderate*	*High*
Violent	*None*	72%	10%	3%
	Moderate	3%	3%	2%
	High	1%	1%	3%

Notes: Cell percentages not add to 100% due to rounding.

4.3 Summary statistics:

The most general sample for analysis is restricted to individuals who participated in Wave 8, to yield long term outcomes, and also participated in Wave 1 and have complete data on the most basic set of covariates we use (N=7,569). As we add further covariates and consider outcomes from various sources in our linked administrative data, the sample reduces. Testing for differences in key characteristics across the different estimation samples does not reveal significant differences.[16] LSYPE contains survey weights, to adjust for the complex survey

[16] Not shown, but available by request.

design (a function of ethnicity, area deprivation and school type, among other factors) and survey drop-out (modelled as a function of observed characteristics in the data). We may wish to use the weights if we suspect they may be correlated with our treatment effects, i.e. that the survey design or survey drop-out may bias our results. In the main analyses, we do not use the weights. However, where we can we have also fitted the models with the survey weights, yielding negligible differences in our parameters estimates, such that we feel confident using the weights would not alter our findings more generally.[17] However, we do adjust the standard errors for clustering by school.

Summary statistics for the outcomes and the control variables are provided in Table 3. These statistics are unweighted and should not be interpreted as population-representative estimates. Some 45% of children are male; 69% self-report white as their ethnicity, 6% of all children report that English is not their first language; the KS2 and KS3 scores are average points scores from the National Pupil Database (NPD), and are recorded at age 10 and 13 respectively; and 16% of children live with just one of their biological parents. Parents were asked if their child was in their first ranked secondary school—which we include because a child might be more likely to be bullied and have lower achievement, irrespective of bullying, if the child has not been able to gain admission to her most favoured school. 82% are placed in their first-choice school. The Index of Deprivation included in the analysis is the IDACI (income deprivation affecting children index), a subset of the Index of Multiple Deprivation, measuring the proportion of children aged 0 to 15 living in income deprived families, defined including people out of work, and people with low income (Department for Communities and Local Government, 2015). Locus of control captures individual beliefs about whether life events are mostly internally or externally determined (Rotter, 1966). People with an external locus of control believe that they cannot have an impact on what happens in life, as events largely depend on circumstances beyond their control. On the other hand, individuals with internal locus of control generally believe that life events are mostly caused by their own decisions and behaviours. We measure locus of control using children's responses to six questions and we use factor analysis to create a continuous index of locus of control. LSYPE includes four questions on working attitudes (see the Appendix for details for these questions) and we use factor analysis to create an index of work ethics from these.

Table 3: Summary statistics for key variables

	Mean	SD	N

[17] Results not shown, but available by request.

Male	0.45	0.50	7,569
Child's ethnic group			
White	0.69	0.46	7,569
Asian	0.17	0.38	7,569
Black	0.06	0.23	7,569
Other ethnic	0.07	0.26	7,569
English second language	0.06	0.24	7,569
Index of area-deprivation	0.22	0.18	7,030
KS2 average points score	27.43	3.92	6,945
KS3 average points score	34.97	6.39	6,960
Highest parental qualification			
Degree or HE	0.25	0.44	7,569
A-level	0.14	0.34	7,569
GCSE	0.26	0.44	7,569
Low or no qualifications	0.30	0.46	7,569
Age of main parent	43	6.0	7,503
Parents separated	0.16	0.36	7,569
At first choice school	0.82	0.39	7,569
Locus of control	0.05	1.00	5,406
Work ethic	0.13	0.96	6,204
5+ GCSE	0.69	0.46	6,698
Best 3 A-level points	2289	99.40	4,018
A-levels	0.51	0.50	7,569
University degree	0.37	0.48	7,569
Income (£ weekly)	303.4	72.5	7,569
Unemployed	0.10	0.31	7,569
Mental health	2.30	3.12	7,234

Note: Weighted using Wave 8 sample weights. Unweighted statistics are provided in Web Appendix Table A1

The parental education variables reflect the rapid expansion that had occurred in HE provision in the late 80's and early 90's so that 37% of the children have gained a HE degree compared to 25% for their mothers - the interviewed "main parent" is the parent most involved in the child's schooling, and is almost exclusively the mother. We have a wide variety of outcomes. The proportion attaining 5+ GCSE passes, 69%, comes from the NPD data and is matched into the LSYPE data. Whether the individual took any A-levels (or equivalent "level 3" qualifications), 51% in Table 3; and the sum of the points of the best three subjects taken (excluding General Studies – a very broad subject that is sometimes taken as a fourth A-level subject) using the grade to points conversion scale prevalent at the time, is taken from wave 7. Income is recorded for the individual in wave 8 of LSYPE. Unemployed is defined to include those not in the labour force (i.e. the unemployed are all who are not self-employed or an employee). Mental health is measured using the General Health Questionnaire (GHQ) index, which is a count of up to 12 conditions where a higher score indicates poorer mental health.

Table 4 reports means and standard deviations of key variables by bullying status: whether a child has *never* been bullied, has been bullied *once*, or has been bullied *multiple* times. Boys are slightly more likely to report being bullied than girls. White families are overrepresented among the repeated bullying category compared with other ethnicities. Children in sole parent families are statistically significantly more likely to face multiple instance of bullying compared with those in two-parent families. There appears to be little difference in propensity to be bullied by measures of socio-economic status, such as the area-based deprivation index (IDACI), or parental education level. This makes sense because a key determinant of being bullied is being *different* from those around you, rather than the levels of any particular variable. There are differences in outcomes by bullying status, especially mental health, unemployment, and income.

Table 4: Differences in key variables by bullying status

	Never bullied	Bullied once	Repeatedly bullied
Male	0.44	0.47	0.42
White	0.62	0.70	0.80
Asian	0.23	0.17	0.09
Black	0.07	0.06	0.04
Other ethnic	0.08	0.07	0.07
ESL	0.07	0.06	0.04
Index of deprivation	0.23	0.22	0.20
KS2 average points score	27.54	27.27	27.46
KS3 average points score	35.23	34.62	34.95
Parental qual = Degree/HE	0.24	0.25	0.28
Parental qual = A-level	0.13	0.13	0.14
Parental qual = GCSE	0.25	0.26	0.28
Parental qual = Low/no qual	0.33	0.31	0.26
Age of main parent	42.54	42.42	42.44
Parents separated	0.14	0.16	0.18
At first choice school	0.82	0.81	0.82
Locus of control	0.12	0.02	-0.01
Work ethic	0.21	0.08	0.08
5+ GCSE	0.74	0.67	0.65
Best 3 A-level points	232.31	227.57	225.04
Any A-levels	0.53	0.48	0.53
Has a degree	0.39	0.35	0.35
ln(Income)	5.67	5.67	5.72
Unemployed	0.09	0.11	0.12
Mental health	1.74	2.44	2.92
N	3,087	2,341	2,141

Figures 1 and 2 give a sense of the distributions of bullying frequency by type of bullying and wave (among those who report both). Figures 1a and 1b shows the extensive margin of victimization experience by type—that is, the proportion of girls and boys reporting (any frequency of) each type of bullying in each Wave. Victimization falls across waves for each type, consistent with the existing literature. Comparing Figures 1a and 1b it is also clear that name-calling and social exclusion are more prevalent for girls and violence more prevalent for boys. Figures 2a and 2b show the intensive margin of victimisation by type and wave—that is, the average numbers of days the youths report experiencing each type of bullying in each wave. Again, victimisation falls over waves and, conditional on positive instances, boys tend to experience more instances, especially of violent types.

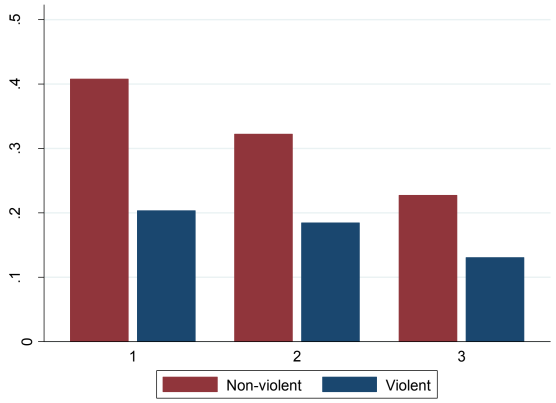
Exploration of the degree of serial correlation in bullying across waves suggested that this was high, for all three main types. For this reason, we feel justified in thinking that frequencies for each type could be aggregated across waves—that is, it may not matter than a bullying instance occurred in Wave 1, what matters is that is the cumulative total of bullying experienced. Figure 3 compares the child and parent reports of experiencing bullying. Typically, the child reports show a higher prevalence of bullying. The reports from both child and parents follow a similar downward trend over the three waves reflecting the decrease in bullying as children mature. It is useful to have a second report of bullying, even though both are subjective, since it allows for a useful sense check on the importance of measurement error on the estimates of the effects on outcomes. We provide descriptive evidence of the outcomes associated with each type and frequency of bullying in Appendix Figures A1 and A2.

We group the days of bullying instances into the three levels defined earlier (None, Low, High), and show, for each intensity group cell, the means for each of our outcomes. These figures show the expected pattern, that increasing bullying intensity is associated with worsening outcomes. This pattern is especially pronounced for unemployment and mental ill-health. The graphs also foreshadow non-linearities in the effects of bullying: moving from moderate to high bullying is associated with a larger drop in outcomes, compared with moving from no bullying to moderate bullying. This is an issue we return to in our modelling.

Finally, although the graphs show important differences in the incidence of different types of bullying by gender this is not something that we find we can pursue in detail. We find that when we split the data by gender there are few statistically significant differences on the effects by gender. However, we also find that the results generally lack precision. So we feel that the data is underpowered to reveal gender differences and this must await larger data.

Figure 1: Bullying participation by wave and type

(a) Girls *(b) Boys*



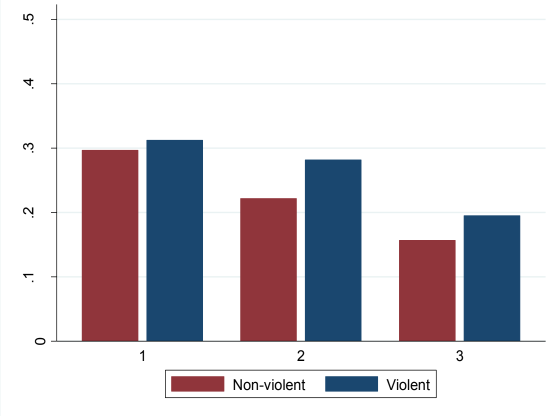
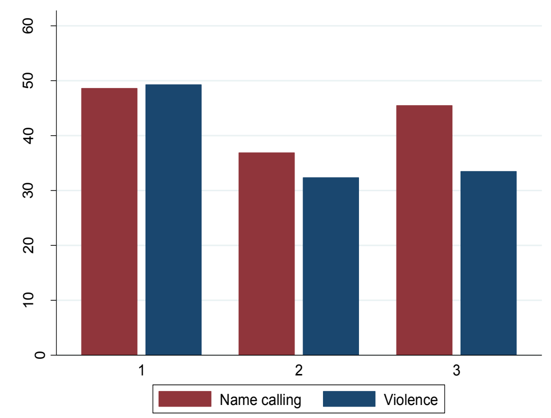


Figure 2: Bullying days by wave and type

(a) Girls *(b) Boys*



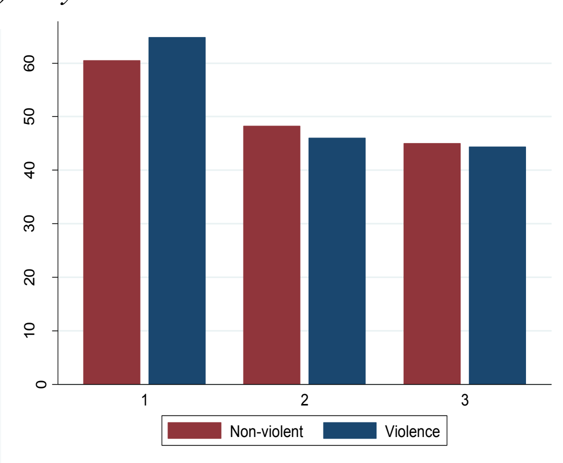
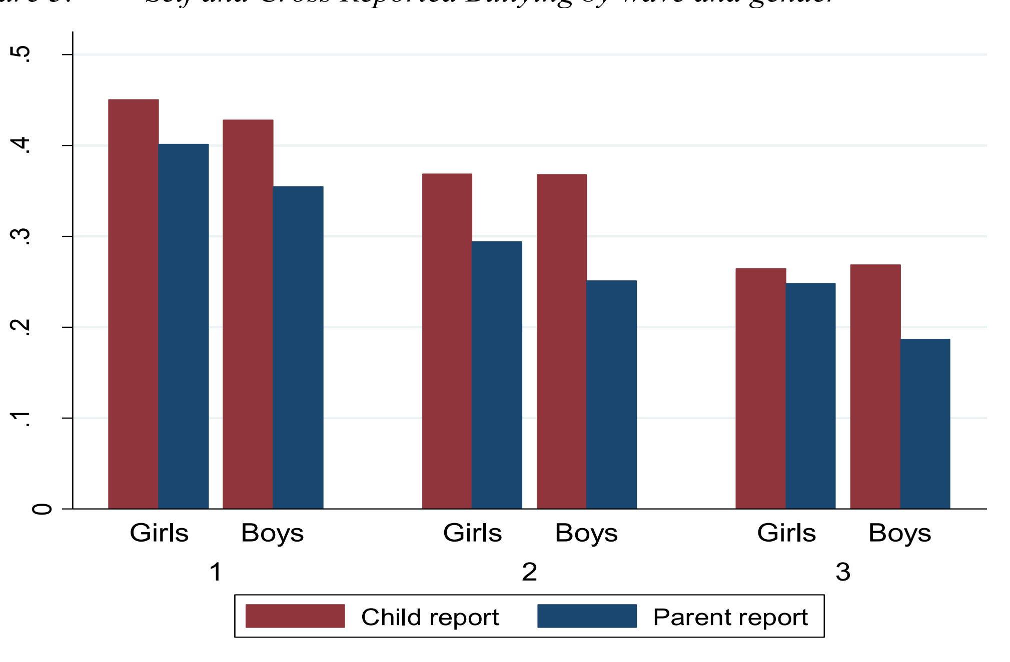


Figure 3: Self and Cross Reported Bullying by wave and gender



Notes: These charts show the unweighted proportions of cohort members experiencing each type of bullying by survey wave (1,2,3) and gender. 'Non-violent' includes social exclusion and/or name calling, 'Violent' includes threats of violence, actual violence and extortion.

5. Estimation

We explore a range of empirical methods, which rely on various assumptions. We first consider OLS estimates, as a benchmark, then propensity score matching (PSM), and finally treatment effects with inverse-probability-weighted regression (IPWRA).

5.1 OLS analysis

We begin by estimating the following simple linear relationship using OLS:

$$Y_{ih} = \mathbf{B}'_{ih}\boldsymbol{\beta} + \mathbf{X}'_{ih}\boldsymbol{\gamma} + \epsilon_h + \omega_{ih} \tag{1}$$

where Y_{ih} represents one of the several outcomes, observed at age 16, 18 or 25 years depending on the outcome in question, for individual i who attended high school h; $\mathbf{B}_{ih}$, represents the bullying treatment variable which may be a scalar of vector, for student i attending high school h; $\mathbf{X}_{ih}$ is a vector of child characteristics (e.g. ethnicity, month of birth, etc), school characteristics (e.g. school type), and family characteristics (e.g. maternal education and marital status), and ϵ_h is a school fixed effect while ω_{ih} captures unobservables that vary across i and across h. The inclusion of the school fixed effects allows us to account for unobserved time-invariant school characteristics, which may affect bullying and students' outcomes at the same time—for example, the disciplinary regime at the school. Using school fixed effects in many of our models allows us to capture the idea that it is the *relative* characteristics of pupils, compared with one's proximate peers, which are important for determining whether a child is bullied.

In this specification the coefficients on our $\mathbf{B}_{ih}$ indicators, $\boldsymbol{\beta}$, are the parameters of interest. While the OLS estimator adjusts for observable factors, the resulting estimates do not necessarily warrant a causal interpretation. The plausibility of the conditional independence assumption required for a causal interpretation depends on the relationship between the outcomes and the covariates $\mathbf{X}_i$. As such, it has become common to explore the stability of the parameters of interest by varying the set of control variables $\mathbf{X}_i$. In particular, $\mathbf{X}_i$ might include pre-treatment controls - specifically, KS2 scores that might reflect pre-treatment bullying in primary school. We use two sets of covariates, each including school fixed effects. The first is a parsimonious specification that includes only those variables that seem plausibly exogenous: gender, ethnicity, month of birth, Government Office Region (GOR) and English being a second language (ESL). The second is an intermediate specification which also includes a set of controls which we think of as being predetermined in Wave 1 of the data (age 14): local area deprivation, parental information including age, education, health, income and marital status, low stakes test scores at age 10 (KS2), and whether the school was the parent's first choice school. We would be concerned that any further extension beyond this specification would run a risk of including "bad controls" which would generate biased coefficients.

We implement a number of falsification, or placebo, tests. We assess the effects of the binary bullying variable on variables which should *not* be impacted by bullying if we have adequately controlled for selection into being bullied, i.e., they are either determined *before* bullying occurred, or are measured afterward but there is no reason to believe that they should be affected by bullying. Therefore, we expect to not see any significant effects of bullying in this analysis, unless our observed effects of bullying are driven, to some extent, by confounding. Finding appropriate predetermined variables in our data is difficult, but we identify the following candidates: the share of pupils in the school gaining 5+ GCSEs in *2001* (the first wave in the estimation sample is 2004); the Key Stage 2 scores of pupils attending the school in 2001; the deviation of the pupil's average height measured at age 25 from their high school peers; and whether the pupil took either the Math or Science Extension Test at primary school. The rationale for the deviation from average height of peers is to pick up children who may have been relatively small at school, and therefore more likely to be bullied due to their physical attributes. We do not observe pupils' heights while they are at school, only at age 25 years, so we need to make the strong assumption that relative height has stayed constant. But if we have adequately controlled for the determinants of being bullied, we should not see an 'effect' of bullying on relative height at age 25 years. Finally, taking the Math or Science Extension tests could represent a proxy for being an intellectual or social outlier, as measured prior to high school.

5.2 PSM analysis

We complement least squares estimation with propensity score matching (PSM). Matching offers a number of advantages compared with OLS: increased similarity (balance) in the distribution of covariates between the treated and control group; explicit consideration of the degree of overlap; and a reduced reliance on a linear functional form. The primary approach we use is kernel propensity score matching. We complement this with a number of alternative estimation methods, to ensure our results are not an artefact of one particular approach: nearest neighbour (NN) propensity score matching, and multivariate distance matching on the Mahalanobis distance (MDM). We also report a histogram showing the resultant overlap between treated and control units, and a plot summarising the balance statistics.

To evaluate the sensitivity of the estimates to confounding, we employ the sensitivity analysis developed in (Nannicini, 2007; Ichino *et. al*., 2008) and applied in other applications in labour economics such as Borra, et. al, (2012). This sensitivity analysis simulates the effects of a potential binary confounder on the average treatment effect on the treated. This method is similar in concept to many other sensitivity analyses in the statistics and econometrics literature who also assess the sensitivity to unobserved confounding (for example, Oster 2019). One advantage of this specific approach is that is does not require a parametric outcome model, making it suitable to use in a matching context

The idea is that we may suspect that the conditional independence assumption may not hold, given the covariates we observe. However, we might think that conditional on an omitted variable, denoted *U*, the assumption would plausibly hold. Matching on *U,* in addition to the vector **X**, would allow us to obtain a consistent estimate of the ATET. By specifying the joint distribution of *U*, the binary treatment, denoted *B*, and the outcome, denoted *Y*, we can compute the "unbiased" ATT, which accounts for the confounding effects of *U*. We can compare this to our original, potentially "biased" estimate, which does not adjust for *U*, to assess the difference made by accounting for the unobserved covariate.

To operationalise the method, one needs to specify the distribution of a hypothesised *U*, in relation to *B* and *Y*. Equation 2 highlights the maintained simplifying assumption that *U* is binary and independent of **X**.

$$p_{ij} \equiv pr(U = 1 \mid B = i, Y = j) = pr(U = 1 \mid B = i, Y = j, \mathbf{X}) \qquad (2)$$

After specifying p_{ij}, the relevant value of *U* is assigned to each observation, depending on which category of *i,j* they are in, and *U* is included in the calculation of the ATET as an additional covariate. For a given set of parameters, the matching procedure is performed multiple times with varying draws of *U*, and the estimate of the ATET is the average over the estimate in each simulation. The standard errors are calculated using Rubin's rules for computing standard errors across multiple datasets.

The first way we operationalise this is to pick *U* such that the unbiased effect would be driven to zero, and then assess the substantive plausibility of such a confounder. A second way to operationalise this is to specify *U* to mimic the distribution of some observed confounder, and this may represent a more plausible scenario.[18] Therefore, in addition to a hypothetical *U*

[18] We use the user-written Stata program *sensatt* to implement these procedures.

which drives the treatment effect to zero, we also look at the potential confounding effects of hypothesised confounders which have the same distribution as some variable that is observable. We choose three such variables to explore this: the "sole parent family" variable; the "English second language" variable; and a binary variable. which we call "outlier", that indicates being in either the top or the bottom decile of the Key Stage 2 distribution in their school (i.e., compared with being in the middle of the distribution as the base category). We choose these particular variables as it seems plausible that they may possibly affect both the probability of being bullied and the outcomes. We then assess the extent to which these hypothetical confounders would reduce the estimated treatment effect.

To assess the economic plausibility the hypothetical confounder *U*, when specified to reduce the treatment effect to zero, we report both two types of odds ratios: the *selection effect* and the *outcome effect (*Nannicini, 2007; Ichino *et. al*., 2008). The *selection* effect quantifies the degree to which the posited unobserved covariate increases selection into being bullied: specifically, the odds of being bullied associated the binary confounder taking the value one, divided by the odds of being bullied associated the binary confounder taking the value zero. The *outcome* effect quantifies the degree to which the posited unobserved covariate increases the average outcome: specifically, the odds of a binary outcome associated with having the confounder taking the value one, divided by the odds of a binary outcome associated with having the confounder taking the value zero. The idea is that if an unobservable must have implausibly large selection and outcome effects to materially change our results then this would provide evidence supporting the robustness of our results. Results from all these tests are presented below, where appropriate, and generally confirm that an implausible level of selection on unobservables would be needed to invalidate the main findings.

The OLS and PSM analysis so far has employed a simple binary treatment. To improve on this, we also consider a continuous treatment constructed using factor analysis on the frequency of each type of bullying in each wave. Beyond this data reduction approach we consider multiple treatments defined by the varying intensities and types of bullying.

5.3 Treatment effects with IPWRA analysis

We also examine the role of different types of bullying using inverse probability weighted regression adjustment (IPWRA) treatment effects estimation based on Imbens and Wooldridge (2009) and its implementation in Cattaneo et al. (2010).[19] We use IPWRA to explore the effects

[19] These effects are estimated using the *teffects ipwra* routine in Stata 15 (STATA Corp, 2017).

of a multi-valued treatment taking nine values: each combination of no bullying, low bullying and high bullying frequency, for two types of bullying (violent and non-violent).

Specifically, the probability of "treatment" (in this context, having a certain combination of violent/non-violent and low/high frequency bullying) is estimated using a multinomial logit specification. The inverse of these predicted probabilities are used as weights in a second-stage regression (Wooldridge, 2007; Wooldridge, 2010; and Imbens and Wooldridge, 2009). IPWRA re-weights the sample based on the inverse probability of treatment, and fits OLS regression on the reweighted sample. The IPWRA estimator has the so-called "double robustness property" (Wooldridge, 2007 and 2010) in that only one of the two equations in the model must be correctly specified to consistently estimate the parameters of interest. That is, estimates in the second stage (the outcome equations) are robust to misspecification of the first stage (the multinomial logit model of treatment propensities) provided that the second stage is correctly specified. Similarly, estimates from the first stage are robust to the second step provided the weighting is correctly specified. Nonetheless, estimation by IPWRA relies on the conditional-independence assumption in order to identify the effect of bullying on long term outcomes. If we have enough information on the observable differences between youths with and without the treatments, we can heavily weight treated observations that have similar observables to untreated individuals and obtain unbiased estimates of the causal relationship between bullying and long term outcomes (Mendolia and Walker, 2015). This approach increases the similarity of the covariate distribution of the treated and control groups via reweighting, leading to reduced reliance on the OLS functional form.

5.4 Instrumental Variable analysis for measurement error

To account for the possibility that the coefficients of interest could be attenuated because of for measurement error, we therefore test our main findings against those from an instrumental variable approach that exploits parental cross-reports of bullying as an instrument for self-reported bullying. Our identification strategy is based on cross-reported bullying types and overall frequency. The estimation model consists of a first stage model of bullying as a function of main parent cross reported bullying (MPB), defined in the same way as the dependent variable. The exclusion restriction rests on the assumption that bullying reported by the main parent does not affect individual's long-term outcomes directly. IV estimation uses a smaller sample, because they rely on the frequency report of both parents and children to be non-missing, in all waves, not just at the extensive margin. The estimated model has the following two stages:

$$B_{ih} = \mathbf{MPB}_{ih}\theta' + \mathbf{X}'_{ih}\gamma_1 + \vartheta_h + \varepsilon_{ih} \tag{3}$$

$$\Upsilon_{ih} = \widehat{\boldsymbol{B}}_{ih}\beta' + \mathbf{X}'_{ih}\gamma_2 + \epsilon_h + \omega_{ih}$$

where $\mathbf{MPB}_{ih}$ is the cross-report by the main parent of child i in school h.

One concern in this analysis is that parents who report bullying may be systematically different from those who do not report it, and that they may put some strategies in place in order to support their child and help her/him navigate through these difficult experiences because of that systematic difference. If these characteristics or strategies also affect long-term outcomes, then our estimates would be biased. A similar argument has been used in other educational production function inputs. This kind of compensating parental behaviour is more likely to be found among parents who are more involved in their children's lives and possibly more able to support their children. We expect these parental characteristics to have a positive effect on children's long-term outcomes, and therefore this is likely to make our OLS estimates more conservative. This is an argument for thinking that our cross reports identify lower bounds to the true effects.

A further concern in the literature is that an IV strategy will only produce consistent results if the measurement error is classical. Although, Light and Flores-Lagunes (2006) show, in their context, that their classical measurement IV error modelling produce comparable estimates to more complex cases, we provide results only for the continuous case where the classical assumption seems more reasonable.

5.5 Instrumental Variable analysis for Selection on Unobservables

Finally, despite our reservations about IV to deal with selection on unobservables, we do explore the possibility of using interactions of observed variables as candidate instruments using an instrumental variables strategy to account for any remaining selection on unobservables. We explored a number of potential instruments, chosen to reflect the *supply* of bullies (unauthorised absences at primary school of your high school peers); *vulnerability* to being bullied (absolute deviation from your high school peers in ability and other characteristics); and the school *environment* (parents' of peers perceptions of discipline at the school). However, we were not able to convince ourselves of the validity and/or power of any interactions. We are therefore pessimistic about being able to use IV to correct for selection on unobservables, in the absence of any natural experiment driven by policy – since policy has been somewhat laissez-faire we are not able conceive of such an experiment to date.

Instead, to examine the *potential* role of unobservable variables, we use recently developed tests that explore the stability of the coefficient(s) of interest in the face of increasing the set of control variables (see Oster, 2019, and Krauth, 2016, which have, in turn, been developed from Altonji *et al*, 2005). We report estimates of the parameter δ, developed in Oster (2019), that indicates the level of selection on unobserved variables, as a proportion to the level of selection on observed variables, that would be required to drive the treatment effect to zero. The assumptions underpinning the calculation of δ can be varied. In particular, the researcher can vary the assumed value of R^2*-max,* the *R*-squared from a hypothetical regression of the outcome on treatment and both observed and unobserved controls. The default option is to set this as 1, which may not be plausible in situations where it is inconceivable that one might be able to explain *all* the variation in the outcome. A rule of thumb proposed in Oster (2019) is to set *R-max* equal to 1.3 times the *R*-squared from a regression of the outcome on the treatment and observed control variables (denoted $\tilde{R}$). The suggested cut-off to define an 'acceptable' level of selection is an estimate of δ (calculated using *R-max* =1.3*$\tilde{R}$) that exceeds 1. This was the level that was found to be consistent with that observed in a sample of papers using RCTs in Oster (2019). Therefore, we report δ based on this level of *R-max*. Results from this test are reported in Section 6 below, and they confirm the credibility of our main estimates.

6. Results

We first present headline results for our two simplest cases: where bullying is a discrete variable corresponding to reporting "any" bullying or not; and a continuous variable derived from factor analysis. In lieu of a convincing identification strategy to deal with the likely endogeneity of bulling we present estimates for a parsimonious specification and a more comprehensive model, but still one where we eschew the temptation to include "bad controls". We also implement Oster (2019) tests of the robustness of the results to potential confounders, using these headline estimates. We go on to explore other robustness checks.

Although we recognise the difficulty of creating a credible identification to overcome the possibility of selection on unobservables our tests provide some reassurance. Nonetheless, we go to include some tentative IV evidence based on using cross-reports (from the main parent) as instruments for the own reports. This is in the spirit of using cross reports in twins studies of returns to education. Finally, report results where we disaggregate our definition to explore the effects of multiple treatments – violent vs non-violent forms of bullying, and different intensities of bullying.

6.1 Headline estimates

Table 5 shows the OLS results for the "Any bullying" measure, which is typically what the existing literature has measured. We report results only for boys and girls pooled (with a gender control included)[20]. The most straightforward specification of the treatment, and the most common in the literature, is the single treatment defined as "Any bullying". OLS results are reported in Table 5 for short term effects on having 5+ GCSE passes at age 16, taking A-levels at age 18, and A-level score which is used as one factor determining university admission; intermediate outcomes associated with university (having a degree by age 25); and long run outcomes at age 25 (log income, being unemployed, and the GHQ score).

Model 1 includes as covariates the child's gender, ethnicity, month of birth, Government Office Region (GOR) and English being a second language, along with the school fixed effects. Adjusting for these basic controls, we observe large detrimental effects of experiencing bullying. The probability of gaining 5+ GCSE passes at age 16 is reduced by 6.3 percentage points (10% reduction from a mean of 0.69). The probability of staying on in school to take a A-levels or an equivalent is reduced by 4.6 percentage points (9.0% reduction from a mean of .51), and the UCAS points gained from those qualifications are reduced by about 5 points (5% of a standard deviation). Turning to longer run outcomes, income at age 25 years is reduced by 2.3% (£7 per week reduction from a mean income of 303.4 in the sample). The probability of being unemployed increases by 3.5 percentage points (35% from a mean of 0.10). Perhaps most strikingly, the GHQ mental ill-health index increases by 0.97, a large effect size of about one third of a standard deviation. Evidently, being subject to any bullying, *within* schools (i.e. with school fixed effects) and controlling for a basic set of covariates, is strongly associated with deleterious outcomes.

However, these effects may be driven, to some extent, by confounding. Model 2 aims to address this by adding a rich set of relevant controls, associated with both being bullied and child outcomes. For the GCSE outcome, A-level participation, income at age 25 years, and university degree, this addition of relevant controls reduces the effects size by about half, and the effects remain statistically significant (aside from having a university degree). For example, the probability of gaining 5+ GCSE passes at age 16 is now 4% lower (and reduced to 2.4 percentage points in Model 2); the probability of staying on in school to take A-levels or equivalent is around -2.5% (reduced by 2 percentage points compared to Model 1); the points

[20] Results by gender are reported in Web Appendix Table WA4.

Table 5 *OLS estimates of the effects of "any bullying" with Oster diagnostics*

Dep Var:	5+ GCSE	Any A-levels	Best 3 A-level points	Ln (income)	University degree	Not employed	Mental health
Model 1							
β	-0.063***	-0.046***	-4.927	-0.023***	-0.023*	0.035***	0.969***
se	(0.012)	(0.011)	(3.412)	(0.004)	(0.012)	(0.007)	(0.075)
N	6,698	7,569	4,018	7,569	7,569	7,569	7,234
δ	39.07	-5.36	783.10	-4.46	10.12	104.20	66.60
Model 2							
β	-0.035***	-0.025**	-5.880*	-0.010***	-0.011	0.028***	0.911***
se	(0.010)	(0.012)	(3.509)	(0.003)	(0.013)	(0.007)	(0.082)
N	6,133	6,413	3,671	6,413	6,413	6,413	6,162
δ	3.47	2.75	673.70	-36.59	1.68	10.25	11.99

Notes: Robust standard errors, clustered by school, in parentheses. *** p<0.01, ** p<0.05, * p<0. School fixed effects are included in all specifications. β = coefficient on bullying treatment; se(β) robust standard error of β; δ is the estimate of delta parameter implemented in —psacalc- and developed in Oster (2019), which indicates how much selection on unobserved variables would be required to drive the beta estimate to zero, measured as a proportion to the selection on observed variables. Results by gender are reported in Web Appendix Table WA4.

Table 6: *OLS estimates of the effects of 'bullying factor' with Oster diagnostics*

Dep Var:	5+ GCSE	Any A-levels	Best 3 A-level points	Ln (income)	University degree	Not employed	Mental health
Model 1							
β	-0.049***	-0.046***	-1.780	-0.018***	-0.027***	0.020***	0.296***
se(β)	(0.008)	(0.007)	(2.930)	(0.003)	(0.006)	(0.006)	(0.049)
N	4,861	4,890	3,053	4,890	4,890	4,890	4,704
δ	26.75	22.60	4.83	-67.67	26.92	-24.51	-18.46
Model 2							
β	-0.014**	-0.023***	-1.203	-0.007**	-0.012*	0.011*	0.297***
se(β)	(0.007)	(0.008)	(2.947)	(0.003)	(0.007)	(0.006)	(0.053)
N	4,450	4,464	2,780	4,464	4,464	4,464	4,307
δ	1.14	2.51	1.14	1.86	1.72	6.51	-49.78

Notes: Robust standard errors, clustered by school, in parentheses. *** p<0.01, ** p<0.05, * p<0. β = coefficient on bullying treatment; se(β) robust standard error of β. δ is the estimate, from Oster (2019), of the ratio of selection on unobservables relative to observables that would be required to drive β to zero. Results by gender are reported in Web Appendix Table WA4.

gained from those qualifications is reduced by about -6 points (6% of a SD) in Model 2; the probability of having a university degree is now reduced to a 1 percentage point fall (although not significantly so); and the effect on income at age 25 years is now to reduce this by 1.0% (a £3 per week reduction from a mean of £303.4 in the sample); and, finally, the GHQ mental ill-health index marginally larger at 0.91, a robustly large effect size of 29% of an SD. However, there may still remain some selection on unobservables, which we explore by reporting the δ parameter proposed in Oster (2019). The estimates of the δ parameter in would be regarded as consistent with an 'acceptable' level of selection according to the rule-of-thumb suggested in Oster (2019). The only exception is log income which has a *negative* value of δ associated with it. A negative estimate of delta can be generated if the observables are positively correlated with the treatment, and the unobservables are negatively correlated with the treatment. The key 'rule of thumb' arguments in Oster (2019) are based on these correlations both being positive, so we do not draw conclusions based on this particular negative estimate.

Table 6 shows the same analysis for the case where bullying is recorded as a continuous variable from a factor analysis exercise. There is a very similar pattern of results relative to Table 5, although the coefficients are not comparable because Table 6 is based on a continuous measure of bullying while Table 5 is a simple dummy variable. Model 2 generally has smaller coefficients than model 1in both tables, as we might expect. The Oster bound generally falls as we move from model 1 to model 2. Yet, in most cases, The results in both tables suggest that most of the model 2 estimates are unlikely to be zero under reasonable assumptions about possible confounders - Oster suggests that if $\delta>1$ it would be reasonable to conclude that it would be unreasonable to believe that the results are driven selection.

6.2 Robustness checks

Table 7 presents results from several falsification tests. We show OLS estimates of the effects of being bullied on various outcomes, which we feel should *not* be affected by the bullying treatment: historical information on school performance, and two individual level outcomes - absolute deviation from average peer height, and whether the individual took KS2 Maths and Science extension tests[21]. Observing an effect on these outcomes would suggest we are

[21] The extension paper was introduced in 2001/2 to measure and stretch the most mathematically able olds (scoring over 90% in the main test). It was then dropped, but was later reintroduced.

conflating the bullying effects on long run outcomes with general omitted variable bias. Conditioning on the variables listed in Model 2, we do not observe any significant effects on these outcomes. This provides further support for the credibility of our results and with the success of Model 2 in controlling for the key determinants of bullying and outcomes.

Aside from the issues associated with identification, the estimates could also be driven by the functional form imposed in the OLS estimation. Therefore, we also investigate propensity score matching. In Table 8, the propensity score findings show a similar pattern to those in Model 1 of Table 5, suggesting that the Table 5 results are not driven by the functional form of the OLS model.

The first row of estimates in Table 8 is copied from the first column of Table 7's PSM estimates. We first consider, in the second row, the plausibility of a binary confounder that would drive our treatment effects to zero Taking the first outcome in Table 7 as an example, gaining 5+ GCSE passes, we see there would need to be large outcome and selection effects to make this effect completely disappear. The binary confounder *U* would need to *increase* the odds of being bullied by a factor of at least 4.5 and *decrease* the odds of gaining 5+ GCSE passes by a factor of 0.2. While this may be plausible, looking across the outcomes it seems that the longer run effects are most robust to selection (and so would require the most extreme confounder). For instance, for mental health, the binary confounder *U* would need to *increase* the odds of being bullied by a factor of at least 9 and *decrease* the odds of being in the top quartile of the GHQ distribution by a factor of 31. This type of extreme confounder seems an unlikely scenario.

To assess more realistic potential confounders, we evaluate the effects of simulated variables that mimic the distribution of relevant observed variables in our data, in relation to the treatment and outcome. The next three rows in Table 8 assess the effects of adding each of our selected simulated variables to be potential confounders in our data: being in a sole parent family, having English as a second language, and being in the top decile or bottom decile of the Key Stage 2 distribution on the child's school. These variables were chosen as variables that may reflect perceived or actual differences from one's classmates, which would shape both the propensity of being bullied <u>and</u> have direct effect on the outcome. While we adjust for school fixed effects in much our analyses, so that the data are in deviations from the school averages, there may be further unobserved confounders based on "being different" which are not captured.

Table 7 OLS linear estimates of "any bullying" (a binary treatment variable) on predetermined variables (falsification tests)

Outcome:	% pupils 5+ GCSEs in 2001	Average KS 2 points in 2001	Absolute deviation from mean height at 25	Took KS2 Math /Science extension tests
	-0.179	0.989	0.0013	-0.0034
se(β)	(0.459)	(1.140)	(0.0017)	(0.0062)
N	6,260	5,702	6,478	6,731

Notes: This tables shows the OLS (linear regression) estimates of four outcomes, which are determined before the bullying variable is measured, on the binary variable "any bullying?". Standard errors are in parentheses. The control variables used in the regressions are from Model 2. For the school-constant outcomes (e.g., % white pupils), school fixed effects are omitted from the specification.

Table 8: Propensity score matching estimates of the effects of "Any bullying" (a binary treatment variable)

	ATT	Std. error	N (control)	N (treated)	Total N
5+ GCSE	-0.071***	0.010	2,401	3,732	6,133
Any A-levels?	-0.053***	0.011	2,526	3,887	6,413
Best 3 A-level points	-7.460**	2.757	1,549	2,122	3,671
ln(Income)	-0.017***	0.005	2,526	3,887	6,413
Has a degree	-0.022**	0.012	2,526	3,887	6,413
Unemployed	0.035***	0.007	2,526	3,887	6,413
Mental health	0.960***	0.080	2,417	3,745	6,162

Notes: Kernel matching estimation is implemented using `attk` in Stata; ATT= average treatment effect on the treated; se, standard error (bootstrapped with 100 replications). *** p<0.01, ** p<0.05, * p<0.1. The covariates included in the propensity score model are from Model 2. The PSM common support graph is available in Web Appendix Figure WA1 and the propensity score graphs are available in Web Appendix Figure WA2. Multivariate distance and Nearest Neighbour matching results are available in Web Appendix Tables WA2 and WA2.

Table 9: Sensitivity analysis for PSM estimates of the effects of "Any bullying" (a binary treatment variable).

Outcomes:	5+ GCSEs	Any A levels	Best 3 Alevel points	Ln Income	Not employed	Mental health
ATT assuming unconfoundedness	-0.071	-0.053	-7.470	-0.017	0.035	0.96
With *U* chosen so ATT ≈ 0						
Selection effect	4.460	2.942	3.541	1.714	3.109	9.015
Outcome effect	0.211	0.314	0.335	0.089	12.86	31.12
ATT, *U* mimicking "Sole parent family"	-0.069	-0.051	-7.433	-0.016	0.035	0.957
ATT, *U* mimicking "English second language"	-0.070	-0.053	-7.383	-0.018	0.036	0.953
ATT, *U* mimicking "Outlier in school KS2 dist^n"	-0.070	-0.053	-7.493	-0.017	0.035	0.959
N	6,133	6,413	3,671	6,413	6,413	6,162

Notes: Kernel matching estimates, ATT = average treatment effect on the treated.

Beginning with effects of the simulated unobserved confounder mimicking the distribution of the sole parent family variable, this would reduce the effect on the GCSE variable by only about 3% i.e. (comparing row 1 with the equivalent row in column 1 of Table 9, -0.071 to 0.069). Adding the simulated variable to the A-level outcomes determinants the bullying estimates is reduced by only 4% and in the ln(income) equation by about 6%, but it has negligible effects on the other treatment effects. The simulated unobserved confounder mimicking the distribution of ESL again has little impact on the estimated treatment effects, aside from reducing the effect of on mental health by 1%.

Finally, we examine the simulated unobserved confounder that mimicks the distribution of being in the tails of the prior ability distribution. This would reduce the effect on the GCSE variable by about 2% and has negligible effects on the other treatment effects. Our conclusion from this analysis is that, overall, scenarios emulating realistic levels of confounding could reduce our treatment effects by between 0%-6%, depending on which outcomes is considered. Therefore, it seems very unlikely that our results could be entirely, or even predominantly, driven by selection. The type of confounder required for this to happen appears to be substantively implausible.

6.3 Measurement error in subjective self-reports

Our results above suggest, at least tentatively, that selection on unobservables is not likely to be sufficiently important to drive our results to zero. Unfortunately, there appears to be no clear identification strategy that might confirm this, at least for the UK, so this remains an open question – and one that has not, with one exception, been satisfactorily addressed in the existing literature.

However, there is an additional source of bias that might also represent a threat – and this is associated with the subjective nature of the self-reported bullying information. This is long familiar from the literature on the effects of health, for example on labour supply in Stern (1989) where self-reports show smaller effects than objective reports. This is consistent with attenuation associated with a measurement error interpretation of subjective measures.[22] We report the results of using IV estimation to resolve attenuation in Appendix Table A1, with a continuous bullying treatment and a similar continuous

[22] See Bingley and Martinello (2017) for an analysis of the use of potentially mis-measured cross-reports in a different context. Light and Flores-Lagunes (2006) suggest that IV based on the presumption of classical measurement error provide a good approximation to more complex measurement error models.

parental cross report that is based on the same factor analysis. Unsurprisingly, the first stage coefficients show that there is a strong correlation between the self and cross reports.[23] The F statistic supports the validity of the instrument since its value exceeds the rule-of-thumb value of 15 (except for the A-level points case). As expected, the effect sizes tend to be larger from IV estimation since, here, we are attempting to correct for attenuation. In fact, they are at least double the corresponding OLS estimates suggesting that this is a serious problem, and it is one that has not been previously addressed in this context. In general, the IV results confirm our OLS findings and show substantial and adverse impacts of bullying on all outcomes. Although the IV estimates are, as usual, less well-determined that their OLS counterparts the long-run adverse effects on mental health, degree, and income are now statistically significant and at least double the OLS estimates, while the IV estimated short run outcomes are not significant, in contrast to their OLS equivalents.

6.4 Multiple treatments

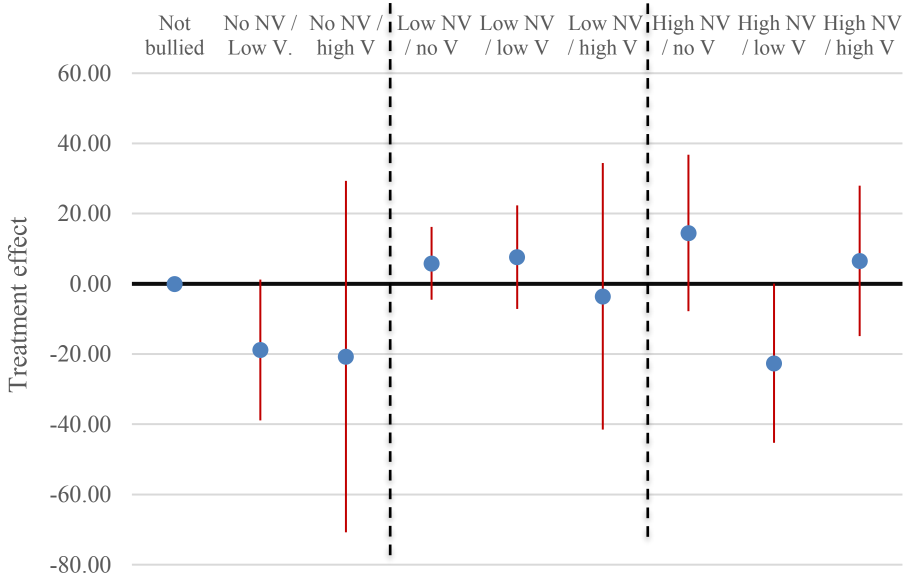
Finally, we explore the role of type and frequency together using treatment IPWRA estimation of multiple treatments. The aim is to show the merit of viewing bullying as a multi-valued treatment problem. Figure 4 summarises the estimates (available in Web Appendix Table WA5) for the four long run outcomes: (a) university degree, (b) income, (c) unemployed, (d) mental health. The dots are point estimates, while the vertical lines represent 95% confidence intervals. Our results, for each outcome variable, are grouped into three groups, where we look at the effect of increasing violence for a given level of non-violent bullying (grouped by the dashed lines). Thus, in Figure 4a we see that with, no violent bullying, increasing the intensity of violent bullying decreases the probably of having a degree. Similarly, in Figure 4d we find that there is (almost) a monotonically increasing effect of violent bullying at any level of non-violent bullying; **and**, looking across groups, we see that as the non-violent bullying level rises this (almost) monotonic pattern of increasing violent bullying across the groups of non-violent levels increases successively. While, most of these individual interacted treatment effects are not individually statistically significant the pattern of results suggests that the interactions of more bulling of each type conditional on the level of the other type is generally adverse.

The estimates for the short-term outcomes are presented in Figure 5. We again visualise the estimates as interactions of more serious bullying and divide these into three

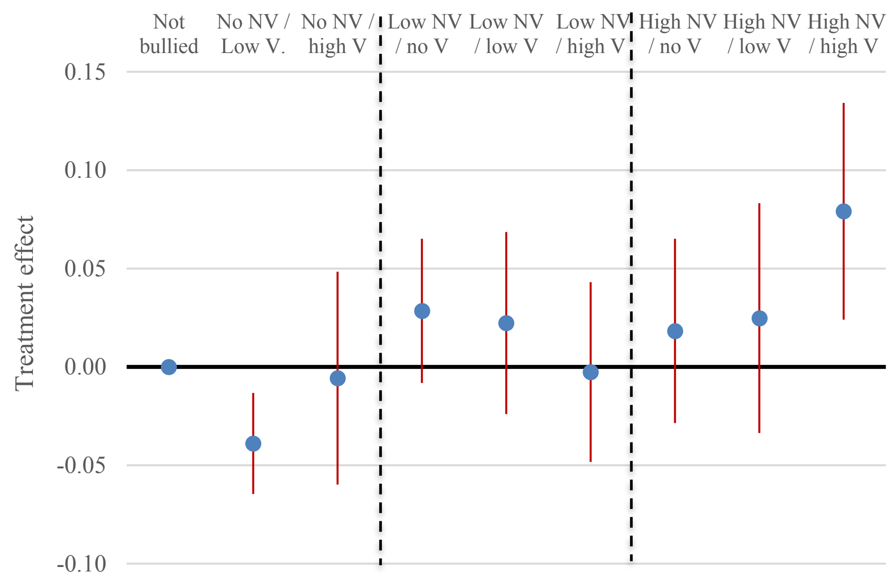
[23] Chrystanthou and Vasilakis (2018) use alternative UK data to suggest that victim non-discolsure is important.

Figure 4 ***Estimated long term effects from IPWRA model***

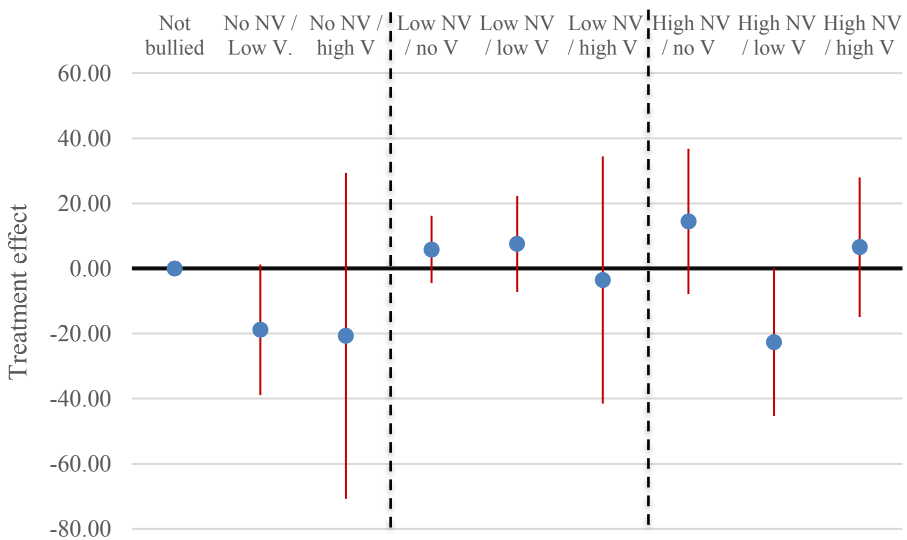
(a) University degree



(c) Unemployed



© Income



(d) Mental health

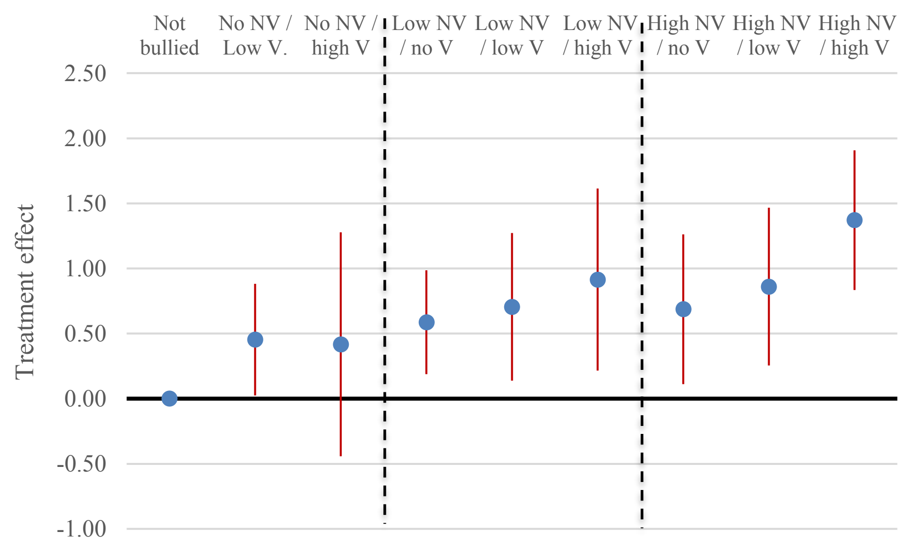
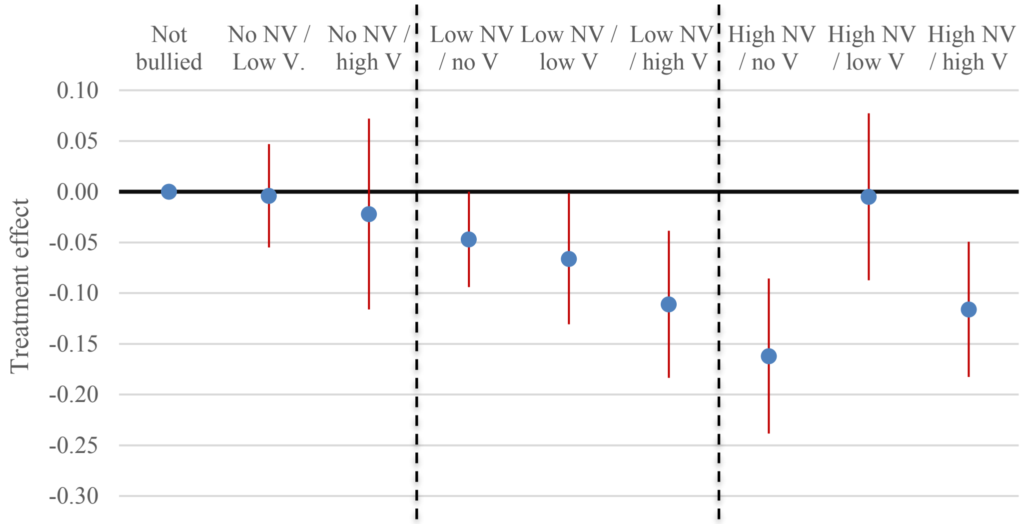
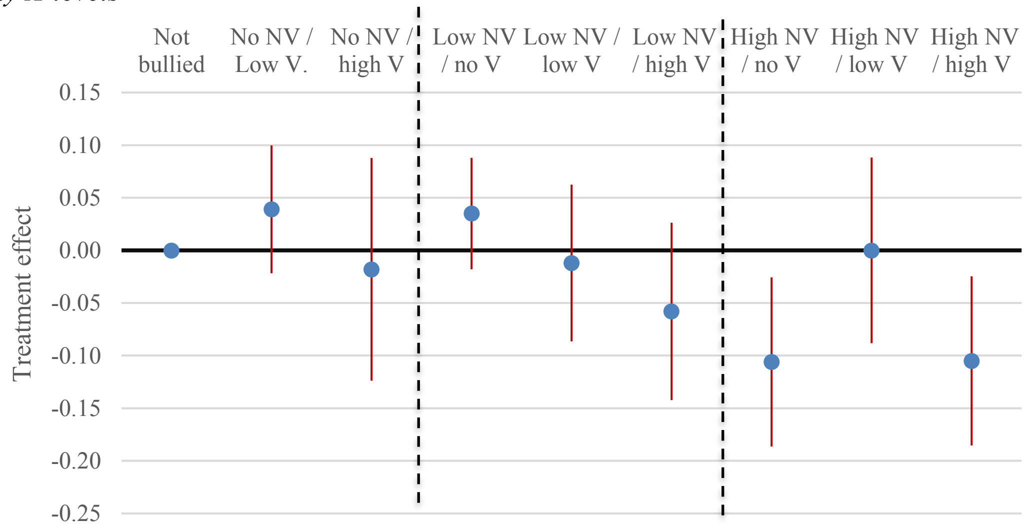


Figure 5 ***Estimated short term effects from IPWRA model***

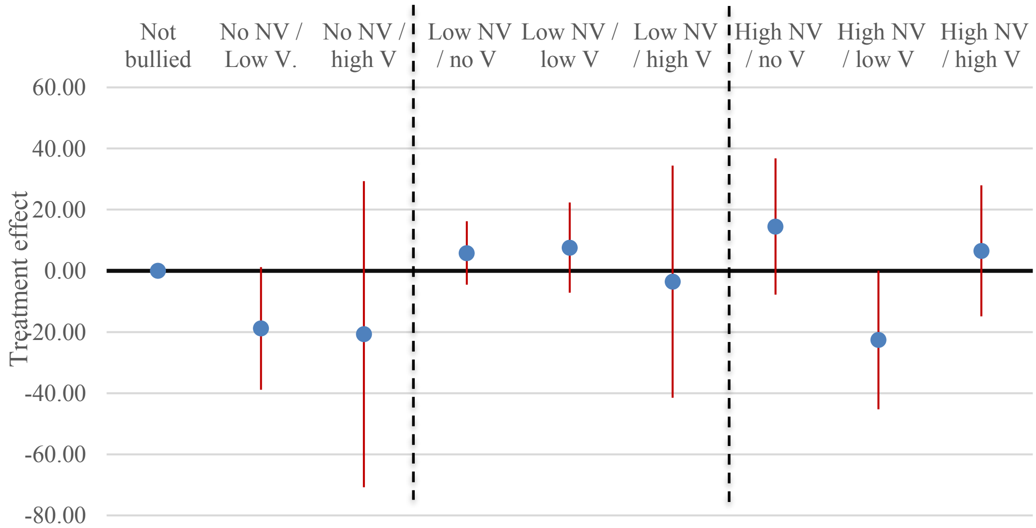
(a) 5+ GCSE



(b) Any A-levels



(b) Best 3 A-level points



groups of successively higher levels of seriousness. Even with this minimal extension that considers just two types of bullying at three levels of intensity (none, low, high), we find systematic effects of both type and frequency using IPWRA. Especially for the longer run outcomes, it appears that much of the effects is driven through the most intense forms of bullying—high intensity, violent bullying. Other types and frequencies also have effects, especially for mental health where any combination of NV and V bullying, whether at high or low intensity, statistically significant adverse effects – raising the mental ill-health count by between 0.5 and 1.5 where the mean is 2.3. The effects on income are large and negative (-4%) only for the relatively small proportion of the population who experience high intensity V bullying and either high or low NV bullying.

The results in Figure 5 suggest that these effects on income in Figure 4 may stem, in part, from negative impacts of bullying combinations on the probability of attaining 5+ GCSEs or any A level. These results strongly reject the idea that a single treatment is sufficient to capture the complex effects of bullying.

7. Conclusion

This paper investigates the effects of bullying in secondary school on later academic and labour market outcomes. We do this by exploiting a rich conditioning set of observables, and using a range of estimation methods: OLS, matching and weighting. The data come from a large high-quality cohort study in England, LSYPE, linked with administrative data on education records. Our empirical findings show that school bullying has negative consequences for short run academic outcomes and persists to have adverse *long-term* effects—the strongest effects are on mental health, and we also find effects on unemployment and income measured at 25 years.

We conduct a comprehensive battery of sensitivity tests to explore our main identifying and estimation assumptions. The results of this indicate that it is unlikely that our effects are entirely driven by selection on unobserved variables. A cautious interpretation of the results is that any of our effect sizes could potentially be reduced*,* but not eliminated*,* by unobserved selection. Even in this scenario, the estimate effects remain large enough to be of substantive importance. The most robust effects are on mental ill-health and unemployment. Being bullied exerts long run adverse effects of children's life outcomes. Based on our analyses, we feel confident that this finding is not an artefact of a particular estimation or identification assumption. If we take the mental health effects alone the costs associated with such an increase would be important enough to justify greater effort in reducing bullying. The results have relevance for policy. Although schools have flexibility in how they deal with bullying, all schools are expected to have a policy. In practice, schools tend to take a zero-tolerance approach to bullying. Our results suggest that low levels of non-violent bullying have modest effects, but higher intensity bullying has much larger effects. We also tentatively suggest that violent bullying has a greater effect than non-violent bullying. These findings suggest that the long run consequences of bullying should not be underestimated, and perhaps policy should be targeted more heavily on the extreme cases of violent and persistent cases.

There are several important ways that this research might, data permitting, be extended to broaden the reach of policy relevance. First, we do not analyse cyber-bullying. The children in the LSYPE data used here do not report explicitly on cyber-bullying. However, child and main parent reports of cyber-bullying are reported in LSYPE2 that was collected for a cohort ten years later than LSYPE, when smartphone use became more prevalent among young people. LSYPE2 has not yet been matched to the administrative data on the KS4 high-stakes educational outcomes. Moreover, there has not yet been a long-run age 25 follow-up to LSYPE2, although one is planned. Secondly, it would be useful to explore workplace bullying and its effects. The age 25 follow up does contain contemporaneous bullying information and it shows a high correlation with bullying at age14 to 16. However, while the age 25 follow-up wave were consented to have subsequent administrative data merged into the data, this has not yet been done. However, the NPD, that provides the school educational outcomes in the data here, has very recently been extended by merging data on pupil's experiences in Further and Higher Education, and data from income taxation records (up to age 39 so far). This database, known as Longitudinal Education Outcomes (LEO), offers the prospect of being extended to wider administrative records that would further extend the scope of future analysis.

Finally, the analysis relies on the Oster (2019) test for bias associated with selection on unobservables. Although these are reassuring findings it would be useful to find confirmation from direct estimates of causal treatment effects. However, it seems unlikely that this would be possible in the absence of school-reported bullying at the individual level in the NPD, and without there being any convincing natural experiment.

REFERENCES

Altonji, J.G., T.E. Elder and C.R. Taber (2005), "Selection on Observed and Unobserved Variables", **Journal of Political Economy***,* 113, 51-184.

Ammermueller, A. (2012), "Violence in European schools: A widespread phenomenon that matters for educational production", **Labour Economics**, 19(6), 908-922.

Anders, J. (2012), "Using the Longitudinal Study of Young People in England for research into Higher Education access," **DoQSS Working Papers 12-13**, Department of Quantitative Social Science , Institute of Education, University College London.

Arseneault, L., Bowes, L., and S. Shakoor, S. (2010). "Bullying victimization in youths and mental health problems: 'Much ado about nothing'? **Psychological Medicine**, 40, 717-729

Ashenfelter, O. and C. Rouse (1998), "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins", **Quarterly Journal of Economics**, 113, 253-284.

Bingley, P. and A. Martinello (2017), "Measurement Error in Income and Schooling and the Bias of Linear Estimators", **Journal of Labor Economics,** 35,1117-1148.

Brown, S., and Taylor, K. (2008) Bullying, education and earnings: Evidence from National Child Development Study, **Economics of Education Review**, 27, 387-401.

Bond, L., Carlin, J., Thomson, L. Rubin, K., and Patton, G. (2001), "Does bullying cause emotional problems? A prospective study of young teenagers", **British Medical Journal**, 323, 480-484.

Borra, C., Iacovou, M., & Sevilla, A. (2012), The effect of breastfeeding on children's cognitive and noncognitive development. **Labour Economics**, 19, 496-515.

Brown, K. (2018), "Bullying: A Review of the Evidence", **Education Policy Institute** blog post, https://epi.org.uk/publications-and-research/bullying-a-review-of-the-evidence/.

Burgess, S., Graves, E., Vignoles, A., and D. Wilson (2015), "What Parents Want: School Preference and School Choice", **Economic Journal**, 124, 1262-1289

Carrell, S., and M. Hoekstra (2010), 'Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids', **American Economic Journal: Applied Economics**, 2(1): 211-228.

Cattaneo, M. (2010), "Efficient semiparametric estimation of multi-valued treatment effects under ignorability", **Journal of Econometrics**, 155, 138-154.

Centre for Longitudinal Studies. (2018). **Next Steps: Sweeps 1-8, 2004-2016, 14th Edition**, Study Number: 5545, UK Data Service. SN: 7104, http://doi.org/10.5255/UKDA-SN-7104-4

Chrystanthou, G.M. and C. Vasilakis (2018), "The Dynamics and Determinants of Bullying Victimisation", **IZA DP** 11092.

Delprato, M., Akyeampong, K. and M. Dunne (2017), 'The impact of bullying on students' learning in Latin America: A matching approach for 15 countries', **International Journal of Educational Development**, 52: 37-57.

Department of Education (2010), "A Review of the Longitudinal Study of Young People in England (LSYPE): Recommendations for a Second Cohort", **Research Report 048**.

Department of Education (2017), **Preventing and Tackling Bullying, Advice for Headteachers, Staff and Government Bodies**, HMSO London.

Department for Communities and Local Government (2015) **The English Index of Multiple Deprivation (IMD) 2015—Guidance**, HMSO, London.

Due, P., B.E. Holstein, J. Lynch, F. Diderichsen, S. N. Gabhain, P. Scheidt, C. Currie, and the Health Behaviour in School-Aged Children Bullying Working Group (2005), "Bullying and symptoms among school-aged children: International comparative cross-sectional study in 28 countries", **European Journal of Public Health**, 15, 128-132.

Eriksen, T., Nielsen, H., and Simonsen, M. (2014), "Bullying in Elementary School", **Journal of Human Resources**, 49 (4), 839-871.

Farrington, D., and M. Ttofi, (2009) School-based programs to reduce bullying and victimization, **Campbell Systematic Review**, 2009.

Fiorini, M., and Keane, P. (2014), "How the allocation of children's time affects cognitive and non-cognitive development?", **Journal of Labour Economics**, 32(4), 787-836.

Ford, R., T. King, N. Priest, and A. Kavanagh (2017), "Bullying and mental health and suicidal behaviour among 14-to 15-year-olds in a representative sample of Australian children", **Australian and New Zealand Journal of Psychiatry**, 51 (9), 897-908.

Harmon, C. (2017), "How Effective is Compulsory Schooling as a Policy Instrument?", **IZA World of Labor**, 348

Harmon, C. and I. Walker (2000), "The Return to the Quantity and Quality of Education: Evidence for Men in England and Wales", **Economica**, 67, 19-35

Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? **Journal of Applied Econometrics**, *23*(3), 305-327.

Imbens, G., and Wooldridge, J. (2009), "Recent Developments in the Econometrics of Program evaluation" **Journal of Economic Literature**, 47(1), 5-86.

Krauth, B. (2016), "Bounding a linear causal model using relative correlation restrictions", **Journal of Econometric Methods**, 5, 117-141.

Ladd, G., I. Ettekal, B., Kochenderfel-Ladd (2017), "Peer Victimization trajectories from kindergarten through high school: Different pathways for children's school engagement and achievement?" **Journal of Educational Psychology**, 109, 826-841.

Lam, J. (2016), "Bullying and human capital accumulation: Evidence from the National Longitudinal Survey of Youth, 1997", **CUNY Academic Works**, http://academicworks.cuny.edu/hc_sas_etds/38.

Light, A. and Flores-Lagunes, A. (2006), "Measurement Error in Schooling: Evidence from Samples of Siblings and Identical Twins", **BE Journal of Economic Analysis & Policy**.

Mendolia, S., and Walker, I. (2015), "Youth unemployment and personality traits", **IZA Journal of Labour Economics**, 4(19), 1-26.

Murphy, R., Scott-Clayton, J. and G. Wyness (2017), "The End of Free College in England: Implications for Quality, Enrolments and Equity". **NBER Working Paper** 23888.

Nannicini, T. (2007), "Simulation-Based Sensitivity Analysis for Matching Estimators," **Stata Journal**, 7, 334-350.

OECD (2017): "How Much of a Problem is Bullying at School?" **PISA in Focus No. 74**, OECD Publishing, Paris.

Oliveira, F.P., T. Almeida de Menezes and G. Irffi (2018), “Bullying effect on student’s performance”, **Economia,** 19, 57-73.

Olweus, D. (1993), **Bullying in school: What we know and what we can do**. Oxford: Blackwell.

Olweus, D. (2013), “School bullying: Developments and some important challenges”, **Annual Review of Clinical Psychology**, 9, 751-780.

Oster, E. (2019), “Unobservable Selection and Coefficient Stability: Theory and Evidence”, **Journal of Business & Economic Statistics**, 37:2, 187-204.

Ponzo M. (2013), “Does bullying reduce educational achievement? An evaluation using matching estimators”, **Journal of Policy Modelling**. 1057-78.

Ren D. and J.G Voelkel (2017), “Bullying”, in Zeigler-Hill V. and T. Shackelford (eds) **Encyclopedia of Personality and Individual Differences**, Springer.

Sarzosa, M. and Urzua, S. (2015), “Bullying among adolescents: the role of cognitive and non-cognitive skills”, NBER Working Paper 21631.

Sharp, S. (1995), “How much does bullying hurt? The effects of bullying on the personal wellbeing and educational progress of secondary aged students”, **Educational and Child Psychology**, 12(2), 81-88.

Smith, P.K. and F. Thompson (2014), “What works best to help stop bullying in schools?”, **The Conversation**, August 7.

STATA Corp (2017), **Stata Statistical Software: Release 15**, College Station, TX: StataCorp LLC.

Steven, S. (1989), "Measuring the Effect of Disability on Labor Force Participation." **Journal of Human Resources** 24, 361-395.

Thompson, F. and P. Smith (2010), “The Use of Effectiveness of Anti-Bullying Strategies in Schools”, Research Report DFE-RR098.

Ttofi, M. and D. Farrington (2011), “Effectiveness of school-based programs to reduce bullying: a systematic and meta analytic review”, **Journal of Experimental Criminology**, 7, 27-56.

Vignoles, A. and Meschi, E. (2010), “The determinants of non-cognitive and cognitive school outcomes”, **Centre for the Economics of Education**, Special Report 004 for the Department of Children, Schools and Families

Woods, S. and Wolke, D. (2004) “Direct and relational bullying among primary school children and academic achievement”, **Journal of School Psychology**, 42, 135-155.

Wooldridge, J.M. (2007). “Inverse probability weighted estimation for general missing data problems”, **Journal of Econometrics** 141: 1281–1301

Wooldridge, J.M. (2010), **Econometric Analysis of Cross Section and Panel Data,** 2nd ed. MIT Press, Cambridge, MA.

APPENDIX

Questions in *LSYPE*

Locus of control

I can pretty much decide what happens in my life
If someone is not a success in life, it is usually his fault
How well you get in this world is mostly a matter of luck
Even if I do well at school, I will have a hard time
People like me do not have much of a chance
If you work hard at something, you will usually succeed

Possible answers: Strongly agree, Agree, Disagree, Strongly disagree

Work ethic

Doing well at school means a lot to me
At school, I work as hard as I can
Working hard at school now will help me to get on later in life
If you work hard at something, you will usually succeed

Possible answers: Strongly agree, Agree, Disagree, Strongly disagree

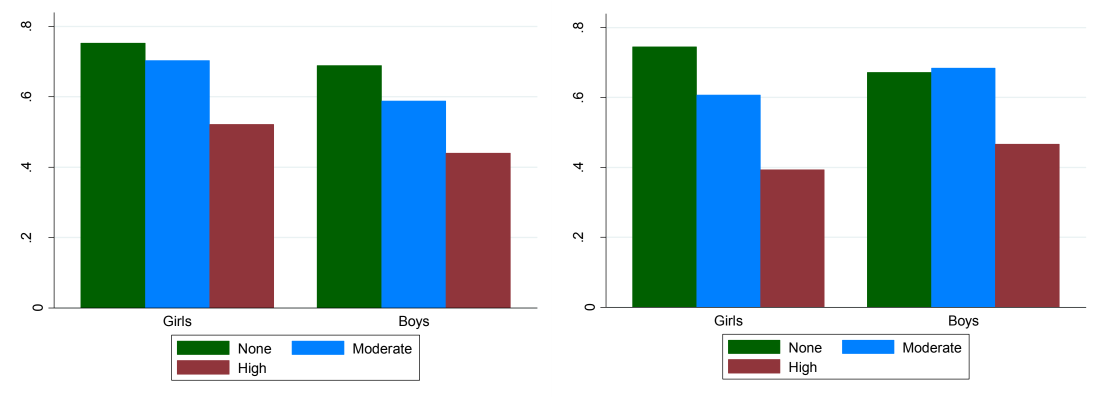
Appendix Table A1: IV estimates of the effects of continuous bullying treatment

Dependent Variable:	5+ GCSE	Any A-levels	Best 3 A-level points	Ln income	University degree	Not employed	Mental health
β	-0.029	-0.050*	-20.20	-0.024***	-0.056***	0.0287	0.775***
(se)	(0.020)	(0.026)	(14.63)	(0.009)	(0.021)	(0.0192)	(0.250)
1st-stage coeff	0.704***	0.592***	0.492***	0.592***	0.592***	0.592***	0.609***
(se)	(0.146)	(0.122)	(0.190)	(0.122)	(0.122)	(0.122)	(0.128)
F-stat	23.35	23.47	6.73	23.47	23.47	23.47	22.61
N	3,444	3,455	2,209	3,455	3,455	3,455	3,337

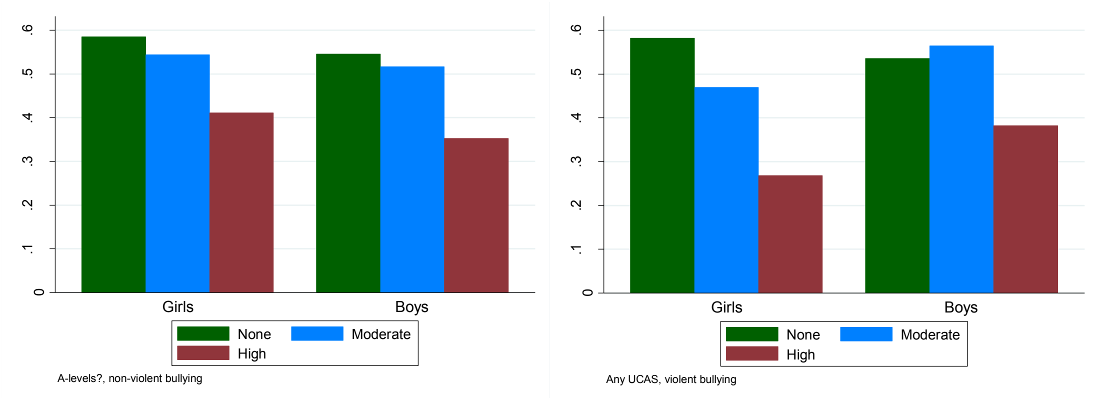
Notes: The first stage coefficient is that on the bullying factor derived from cross-reported bullying by the main parent. Robust standard errors, clustered by school, in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The covariates included, in both stages, are from Specification 2.

Appendix Figure A1 Outcome means by type of bullying and level of bullying.

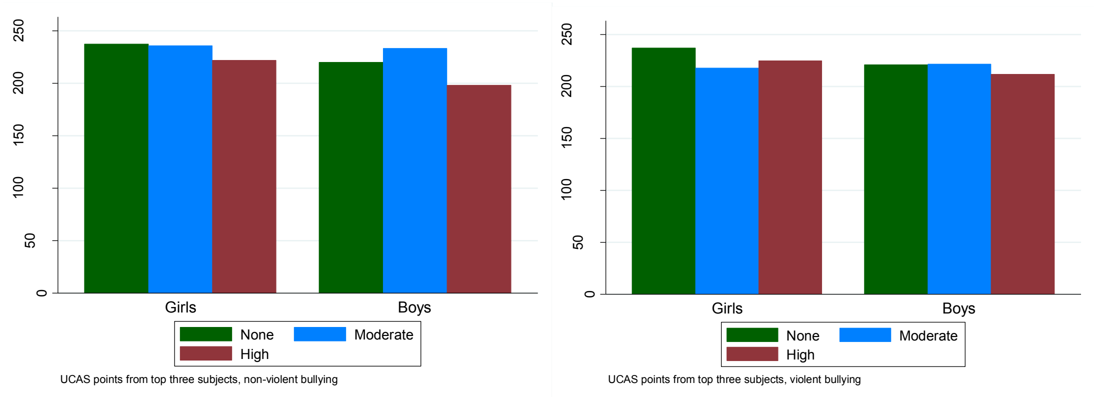
a) *5+ GCSEs - Non-violent and violent*



b) *Any A-levels - Non-violent and violent*

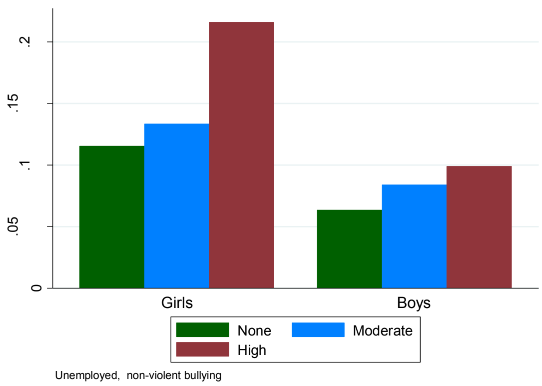


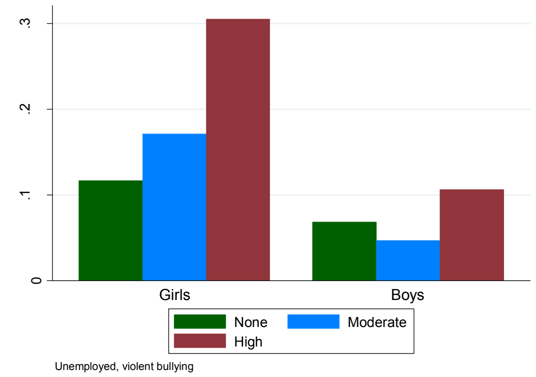
c) *Best 3 A-level points- Non-violent and violent*



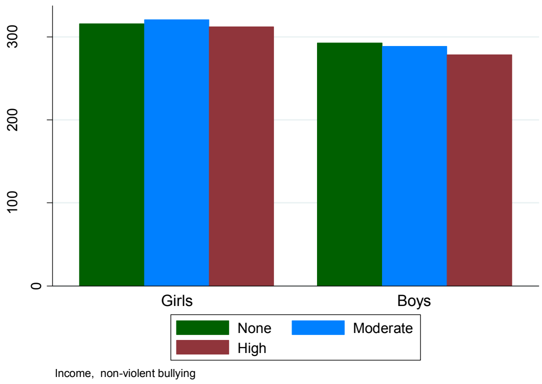
Appendix Figure A2 contd.   Outcomes by type of bullying and level of bullying.

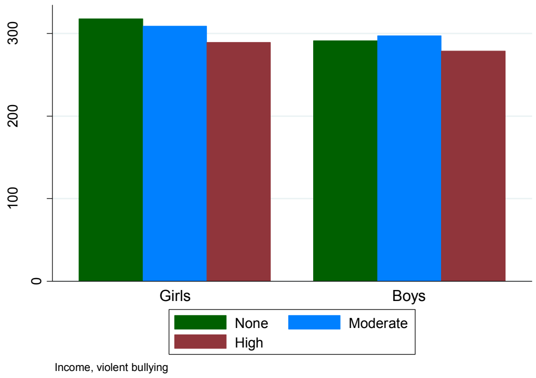
d) *Unemployed - Non-violent and violent*



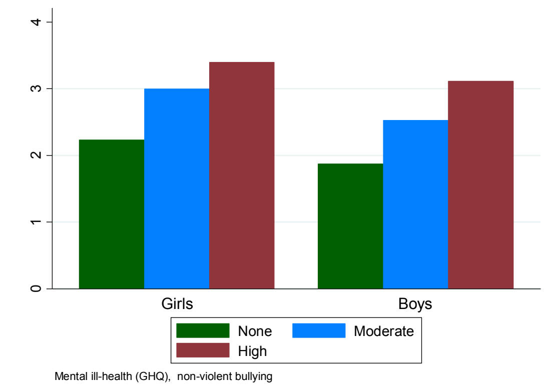


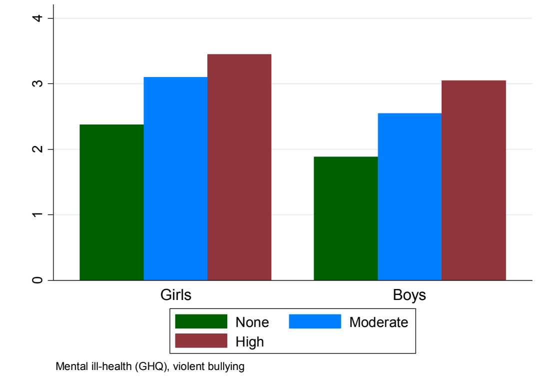
e) *Weekly income - Non-violent and violent*



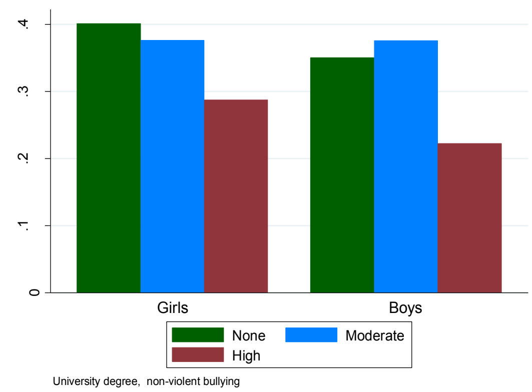


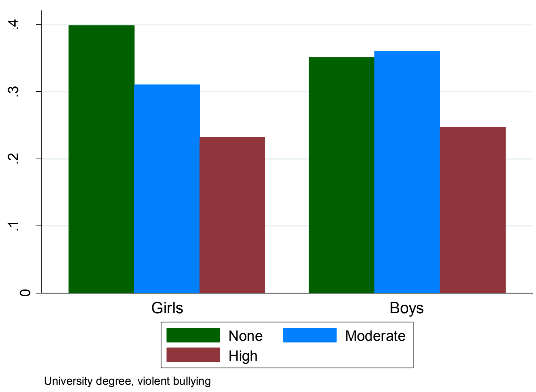
f) *Mental health - Non-violent and violent*





g) *University degree - Non-violent and violent*





WEB APPENDIX

Web Table WA1: Unweighted summary stats for bullying variables

	Mean	SD	Unweighted N
Parent report of type:			
Called names	0.30	0.46	6,885
Excluded from groups	0.11	0.31	6,885
Made to hand over money or items	0.01	0.12	6,885
Threatened with violence	0.12	0.32	6,885
Experienced violence	0.10	0.30	6,885
Child report of type:			
Called names	0.30	0.46	7,102
Excluded from groups	0.16	0.37	7,135
Made to hand over money or items	0.03	0.18	7,297
Threatened with violence	0.19	0.39	7,236
Experienced violence	0.16	0.37	7,250
Child report (factor)	-0.07	0.82	4,890
Parent report (factor)	-0.09	0.76	4,971

Web Table WA2: Multivariate distance matching

	ATT	std.err	N (control)	N (treated)	Total N
5+ GCSE	-0.0452	0.0127	2,401	3,732	6,133
Any A-levels?	-0.0322	0.0153	2,526	3,887	6,413
Best 3 A-level points	-9.3701	3.9647	1,549	2,122	3,671
ln(Income)	-0.0147	0.0048	2,526	3,887	6,413
Has a university degree	-0.0081	0.0157	2,526	3,887	6,413
Unemployed	0.0260	0.0088	2,526	3,887	6,413
Mental health	1.0121	0.0903	2,417	3,745	6,162

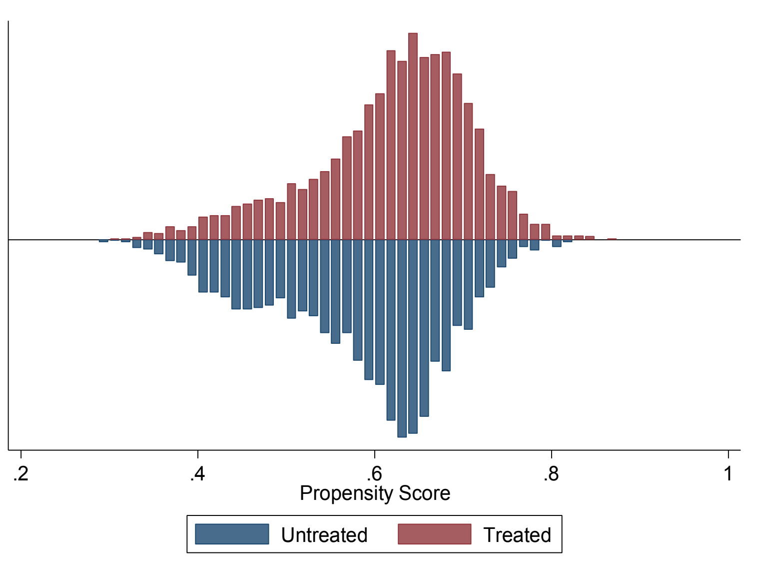
Notes: ATT: Average Treatment Effect on the Treated; Std. err: Robust Abadie/Imbens standard errors. The covariates included are from Model 2**.**

Web Table WA3 Nearest neighbour propensity score matching

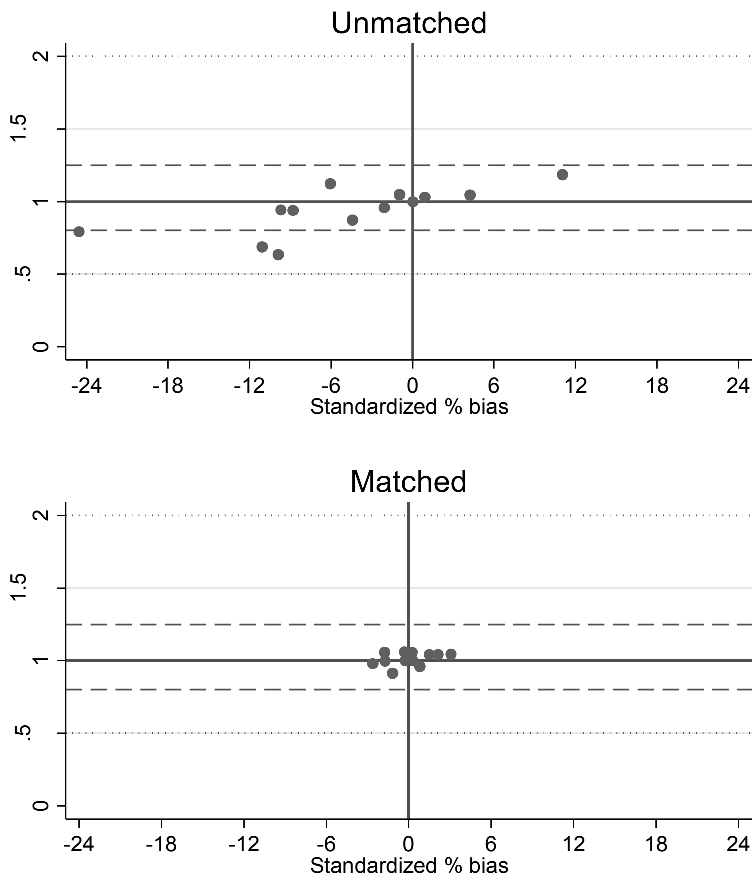
	ATT	std.err	N (control)	N (treated)	Total N
5+ GCSE	-0.0377	0.0104	2,401	3,732	6,133
Any A-levels?	-0.0379	0.0126	2,526	3,887	6,413
Best 3 A-level points	-8.3539	3.3439	1,549	2,122	3,671
ln(Income)	-0.0133	0.0046	2,526	3,887	6,413
Has a degree	-0.0102	0.0125	2,526	3,887	6,413
Unemployed	0.0285	0.0078	2,526	3,887	6,413
Mental health	0.9584	0.0818	2,417	3,745	6,162

Notes: ATT: Average Treatment Effect on the Treated; Std. err: Robust Abadie/Imbens standard errors; nn=5; caliper=0.15. The covariates included are from Model 2.

Web Figure WA1: Histogram showing common support and balance of the matched sample. All observations are on the common support.



Web Figure WA2: Plot summarizing the balance statistics comparing the unmatched and matched sample (from –psgraph-)



Web Table WA4: OLS estimates of the effects of bullying (binary and continuous) on boys vs girls.

	Dependent Variable:	5+ GCSE	Any A-levels	Best 3 A-level points	Ln(income)	University degree	Unemployed	Mental health
	Binary treatment							
Boys	B	-0.0355***	-0.0309*	-4.522	-0.00710	-0.0456***	0.0210*	1.019***
	se(β)	(0.0126)	(0.0163)	(4.970)	(0.00478)	(0.0174)	(0.0118)	(0.119)
	N	3,416	3,570	2,085	3,570	3,570	3,570	3,436
Girls	B	-0.0208	-0.0201	-6.808	-0.0108*	0.0313	0.0383***	0.660***
	se(β)	(0.0161)	(0.0183)	(5.830)	(0.00592)	(0.0201)	(0.00972)	(0.130)
	N	2,717	2,843	1,586	2,843	2,843	2,843	2,726
	Gender diff. p-value	*0.418*	*0.475*	*0.896*	*0.274*	*0.029*	*0.376*	*0.082*
	Continuous treatment							
Boys	B	-0.0235	-0.0165	-4.299	-0.00556	-0.0166*	0.0103	0.224***
	se(β)	(0.0143)	(0.0113)	(5.190)	(0.00419)	(0.00982)	(0.0115)	(0.0798)
	N	2,430	2,437	1,544	2,437	2,437	2,437	2,359
Girls	B	-0.00599	-0.0317***	-1.501	-0.00549	-0.0192*	0.00912	0.309***
	se(β)	(0.00857)	(0.0113)	(4.427)	(0.00351)	(0.0104)	(0.00632)	(0.0848)
	N	2,020	2,027	1,236	2,027	2,027	2,027	1,948
	Gender diff. p-value	*0.529*	*0.287*	*0.967*	*0.740*	*0.713*	*0.603*	*0.848*

Notes: Gender diff': the p-value from a test for differences in the effect between the boys and girls subgroup. Robust standard errors, clustered by school, in parentheses. *** p<0.01, ** p<0.05, * p<0.10. The covariates included are from Model 2

Web Table WA5: Full IPWRA results

Dependent Variable:	5+ GCSE	Any A-levels	Best 3 A-level points	Ln (income)	University degree	Not employed	Mental health
No bullying	-	-	-	-	-	-	-
No non-violent Low violent	-0.004	0.039	-18.84*	-0.022	-0.0090	-0.039**	0.453*
s.e.	0.026	0.031	10.233	0.031	0.0107	0.0131	0.2187
No non-violent High violent	-0.022	-0.018	-20.724	-0.037	-0.0289*	-0.0057	0.4176
s.e.	0.048	0.054	25.542	0.053	0.0188	0.0276	0.4387
Low non-violent No violent	-0.05**	0.035*	5.847	0.089***	-0.0115	0.0285*	0.5870**
s.e.	0.024	0.027	5.295	0.030	0.0106	0.0187	0.2035
Low non-violent Low violent	-0.07**	-0.012	7.602	0.030	0.0070	0.0223	0.7052**
s.e.	0.033	0.038	7.526	0.039	0.0128	0.0236	0.2893
Low non-violent High violent	-0.1***	-0.058	-3.547	-0.035	-0.045**	-0.0026	0.9149**
s.e.	0.037	0.043	19.371	0.050	0.0168	0.0233	0.3567
High non-violent No violent	-0.16**	-0.11**	14.500	-0.046	-0.0099	0.0183	0.6867**
s.e.	0.039	0.041	11.371	0.049	0.0170	0.0239	0.2934
High non-violent Low violent	-0.005	0.000	-22.6**	0.035	-0.021**	0.0248	0.861**
s.e.	0.042	0.045	11.567	0.048	0.0141	0.0298	0.3092
High non-violent/high violent	-0.12**	-0.11**	6.545	-0.050	-0.037**	0.079**	1.371***
s.e.	0.034	0.041	10.933	0.043	0.0133	0.0281	0.2736
N	5,924	6,650	3,531	6,650	6,650	6,650	6,378

Notes: Robust standard errors, clustered by school, in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The covariates included are from Specification 2.

@cepeo_ucl

ucl.ac.uk/ioe/cepeo