

Special Educational Needs and Youth Offending

Secondary Data Project Analysis Plan

Evaluating institution: University of York

Principal Investigator(s): Prof. Umar Toseeb



Analysis Plan for YEF Secondary Data Analysis Projects

Project summary

Project title	Special Educational Needs and Youth Offending
Research Team	Jo Hutchinson – Education Policy Institute Anna Leyland – University of Sheffield Nathan Hughes – University of Sheffield Emre Deniz – University of York Megan Frith – University of York
Principal investigator	Umar Toseeb
Analysis plan author(s)	Umar Toseeb, Jo Hutchinson, Anna Leyland, Nathan Hughes
Overarching research question	What is the relationship between special educational needs (SENs) and violent youth offending in England?
Supporting research question(s)	1. Are there differences between youth with and without SENs based on patterns of violent offending and disposals? 2. Which school-based activities lead to reduced violent offending for youths with SENs? 3. What is the impact of cautions on re-offending in youth with identified SENs involved in violent offending?
Dataset(s) to be used	1. Ministry of Justice and Department for Education Linked Dataset 2. Avon Longitudinal Study of Parents and Children
Population characteristics	Youth with and without SENs up to the age of 18 years
Years data spans	1. DfE-MoJ – birth years 2002-2004, data years all records up to 2021 (offending data) and 2022 (education data) 2. ALSPAC – birth years ~1990, data years ~1997-~2008
Geographic coverage	England
Primary outcome(s) investigated	Violent Offending
Main method(s) to be used or tested	Propensity score matching

--	--	--

Analysis plan history

Version	Date	Reason for revision
1.0 [original]	10.07.24	<i>[leave blank for the original version]</i>

Any changes to the design or methods need to be discussed with the YEF. Describe in the table above any agreed changes made to the design.

Table of contents

1. About the project.....	4
1.1. Background to the project	4
1.2. Research question(s)	6
1.3. Hypotheses.....	8
1.4. Key concepts.....	11
2. About the datasets.....	13
2.1. Overview of datasets used	13
2.2. Secondary data source(s)	14
2.3. Primary data collection.....	16
2.4. Linking datasets.....	16
2.5. Access and data protection	16
3. About the data	18
3.1. List of variables	18
3.2. Measurement of key concepts.....	33
3.3. Missing data and attrition.....	36
3.4. Other sources of bias.....	36
4. About the analysis.....	37
4.1. Overview of analytical approach	37
4.2. Approach to addressing research question(s).....	37
5. Project management	53
5.1. Risks and mitigations.....	53
5.2. Timeline	53
6. References.....	55

1. About the project

1.1. Background to the project

Youth with special educational needs (SENs) are disproportionately represented in the criminal justice system. Eighty seven percent of all violent youth offences in England are committed by those with an identified SEN (DfE, 2022). Knowing the link between age of first violent offence and SEN might help to tailor interventions. For example, if youth with SEN commit their first offence earlier than those without SEN, the window for intervention and prevention might need to be shorter for youth with SEN and diversionary schemes need to be tailored to younger groups to ensure earlier diversion from violent offending.

Youth with SENs may also have different experiences once in the criminal justice system. For example, those with speech, language, and communication needs may struggle with police questioning and subsequent court hearings affecting their disposal types (Hughes et al., 2020), which in turn may affect their subsequent likelihood of violent offending. We will test these trends and test hypotheses for how youth with SENs become involved in violent offending. This will lay the groundwork for further investigations of how practices within, and experiences of, the criminal justice system might increase violent offending for youth with SENs.

Children's SEN diagnostic labels can sometimes be problematic. In schools in England, most children with SEN are given a single label (primary need) and some are given a second label (secondary need); these can change over time resulting in more than two need types. Whilst the various SEN categories allow for some degree of specificity (e.g., speech, language, and communication needs), these may not capture the full extent of the child's challenges; in many cases, the primary labels mask the underlying difficulties. For example, many children identified as having behavioural difficulties have unidentified speech, language, and communication needs (Senedd, 2023). Therefore, relying solely on youth's diagnostic labels does not provide a full understanding of the underlying skills deficits that might make youth with SENs more prone to violent offending.

Our project will focus on violent offending from middle childhood onwards. Whilst younger children can be aggressive, we are interested in violent offending in the sense of illegal activity (i.e., committed after the age of criminal responsibility) or antecedents to illegal activity (e.g., anti-social/violent behaviour). Specifically for violent offending, we are concerned with recorded offences (as in charged or cautioned by the police), which cannot happen before the age of 10 years in England. In some cases there may be data on whether children have been involved in violence at school (exclusions data) before the age of 10 years. For lower age children and toddlers, some aggression is expected as part of typical development, for example it may be a sign of poor self-regulation (Tremblay et al., 2004). Very young children probably do not have the cognitive ability to intend to be violent, which is a key feature of aggression and violence according to one body of researchers (e.g., Shaw et al., 2002). Intentionality is likely to begin in middle childhood (Coie and Dodge, 1998). Therefore, whilst the age of criminal responsibility of age 10 years is not ideal from a

developmental perspective (see section on sources of bias), it broadly fits the timing of when violence can be intentional.

We will use data on children's social, emotional, cognitive, and learning profiles to identify specific challenges that are associated with subsequent violent behaviour. This will provide novel targets within schools for the reduction of violent offending. Such school-based targets for intervention are likely to be particularly appealing for policy makers as, in addition to reducing violence, they are likely to also improve educational attainment.

Existing evidence is limited as it does not provide potential levers for change. We will directly test possible levers that might help to prevent youth with SENs from violent offending.

- SEN identification. The rates of identified SENs have been increasing in recent years. This is likely because historically many youth with SENs went through school without ever being identified. Therefore, we will compare violent offending for youth with an identified SEN with those with an unidentified SEN. It may be that identification and support for SENs leads to reduced violent offending.
- Timing of SEN identification. International evidence suggests early identification of SENs reduces offending (Cronin & Addo, 2021). Therefore, we will compare youth with SEN who are identified earlier vs later to determine the impact of early identification on violent offending in England.¹
- Consistency of label. Many children lose their SEN label and support during school transition. We will compare children who retained their SEN label during school transition with those who lost their label to determine the impact of losing a SEN label during school transitions on subsequent violent offending. Additionally, we will compare those whose label changed between primary and secondary school (e.g., speech, language, and communication needs to social, emotional, and behavioural difficulties).
- A statement/EHCP. Children with SENs whose needs cannot be met by existing provision are provided additional support detailed in a statement of SEN/EHCP. We will compare children with SEN who have a statement of SEN/EHCP with those who do not to determine whether having a statement of SEN/EHCP prevents youth with SENs from violent offending. We recognise that there are differences *between* EHCPs and SEN statements and also differences *within* EHCPs and *within* SEN statements.

¹ There are a number of reasons why children may be identified earlier or later depending (e.g., type of difficulty or severity of need). Whilst we will not be able to distinguish between need severity using the data we have, we will be able to distinguish the different types of SENs, and we will do this. This will allow us to determine whether the timing of SEN identification affects subsequent offending, when the type of SEN is kept constant. Additionally, we will control for strong proxies for SEND such as primary attainment, attendance and exclusions in the statistical models.

There are lots of differences in how EHCPs/SEN statements are applied. We are not necessarily interested in EHCPs/SEN statements per se or their content; rather we are interested in what they represent (i.e., children who's needs cannot be met by existing provision). We will be effectively testing the effect of enhanced SEN support (whatever that might look like).

- Type of school provision. The type of school a child attends (e.g., mainstream school with/without SEN unit, special school, alternative provision etc.) might lead to different levels of SEN support. Therefore, we will compare youth with SENs in different types of school provision to determine whether school provision prevents youth with SENs from becoming involved in violent offending.

The use of cautions after a first offence might alter the developmental trajectory of violent offending. Evidence for the impact of the use of cautions for youths with SENs is lacking.

- Cautions. Little is known about the effectiveness of the different types of cautions (unconditional vs conditional) or the use of cautions instead of charging youth, to divert youth with SENs from the violence. We will compare youth with SENs who received different types of cautions to determine their effectiveness for reducing re-offending. Additionally, we will compare outcomes for those given a caution for a first offence compared to those who were charged.

1.2. Research question(s)

We will investigate the relationship between special educational needs (SENs) and violent youth offending in England.

We will do this by answering the following research questions:

1. Are there differences between youth with and without identified SENs based on:
 - a. Age of first violent offence,
 - b. Number of violent offences (single offence vs multiple offences),
 - c. Caution for first violent offence (no or yes),
 - d. Disposal type (custodial sentence vs other).
2. To what extent do social, emotional, cognitive, and learning skills in childhood deficits predict violent behaviour in adolescence?
3. To what extent do the following lead to reduced violent offending for youth with SENs:
 - a. Identification of SEN (identified vs unidentified),
 - b. Timing of SEN identification (relative to the typical time of identification for that SEND need type),
 - c. Consistency of SEN label during school transition (consistent vs inconsistent)
 - d. Having a statement of SEN/education health and care plan (EHCP versus SEND support),
 - e. Type of school provision:

- i. special vs any mainstream (with an EHCP),
 - ii. alternative provision vs any mainstream (with an EHCP),
 - iii. mainstream with a SEN unit vs mainstream with neither a SEN unit nor resourced provision (with EHCP)
 - iv. mainstream with a SEN unit vs mainstream with neither a SEN unit nor resourced provision (without EHCP)
 - v. mainstream with resourced provision vs mainstream with neither resourced provision nor a SEN unit/mainstream without SEN unit other (with EHCP)
 - vi. mainstream with resourced provision vs mainstream with neither resourced provision nor a SEN unit (without EHCP).
4. What is the impact of cautions on re-offending in youth with identified SENs involved in violent offending. Specifically:
- a. Does being cautioned, after first offence, reduce violent re-offending for youth with SENs? (youth with identified SENs cautioned vs youth with identified SENs charged)
 - b. Is there a difference in re-offending rates for youth with SENs depending on the type of caution? (unconditional vs conditional)

Table 1.2. How will the questions be addressed at each stage?

Question Number ²	Interim report	Final report
1	DfE-MoJ Dataset ● Extract individuals with identified SENs from DfE-MoJ dataset ● Produce description of DfE-MoJ dataset and write methodology section ● Clean variables and produce descriptive statistics for all variables of interest for RQ1 ● Fit inferential statistical models for RQ1 ● Provide high level summary of findings	DfE-MoJ Dataset ● Revise the interim report in light of reviewer comments ● Write detailed description of the findings ● Summarise results in tables/figures ● Write discussion and implications of the findings ● Make recommendations for criminal justice services
2	ALSPAC Dataset ● Clean variables and produce descriptive statistics for all	ALSPAC Dataset ● Revise the interim report in light of reviewer comments

²Question numbers should follow the ordering in the section above.

	variables of interest for RQ2 using the ALSPAC dataset ● Produce description of ALSPAC dataset and write methodology section ● Fit inferential statistical models for RQ2 ● Provide brief summary of findings ● Summarise results in tables/figures	● Write detailed description of the findings ● Write discussion and implications of the findings ● Make recommendations for education providers
3-4	No results will be included in interim report	DfE-MoJ Dataset ● Clean variables and produce descriptive statistics for all variables of interest for RQs3-4 ● Fit inferential statistical models for RQs3-4 ● Write detailed description of findings ● Summarise results in tables/figures ● Write discussion and implications of the findings ● Make recommendations for education providers (RQ3) ● Make recommendations for education providers (RQ4)

1.3. Hypotheses

1. We expect there to be a number of differences between youth with and without identified SENs based on:
 - a. Age of first violent offence. Youth with identified SENs will become involved in violent offending earlier than those without SENs. We expect this given that youth with identified SENs are up to 5 times more likely to be excluded from school than those without SENs (DfE, 2022) and their difficulties often go unnoticed and so more likely to engage in anti-social behaviour (Taylor, 2016) at an earlier age, which may then escalate to violent offending.
 - b. Number of violent offences (single offence vs multiple offences). Youth with identified SENs will be more likely to engage in multiple violent offences than those without SENs. We expect this given the lack of appropriate support after coming into contact with the criminal justice system (Centre for Justice and Innovation, 2024). Additionally, even when support is offered, those with identified SENs might struggle to access support services due to their difficulties (e.g., interventions might rely on social skills or a particular level of language

proficiency). For example, those with dyslexia might struggle to understand the terms of their parole leading to violations and so further criminalising them (Baldry et al., 2017).

- c. Caution for first offence (no vs yes). We expect youth with identified SENs to be less likely to be cautioned for their first offence compared to youth without SENs. Youth with identified SENs might struggle to understand what is being asked of them during police questioning (Hughes et al, 2020). In many cases, an appropriate adult, who should be there to support them, is not provided (Jessiman and Cameron, 2017) and thus might be less likely to meet some of the criteria for being given a caution (e.g., admission of guilt and willingness to accept a caution, CPS, 2022).
 - d. Disposal type (custodial sentence vs other). We expect youth with identified SENs to be more likely to receive a custodial sentence than youth without SENs. Young people with SENs might find it more difficult to navigate the legal proceedings given their vulnerabilities. For example, those with communication difficulties might struggle to recall and narrate an account of what happened, elaborate when questioned, and make eye contact, which may be perceived as non-compliant or deliberately difficult (Hughes et al., 2020). Additionally, previous diversionary exercises might be less successful with youth with SENs due to how they are delivered (Centre for Justice Innovation, 2023). Verbally mediated interventions may be less effective for those with speech, language, and communication needs. Therefore, subsequent offending might be met with harsher disposal types (i.e., custodial sentence).
2. To what extent do social, emotional, cognitive, and learning skills deficits in childhood predict violent behaviour in adolescence? We expect social, emotional, cognitive, and learning skills deficits to predict violence to a greater or lesser extent. We expect this given that a) not all SENs are identified and b) SEN labels do not capture the breadth of difficulties experienced by young people. We expect all of the different types of difficulties to predict violent offending to a greater or lesser extent but we are unable to predict which type of difficulty is likely to be most predictive of violent behaviour.
 3. To what extent do the following lead to reduced violent offending for youth with SENs:
 - a. Identification of SEN (identified vs unidentified). We expect those with identified SENs to be less likely to be involved in violent offending than those with unidentified SEN. We know that identification of SEN is dependent on much more than the child's needs; school, neighbourhood, and family factors are also important (Hutchinson, 2021). We assume that those who are identified with SENs are also supported (see analysis section for further details on how we will quantify this), although we acknowledge that this is not universally true. This additional support, in theory, means young people are less likely to be excluded and more likely to remain engaged with and feel part of school. Additionally,

having a label might highlight to parents and teachers that the young person is vulnerable and additional protective measures might be put in place.

- b. Timing of SEN identification (relative to the typical timing of identification for the type of SEND need). We expect there to be differences in violent offending based on when the SEN was identified but these are likely vary depending on the types of SEN. For example, speech, language, and communication needs / autism spectrum conditions are likely to manifest in primary school whereas social, emotional, and mental health difficulties are more likely to manifest in secondary school. Earlier identification and support will reduce the likelihood of becoming involved in violent behaviour and subsequently violent offending. In theory, additional resources and plans are put in place, which will divert young people from violent behaviour and subsequent violent offending. The later this happens, the more opportunity there is for youth with SENs to become involved in anti-social behaviour; especially if it does not happen before the onset of adolescence. A key developmental period during which youth are particularly susceptible to social and emotional difficulties and peer influence.
- c. Consistency of SEN label during school transition (consistent vs inconsistent). We expect those who lose their SEN label in the transition from primary to secondary school will be more likely to be involved in violent offending. The definition of consistency can be thought of in two ways 1) whether youth keep *any* label in the transition from primary to secondary or 2) whether they keep the *same type of SEN* during transition. We assume that a continuation of a label means a continuation of support. Therefore, if there is cessation of support, youth with SENs are more likely to become involved in violent behaviour and subsequently violent offending.
- d. Having a statement of SEN/education health and care plan (EHCP versus SEND support). We expect those with a statement of SEN/EHCP to be less likely to be involved in violent offending. Those with EHCPs often have more severe needs, which cannot be met by existing provision. Therefore, they may receive additional support and supervision preventing them from being excluded from school and getting involved in violence. We recognise that certain types of SENs might be more likely to have a statement of SEN/education health and care plan. We will, therefore, conduct sub-group analysis comparing those with similar types of SEN need.
- e. Type of school provision (special vs any mainstream, alternative provision vs any mainstream, mainstream with an SEN unit vs mainstream with NEITHER a SEN unit nor a resourced provision, mainstream with resourced provision vs mainstream with NEITHER resourced provision nor a SEN unit). We expect those attending any sort of specialist provision to be less likely to be involved in violent offending. We assume that specialist provision is better resourced and more

attuned to the challenges faced by youth with SENs, and, therefore, are better equipped to provide support and foster a sense of belonging.

4. What is the impact of cautions on re-offending in youth with identified SENs involved in violent offending. Specifically:
 - a. Caution after first offence. We expect those who are cautioned to be less likely to reoffend than those who are charged with an offence. We expect this given that a caution has fewer social (e.g., stigma) and economic consequences (e.g., fewer employment issues arising from a criminal record). Whilst the evidence of the effectiveness of cautions is mixed, a recent overview by HM Inspectorate of Probation (2023) suggests there are likely to be positive effects.
 - b. Type of caution. We expect youth with SENs who are given a conditional caution to be more likely to reoffend compared youths with SENs who are given unconditional cautions ³. Youth with specific types of SENs might struggle to understand the conditions of their caution, or the implications of violating the conditions, and therefore more likely to breach the conditions of their caution.

1.4. Key concepts

Table 1.4 Definitions of key concepts

Terms	Definition used
<i>Special Educational Needs</i>	We will use the DfE definition of SEN as outlined in their Code of Practice (2015). Specifically • “a child or young person has SEN if they have a learning difficulty or disability which calls for special educational provision to be made for him or her. • A child of compulsory school age or a young person has a learning difficulty or disability if he or she: • has a significantly greater difficulty in learning than the majority of others of the same age, or • has a disability which prevents or hinders him or her from making use of facilities of a kind generally provided for others of the same age in mainstream schools or mainstream post-16 institutions”
<i>Identified Special Educational Needs</i>	A young person has an identified special educational need if they have a record of having a special educational need on their school SEN register; specifically in the school census.

³ As we are only interested in offending up to the year 2021, changes in caution options after this date are unlikely to apply to the data.

<i>Statement of Special Educational Needs</i>	We define a statement of special educational needs as “a formal document issued by the local authority that describes a child’s learning difficulties and the special educational provision required to meet those needs” Statements were phased out and replaced with education health and care plans from September 2014.
<i>Education Health and Care Plan</i>	We will use the DfE definition of an Education Health and Care plan as “An education, health and care (EHC) plan is for children and young people aged up to 25 who need more support than is available through special educational needs support. EHC plans identify educational, health and social needs and set out the additional support to meet those needs”
<i>Violent Offending</i>	We will use YEF’s definition of violent crime, as described in the <u>outcomes framework</u> , to define violent offending as “Criminal acts involving harm against another person (e.g. assault, robbery using threat or force, homicide)” In addition to this, we will also use YEF’s definition of sexually violent crime to supplement our definition “There are many forms of sexually violent crimes that can take place in a range of settings. We focus on two forms of sexual violence most relevant to YEF programmes and the target age range of young people: sexual violence in a relationship and sexual harassment”. Violent offences will be those where the young person has been cautioned or convicted.
<i>Violent Behaviour</i>	We will use YEF’s definition of violent crime, as described in the <u>outcomes framework</u> , to define violent behaviour as “Criminal acts involving harm against another person (e.g. assault, robbery using threat or force, homicide)”. To be counted as violent behaviour, the young person does not need to have been cautioned or convicted.
<i>Disposal Type</i>	For offenders who were convicted of an offence, we define disposal type as the type of sentence imposed by the court.
<i>Caution Type</i>	For offenders who were cautioned for an offence, we define caution as the type of caution given.
<i>Re-offending</i>	We will use the MoJ definition of re-offending as outlined in their <u>re-offending statistics document</u> . Specifically: “Offences are counted as re-offences if they meet all of the following criteria: • They are recordable • The offences were committed in England and Wales • They are offences that were prosecuted by the police.

	• Offences are only counted if they are proven through caution (for adults), reprimands or final warnings (for juveniles) and court convictions. • The offence is not a breach offence (i.e., breach of a court order)” Additional considerations The focus of the re-offending analysis will be recidivism. Additionally, we will aim to use 12 months as the definition of time at liberty. We recognise that using the follow-up period 12 months for a re-offence will likely yield different results to using a one calendar year.
<i>Social Difficulties</i>	Social difficulties are any behaviours that affect a young person’s interactions with those around them. More broadly, this can be thought of as underlying psychological distress manifesting outwards. Specific examples include, conduct problems (e.g., temper tantrums), hyperactivity (e.g., restless or overactive), peer problems (e.g., solitary), bullying, and autistic characteristics.
<i>Emotional Difficulties</i>	Emotional difficulties are any emotions/behaviours that are indicative that a young person’s emotional well-being is suffering. More broadly, this can be thought of as underlying psychological distress manifesting inwards. Typically, emotional difficulties can be thought of as symptoms of depression (e.g., low mood, withdrawal), anxiety (e.g., worry), and psychosomatic symptoms (e.g., often complaining of headaches)
<i>Cognitive and Learning Difficulties</i>	Cognitive and learning difficulties are difficulties that affect a young person’s ability to process, retain, and use information. These typically include reading, structural language (e.g., being able to understand others), pragmatic language (e.g., being able to use language appropriately in a social situation), non-verbal ability (commonly known as IQ), and short-term memory.

2. About the datasets

2.1. Overview of datasets used

We will rely upon two data sources: Department for Education-Ministry of Justice Linked (MoJ-DfE) Dataset and the Avon Longitudinal Study of Parents and Children (ALSPAC).

DfE-MoJ Linked Dataset

The DfE-MoJ dataset is a linkage of two administrative datasets; the National Pupil Database (NPD) and the Police National Computer (PNC). The NPD contains routinely collected

administrative data on all children in state-funded schools, academies, and free schools in England. The PNC contains data on any individual who has received a caution or been convicted of an offence in England.

ALSPAC

The ALSPAC cohort is a birth cohort study of approximately 15,000 children born in the early 1990s in the Avon region of England. Pregnant women resident in Avon, with expected dates of delivery between 1st April 1991 and 31st December 1992, were invited to take part in the study. Some booster recruitment was carried out in subsequent years. The parents and children were contacted periodically and asked to complete questionnaires and assessments of their social, emotional, and cognitive functioning.

2.2. Secondary data source(s)

Table 2.2a Dataset Description - Department for Education-Ministry of Justice Linked (MoJ-DfE)

Name of dataset	Department for Education-Ministry of Justice Linked (MoJ-DfE)
Data owner(s)	Department for Education and Ministry of Justice
Type of data	Administrative Records
Availability of data	Data can be accessed by accredited researchers who want to use the data for public good.
Team member(s) who will have access	Umar Toseeb, Emre Deniz, & Megan Frith
Population/geographic coverage or sampling frame	The dataset covers young people who have been convicted or cautioned for an offence and have records in the NPD. The NPD includes data on students from state-funded schools, academies, and free schools in England.
Years covered or survey waves	The data relates to those offenders with at least one record from 2000 or later, who were on the PNC at the end of 2021 and were matched to individuals on the NPD. Only offenders who were born on, or after 31 August 1985 were matched, because earlier groups do not have a realistic chance of matching.
Exclusion criteria	We will include in our analysis individuals born between 2002-2004 who committed an offence between ages of 10-17 years.

Expected population/sample size (following exclusion criteria)	There are approximately 2 million individuals who were born between 2002-2004 who have a record on the NPD. According to the DfE (2022): • Approximately 5% would have offended between the ages of 10-17 = ~100,000 • Approximately 23% of these offences are violent offences = ~23,000 • Approximately 83% of these violent offences were committed by those with an identified SEN = ~19,000 Therefore, for those involved in violent offending, we expect a sample size of ~23,000 (of which ~19,000 will have an identified SEN). This is the minimum sample size for any of the analysis using the linked DfE-MoJ dataset. For the broader analysis, involving those who have and have not offended, we expect that of the 2 million individuals, approximately 700,000 (~35%) would have ever been identified as having a SEN.
Documentation	A user guide for the dataset can be found here .

Table 2.2b Dataset Description – Avon Longitudinal Study of Parents and Children (ALSPAC)

Name of dataset	Avon Longitudinal Study of Parents and Children
Data owner(s)	University of Bristol
Type of data	Longitudinal cohort study with data to the NPD
Availability of data	Data can be accessed by bona fide researchers who follow ALSPAC's data access policy .
Team member(s) who will have access	Umar Toseeb, Emre Deniz, & Megan Frith
Population/geographic coverage or sampling frame	The cohort is a community sample of families in the former English county of Avon. Broadly, this includes the cities of Bristol, Bath, and some parts of Somerset and Gloucestershire.

Years covered or survey waves	Mothers were recruited to the study during pregnancy during 1991-1992. For this analysis we will primarily use data collected during the “in focus” clinics between ages 7-9 years and questionnaire data when the young person was 17-18 years old.
Exclusion criteria	Some families have multiple children per family (in the case of twins and triplets). To avoid intra-family clustering, we will only use data from one child per family. In most cases these are twins and triplets, who account for approximately 1% of the sample.
Expected population/sample size (following exclusion criteria)	The sample sizes for each of the waves we will be using are: • Age 7-9 in Focus: ~8,000 • Age 17-18 Questionnaire: ~3,500 • Overlap: ~3,000 young people took part at both time points. Approximately 20% has an identified SEN (according to linked NPD records). Therefore, our predicted sample size is ~3,000 young people (of which ~600 have an identified SEN).
Documentation	Further details of the cohort can be found on the ALSPAC webpage .

2.3. Primary data collection

No primary data will be collected

2.4. Linking datasets

Whilst our project involves linked datasets, these linkages have already been done. Therefore, for our project, no further linking is needed.

2.5. Access and data protection

Accessing the DfE-MoJ Linked Dataset.

The data access process for the DfE-MoJ linked dataset is outlined in the [Data First User Guide](#). Briefly, this consists of four steps:

1. Submission of a research application. This includes working with the relevant members of the data team to make the best case for the project. This stage also includes the identification of relevant variables.

2. Review of applications by the Data Access Group and Data Sharing Approval Panel. The groups include subject matter experts, independent analysts, and social researchers.
3. Approval by Data Access Governance Board. This group consists of senior analysts and is chaired by the Chief Statistician. They decide whether or not to approve data access based on recommendations from the Data Access Group.
4. Legally binding data access agreements are then signed by the data holders and the home institution of the lead researcher.

Accessing the ALSPAC Dataset.

The data access process for the ALSPAC dataset is outlined in their [Data Access Policy](#).

Briefly, this consists of the following four steps:

1. Submission of an online proposal form. This proposal typically includes clear aims and hypotheses with a description of the requested variables.
2. The proposal is reviewed by the Data Access Committee who ensure that the proposal meets ethical guidelines and complies with ALSPAC's data access policies.
3. Legally binding data access agreements are then signed by the data holders and the home institution of the lead researcher.
4. The ALSPAC data team calculates the cost of the requested variables and the host institution of the lead researcher pays this amount.

Data Protection

All data will be handled in line with local policies. Specifically, the policies relevant for the proposed project are:

1. University of York Data Protection Policy
(<https://www.york.ac.uk/media/recordsmanagement/documents/dataprotectiondocs/University%20of%20York%20Data%20Protection%20Policy%20v1.2.pdf>).
2. University of York Special Categories and Personal Data Policy
(<https://www.york.ac.uk/media/recordsmanagement/documents/dataprotectiondocs/University%20of%20York%20Data%20Protection%20Policy%20v1.2.pdf>). This is to ensure criminal convictions and offences data is handled in line with Article 10 of GDPR.

For the MoJ-DfE dataset, we will adopt the five safes framework, which is recommended by the data holders. We will implement the framework as follows:

- Safe People. All of the research team with access to the data will become accredited researchers by completing the ONS accredited researcher training.
- Safe Projects. The project will be considered by the MoJ Data Access Group and DfE Data Sharing Approval Panel to ensure that data is only used for valuable and ethical research that delivers clear public benefits.
- Safe Settings. We will access the data via an ONS Secure Service, which operates under standardised policies and procedures approved by data holders such as the DfE and MoJ.

- **Safe Outputs.** The Office for National Statistics safe outputs protocols will be followed. This means that any outputs (e.g., tables, graphs etc.) will be sent for approval before they are taken out of the secure environment (all research outputs are checked to ensure they cannot identify individuals).
- **Safe Data.** All data is provided in an anonymised format so the risk of re-identification is low.

Data Processing Roles

The University of York will act as the data controller. The purpose and means of processing data is outlined in the [University of York Data Protection Policy](#) and [Special Categories and Personal Data Policy](#).

The following team members will act as data processors: Umar Toseeb, Emre Deniz, and Megan Frith.

Data Access Termination

DfE-MoJ Dataset. Access to the secure server will cease after the project end date.

Therefore, it will no longer be possible to access data.

ALSPAC. After the outputs for the project have been finalised, data will be encrypted and then deleted. This will ensure that the deleted files cannot be recovered.

3. About the data

3.1. List of variables

Table 3.1a: Variable definitions for DfE-MoJ Dataset

Variable abbreviation	Variable definition	Variable source	Derivation or specification
VO	A categorical variable identifying individuals who have been convicted of a violent offence (0=no, 1= yes).	PNC	We will use the Offence_group variable to derive the violent offence variable. See Table 3.2 for a full description.
I-SEN	A categorical variable identifying individuals who have an identified special educational needs (0=no SEN, 1=identified SEN).	SC-school census	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. See Table 3.2 for a full description.
T-SEN	A categorical variable describing the timing of SEN identification. It will consist of two levels (0 = primary	SC	We will use the PrimarySENtype and SecondarySENtype, as well as school year group data, to derive the T-SEN variable.

	school, 1 = secondary school).		
C-SEN	A categorical variable describing the consistency of SEN labels between primary school (year 6) and secondary school (mid year 8). It will consist of 2 levels (0 = SEN label in primary school but not secondary school, 1 = SEN label in primary school and secondary school).	SC	We will use the T-SEN variable, described previously to derive the C-SEN variable.
C-SEN-TYPE	A categorical variable describing the consistency of the type of SEN label between primary school (year 6) and secondary school (mid year 8). It will consist of 2 levels (0 = SEN label type the same in primary and secondary school, 1 = SEN label type different in primary and secondary school).	SC	We will use the C-SEN variable, described previously to derive the C-SEN-TYPE variable. Those who have a flag of 1 = SEN label in primary school and secondary school in the C-SEN variable, will included in this variable using the data from the following variables SPLD, MLD, SLD, PMLD, SEMH-BESD, SLCN, HI, VI, MSI, PD, ASD, OTH, SEN-NSA.
SLD-PMLD	A categorical variable identifying individuals who have ever had a severe learning difficulty or profound and multiple learning difficulty (0=no, 1=yes).	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. Those with the label “SLD = Severe Learning Difficulty” and “PMLD = Profound & Multiple Learning Difficulty” will be included.
SENS	A categorical variable identifying individuals who have ever had	SC	We will use the PrimarySENtype and SecondarySENtype to derive the SENS variable. Those with

	sensory impairments (0=no, 1=yes).		the label “VI = Visual Impairment”, “MSI = Multi-Sensory Impairment”, or “HI = Hearing Impairment” will be included.
PD	A categorical variable identifying individuals who have ever had a physical disability (0=no, 1=yes).	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. Those with the label “PD = Physical Disability” will be included.
OTH-SEN	A categorical variable identifying individuals who have ever had any other disability (0=no, 1=yes).	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. Those with the label “OTH = Other Difficulty/Disability” will be included.
ASD	A categorical variable identifying individuals who have ever had an identified autism spectrum disorder (0=no, 1=yes).	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. Those with the label “ASD = Autistic Spectrum Disorder” will be included.
SLCN	A categorical variable identifying individuals who have ever had a speech, language, and communication need (0=no, 1=yes).	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. See Table 3.2 for a full description. Those with the label “SLCN = Speech, Language and Communication Needs” will be included.
SEN-NSA	A categorical variable identifying individuals who have not yet had an assessment of their need (0=no, 1=yes).	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. See Table 3.2 for a full description. Those with the label “NSA = SEN support but no specialist assessment of type of need” will be included.
SPLD	A categorical variable identifying individuals who have ever had an identified specific	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. See Table 3.2 for a full description. Those with

	learning disability (0=no, 1=yes).		the label “SPLD = Specific Learning Difficulty” will be included.
MLD	A categorical variable identifying individuals who have ever had an identified moderate learning difficulty (0=no, 1=yes).	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. See Table 3.2 for a full description. Those with the label “MLD = Moderate Learning Difficulty” will be included.
SEMH-BESD	A categorical variable identifying individuals who have ever had a social emotional and mental health difficulty or behavioural, emotional, and social difficulties (0=no, 1=yes).	SC	We will use the PrimarySENtype and SecondarySENtype to derive the I-SEN variable. See Table 3.2 for a full description. Those with the labels “Behaviour, Emotional & Social Difficulties” or “SEMH = Social, emotional and mental health” will be included.
STAT	A binary variable describing whether a young person with SEN has ever had a statement. It will consist of two levels (0=no, 1 = yes).	SC	We will use the SENprovision variable to derive the STAT variable. See Table 3.2 for a full description.
EHCP	A binary variable describing whether a young person with SEN has ever had an education health and care plan. It will consist of two levels (0=no, 1 = yes).	SC	We will use the SENprovision variable to derive the EHCP variable. See Table 3.2 for a full description.
STAT_EHCP	A binary variable describing whether the young person with SEN has ever had either a statement of SEN or EHCP (0=no, 1 = yes).	SC	We will derive this variable using the STAT and EHCP variables. If either of the STAT or EHCP variables is scored 1, then the score on this variable will be scored 1.

VO-AGE	A continuous variable describing the age of each violent offence.	PNC	We will use the OffenceStartAge variable to determine the age of the violent offence. We will extract all ages between 10 and 17 years.
NO-VO	A continuous variable describing the number of violent offences a young person has been convicted of or cautioned for.	PNC	We will use the CaseIDs variable to derive the number of violent offences. The number of unique CaseIDs will be used as an indicator of the number of offences.
NO-VO-POST	A continuous variable describing the number of violent offences a young person has been convicted or cautioned for, that they committed <i>after</i> their first conviction or caution.	PNC	We will use the CaseIDs variable to derive the number of violent offences. The number of unique CaseIDs will be used as an indicator of the number of offences. We will use variable CourtCautionDate to determine when the caution/conviction took place and OffenceStartAge to determine when each offence was committed, and which was the first offence.
DISP	A binary variable describing the type of sentence given to the young person for their first offence (0=custodial, 1 = other).	PNC	We will derive this variable using the PNCDisposalCode variable. See Table 3.2 for a full description. We will use OffenceStartAge to determine when each offence was committed, and which was the first offence.
CAUT	A binary variable describing the type of caution given to a young person for their first offence. It will consist of two levels (0=unconditional, 1 = conditional)	PNC	We will use the Cautiontype variable to derive this variable. See Table 3.2 for a full description. We will use OffenceStartAge to determine when each offence was committed, and which was the first offence.

S-SPE	A binary variable describing whether a young person has ever attended a special school (0=no, 1= yes).	SC	We will use the Phase variable to derive this variable. Young people with “SP = Special” in this field will be extracted as those attending a special school.
S-PRU	A binary variable describing whether a young person has ever attended a pupil referral unit (0=no, 1= yes).	SC	We will use the Phase variable to derive this variable. Young people with “PR = Pupil Referral Unit” in this field will be extracted as those attending a pupil referral unit.
S-MSS	A binary variable describing whether a young person has ever attended a mainstream school with an SEN unit. (0=no, 1= yes).	SC	We will use the SENUnitIndicator variable to derive this variable. Young people with “1 = true” in this field will be extracted as attending a mainstream school with an SEN unit.
S-MSR	A binary variable describing whether a young person has ever attended a mainstream school with resourced provision. (0=no, 1= yes).	SC	We will use the ResourcedProvisionIndicator variable to derive this variable. Young people with “1 = true” in this field will be extracted as attending a mainstream school with resourced provision.
S-AP	A binary variable describing whether a young person has ever attended alternative provision. (0=no, 1= yes).	AP – Alternative Provision Census	If the young person appears in the alternative provision census they will be categorised as attending alternative provision.
SC-HL	A categorical variable describing the highest level of social care intervention the child has ever been under for any reason	CIN, CLA	1. We will use the variables cla_CAT_NEED to derive cases where the highest level is ‘looked after’, and the variables cla_POC_START, cla_DATE_EPI_COMM, and cla_PROCESSING_YEAR to determine the timing of the the care episode with reference to the violent offending.

			2. We will use the variable cin_CPPstartDate to derive cases with one or more CPPs where the highest level is 'Child Protection Plan', and to determine the timing of the earliest CPP's commencement with reference to the violent offending. 3. We will use the variable CIN_PrimaryNeedCode to derive cases where the highest level is 'child in need', and CIN_CINReferralDate to determine the timing of the earliest period of CIN with reference to violent offending. We will use the variables CIN_ReasonForClosure and CIN_ReferralNFA to exclude cases which were referred but dropped after an initial assessment so that no support was provided.
CIN-NA	A binary variable describing whether the young person has been referred as a 'Child in Need' at least once but never (for any referral) received any support after initial assessments.	CIN	We will derive this using the variables CIN_ReasonForClosure and CIN_ReferralNFA to identify cases where all referrals were closed without action, and CIN_CINReferralDate to determine the timing of the earliest referral with reference to violent offending.
CI-REAS	A categorical variable describing the reason (category of need) for the earliest period as a 'child in need'.	CIN	We will derive this from the variable CIN_PrimaryNeedCode, using CIN_CINReferralDate to determine the timing of the earliest referral, and applying the baseline reference value = never CIN.
CIN	A binary variable describing whether a young person has ever been identified as	CIN - Child in Need dataset	If the young person appears in the child in need dataset they will be categorised as ever been a child in need.

	being in need (0=no, 1 = yes)		
ETH	A categorical variable describing the young person's ethnicity (1=Asian or Asian British, 2=Black, Black British, Caribbean, or African, 3=Mixed or multiple ethnic groups, 4= White, 5=other)	SC	We will derive this variable using the EthnicGroupMajor variable following the <u>2021 Census definitions</u> .
ETH-18	A categorical variable describing the young person's ethnicity using the 23 code list.	SC	Derived from EthnicGroupMinor_[term][yy] and ordered so that White British appears first and becomes the reference value when used as a regression covariate.
BIRTHY	A categorical variable describing the <i>academic</i> year of birth of the young person, and therefore which cohort they fall into.	SC	Derived from the variables YearOfBirth_[term][yy] and MonthOfBirth_[term][yy].
BIRTHM	A categorical variable describing the month of birth of the young person.	SC	Derived from MonthOfBirth_[term][yy] and ordered from September-born children (the eldest in each school year group) to August-born children.
MALE	A binary variable where 0 = female and 1 = male.	SC	Derived from the variable Gender_[term][yy].
EAL	A binary variable where 1 = a young person known or believed to speak English as an Additional Language and 0 = any other young person.	SC	Derived from the variable LanguageGroupMajor_[term][yy].

FSM	A categorical variable describing whether a young person has ever been eligible for free school meals (0=no, 1 = yes)	SC	Derived from the variables FSMeligible_[term][yy] and FSMeligibility_[term][yy].
FSM-DUR	A categorical variable describing FSM duration, with a baseline value of 0 = never FSM and five further categories of 1-19% of possible years eligible for FSM, 20-39%, 40-59%, 60-79% and 80-100% of possible years eligible.	SC	Derived from the variables FSMeligible_[term][yy] and FSMeligibility_[term][yy].
IDACI	A continuous variable describing the neighbourhood deprivation rate for families with children in the LSOA of the young person's residence.	SC	Derived from the variable IDACIScore_[term][yy], and BIRTHY.
IMD-EMP	A continuous variable describing the neighbourhood employment deprivation rate for adults in the LSOA of the young person's residence.	SC, IMD	Derived from the variables LLSOA_[term][yy] LSOA11_[term][yy] LSOA21_[term][yy] used to link to the IMD Proportion of working age adults in employment deprivation.
IMD-ED	A continuous variable describing the neighbourhood education, skills and training deprivation rate in the LSOA of the	SC, IMD	Derived from the variables LLSOA_[term][yy] LSOA11_[term][yy] LSOA21_[term][yy] used to link to the IMD Proportion in education deprivation.

	young person's residence.		
IMD-HEA	A continuous variable describing the neighbourhood health deprivation rate in the LSOA of the young person's residence.	SC, IMD	Derived from the variables LLSOA_[term][yy] LSOA11_[term][yy] LSOA21_[term][yy] used to link to the IMD Proportion in health deprivation.
IMD-CRIM	A continuous variable describing the neighbourhood crime victimisation deprivation rate in the LSOA of the young person's residence.	SC, IMD	Derived from the variables LLSOA_[term][yy] LSOA11_[term][yy] LSOA21_[term][yy] used to link to the IMD Proportion in crime deprivation.
IMD-HS	A continuous variable describing the neighbourhood housing and services deprivation rate in the LSOA of the young person's residence.	SC, IMD	Derived from the variables LLSOA_[term][yy] LSOA11_[term][yy] LSOA21_[term][yy] used to link to the IMD Proportion in housing and services deprivation.
IMD-ENV	A continuous variable describing the neighbourhood living environment deprivation rate in the LSOA of the young person's residence.	SC, IMD	Derived from the variables LLSOA_[term][yy] LSOA11_[term][yy] LSOA21_[term][yy] used to link to the IMD Proportion living environment deprivation.
ABS-UA	A continuous variable describing the percentage of possible half days that were missed due to unauthorised absence.	SC ABS	Derived from the variables TermlyReasonU_Summer_ab[yy] TermlyReasonU_Spring_ab[yy] TermlyReasonU_Autumn_ab[yy].
ABS-ILL	A continuous variable describing the percentage of possible half days that were	SC ABS	Derived from variables TermlyReasonI_Summer_ab[yy] TermlyReasonI_Spring_ab[yy] TermlyReasonI_Autumn_ab[yy].

	missed due to sickness absence.		
ABS-ALL	A continuous variable describing the percentage of possible half days that were missed due to sickness absence.	SC ABS	Derived from variables for all reason codes for the Summer, Spring and Autumn terms.
MOB	A continuous variable describing how many schools the young person has attended.	SC, GIAS	URN_[term][yy] linked to GIAS institution links file.
SUS	A continuous variable describing how many suspensions the young person has received.	SC EXC	Derived from TotalFixedExclusions_ex[yy]
SUS-V	A continuous variable describing how many suspensions the young person has received for physical assault against a pupil or adult, or sexual misconduct.	SC EXC	Derived from Reason_ex[yy] and Reason_A_ex[yy]
SUS-SESS	A continuous variable describing how many half days the young person has been suspended for.	SC EXC	Derived from TotalFixedSessions_ex[yy]
PX	A continuous variable describing how many permanent exclusions the young person has received.	SC EXC	Derived from PermanentExclusionCount_ex[yy]
EY-COM	A standardised score based on EYFSP communication and literacy scores.	EYFSP	Derived from FSP_CLL_TOTAL FSP_COM_G01 FSP_COM_G02 FSP_COM_G03 FSP_LIT_G09 FSP_LIT_G10

EY-MAT	A standardised score based on EYFSP numeracy scores.	EYFSP	Derived from FSP_PSRN_TOTAL FSP_MAT_G11 FSP_MAT_G12
EY-PSE	A standardised score based on EYFSP personal, social, emotional scores.	EYFSP	Derived from FSP_PSE_TOTAL FSP_PSE_G06 FSP_PSE_G07 FSP_PSE_G08
EY-PHY	A standardised score based on EYFSP physical development scores.	EYFSP	Derived from FSP_RIPD FSP_ PHY_G04 FSP_PHY_G05
EY-UW	A standardised score based on EYFSP understanding of the world scores.	EYFSP	Derived from FSP_RKUW FSP_UTW_G13 FSP_UTW_G14 FSP_UTW_G15
EY-CD	A standardised score based on EYFSP creative development scores.	EYFSP	Derived from FSP_RICD FSP_EXP_G16 FSP_EXP_G17
K1-READ	A standardised score based on KS1 reading assessments.	KS1	Derived from KS1_READING KS1_READ_OUTCOME
K1-WRIT	A standardised score based on KS1 writing assessments.	KS1	Derived from KS1_WRITING KS1_WRIT_OUTCOME
K1-MATH	A standardised score based on KS1 maths assessments.	KS1	Derived from KS1_MATHS KS1_MATH_OUTCOME
K1-SCI	A standardised score based on KS1 science assessments.	KS1	Derived from KS1_SCIENCE KS1_SCI_OUTCOME
K2-READ	A standardised score based on KS2 reading assessments.	KS2	Derived from KS2_READOUTCOME KS2_READSCORE KS2_READTAOUTCOME KS2_READMARK

			KS2_READMRK
K2-WRIT	A standardised score based on KS2 writing assessments.	KS2	Derived from KS2_MATOUTCOME KS2_MATSCORE KS2_MATTAOUTCOME KS2_MATMRK KS2_MATTOTMRK
K2-MATH	A standardised score based on KS2 maths assessments.	KS2	Derived from KS2_GPSOUTCOME KS2_GPSSCORE KS2_WRITTAOUTCOME KS2_WRITMARK KS2_GPSMRK

Table 3.1b: Variable definitions for the ALSPAC Dataset

Variable abbreviation	Variable definition	Variable source	Derivation or specification
I-SEN	A categorical variable describing individuals who have an identified special educational need.	NPD-linked data	This will be derived using the plasca41 variable. See Table 3.2 for a full description.
C&L	A categorical variable describing individuals who have a cognition and learning need	NPD-linked data	We will use the variable plascc62
BESD	A categorical variable describing individuals who have behaviour, emotional, and social difficulties	NPD-linked data	We will use the variable plascc63
C&I	A categorical variable describing individuals who have a communication and interaction impairment	NPD-linked data	We will use the variable plascc64
S&P	A categorical variable describing individuals	NPD-linked data	We will use the variable plascc65

	who have sensory and/or physical needs		
STAT	A categorical variable describing individuals who have a statement of SEN (0=no, 1 = yes)	NPD – linked data	This will be derived using the plasca41 variable. Fields populated with “Statement of SEN” will be coded as 1. Those coded as “SEN without statement” will be coded as 0.
S-SPE	A categorical variable describing individuals who attend (or have attended) a special school.	Parent-report	This will be derived using a number of parent report variables where they were asked if their child attends a special school: ● KM4164 – Child attends special school ● KP1221 – Child attends special school ● KR560 – type of school child attends ● KU020, KU030, KU040 – child goes to special school ● KV9000 – type of school child attends
VIOB	A continuous variable describing self-reported violent behaviour at age 17-18 years.	TF4	We will derive this variable from the following items: ● AS1038: Number of occasions YP used threats or actual force or violence against the other person when stole money or property. ● AS1041: Number of times during last year YP hurt or injured any animals or birds on purpose. ● AS1047: During last year YP carried knife or other weapon for protection or in case it was needed in a fight

			AS1049: YP actually used a weapon against somebody in the last year.
NVIQ	This is a continuous variable describing the young person's non-verbal IQ at age 8 years.	F8	We will use the f8ws111 variable.
REC	This is a continuous variable describing the young person's receptive language abilities at age 8 years	F8	We will use the f8sl040 variable.
EXP	This is a continuous variable describing the young person's expressive language abilities at age 8 years.	F8	We will derive this variable from the following variables: f8sl051 f8sl052 f8sl053 f8sl054 f8sl055 f8sl056 f8sl057 f8sl058 f8sl059 f8sl060.
STM	This is a continuous variable describing the young person's short-term memory at age 8 years.	F8	We will use the f8s1105 variable.
REA	This is a continuous variable describing the young person's reading ability at age 9 years	F9	We will derive this variable using the following variables f9mw031 f9mw061 (word reading and non-word reading).
PRAG	This is a continuous variable describing the young person's pragmatic language abilities at age 9 years	KU	We will use the ku510a variable.
AUT	This is a continuous variable describing the young person's autistic characteristics at age 9 years.	KR	We will use the kr554a variable
EMO	This is a continuous variable describing the	KU	We will use the ku707a variable.

	young person's emotional problems at age 9 years.		
CON	This is a continuous variable describing the young person's conduct problems at age 9 years.	KU	We will use the ku708a variable.
HYP	This is a continuous variable describing the young person's hyperactivity/inattention at age 9 years.	KU	We will use the ku706a variable.
PEE	This is a continuous variable describing the young person's peer problems at age 9 years.	KU	We will use the ku709a variable.
PRO	This is a continuous variable describing the young person's prosocial behaviour.	KU	We will reverse score the ku705a variable.
BUL-V	This is a continuous variable describing the extent to which the young person has been the victim of bullying at age 8 years.	F8	We will derive this variable using the following variables: f8fp140, f8fp150, f8fp160, f8fp170, f8fp180, f8fp330, f8fp340 , f8fp350 ,f8fp360.
BUL-P	This is a continuous variable describing the extent to which the young person has been the perpetrator of bullying at age 8 years.	F8	We will derive this variable using the following variables: f8fp240 , f8fp250 , f8fp260 , f8fp270 , f8fp280, f8fp410, f8fp420 ,f8fp430.

3.2. Measurement of key concepts

Table 3.2 Measurement of key concepts

Concept	How the concept will be measured and encoded
---------	--

Identified Special Educational Needs (DfE-MoJ only)	A child or young person with an identified SEN is one who has any SEN type listed in the “Primary SEN Type” or “Secondary SEN Type” in the school census at any point during primary or secondary school. We will combine all of the categories into a single “SEN” category. This will result in a binary variable (0=no I-SEN, 1 = I-SEN).
Identified Special Educational Needs (ALSPAC only)	A child or young person with an identified SEN is one who has any SEN type listed in plasca41 variable. This will result in a binary variable (0= no SEN, 1 = SEN (with or without statement)).
Unidentified Special Educational Needs	A child or young person with an unidentified SEN is one who would been given an SEN label but was not due to school level or other factors not related to the child’s abilities. It is this implementation variation which is the source of the counterfactual required for the propensity score analysis. Therefore it is a necessary feature of the analysis. We will use an approach we adopted in our previous work (Hutchinson, 2021). Specifically, we will use children’s gender, ethnicity, prior attainment in primary school, their attendance and exclusions histories in upper primary school (KS2), their social care records, and the detailed individual and geographical deprivation measures outlined to identify young people with unidentified SEN. Our previous work shows that it is possible to determine with 95% accuracy children’s SEN identification status without knowing their SEN status. Much of the variation in whether a child is identified as having a SEN is at the school level. It is important to note here, we are interested in those who have been identified vs not identified. This is not necessarily the same group as who have SEN and who do not have SEN. Even without school effects (since we need to find the missing counterfactual cases (unidentified)), there is 70% chance of correct classification of a random pair of cases with one identified and one unidentified. In practice, the classification accuracy will be between 70%-90%.
Statement of Special Educational Needs	A child or young person with a statement of special educational needs is anyone who has an entry of “S= Statement” in the “SENprovision” variable.
Education Health and Care Plan	A child or young person with a statement of special educational needs is anyone who has an entry of “E= Education Health and Care Plan” in the SENprovision variable.
Violent Offending	We will use the Offence Group Coding. In line with the <u>YEF’s guidance</u> , we will define violence as “01 Violence against the person” (all instances). We will also use some instances of “02 Sexual Offences” and

	use the “main offence code” to determine which sexual offences are violent vs non-violent.
Violent Behaviour	We will recode four variables (AS1038, AS1041, AS1047, AS1049) into a binary scale (0=never, 1= all other responses). The four questions will be used to generate a single violent behaviour score of between 0-4, higher scores indicating higher levels of violent behaviour. This will be used as one of the outcome variables.
Disposal Type	We will generate a binary variable using the PNCDisposalCode variable (0=custodial, 1= other).
Caution Type	We will generate a binary variable using the Cautiontype variable (0=unconditional, 1 = conditional).
Re-offending	In line with the MoJ definition, we will count an offence as a re-offence if all of the following criteria are met: • The offence is recordable. We assume that if the offence appears on the PNC it is recordable • The offence was committed in England or Wales. We will exclude offences with an Offence Group code of “21 Offences outside England and Wales” • The offence is prosecuted by the police. The offence is proven through caution, reprimand, or court convictions. If either the DISP or CAUT variable =1. • The offence is not a breach offence. We will exclude offences with an Offence Group code of “23 Breach Offences”
Social Difficulties	These difficulties will captured using the following continuous variables; higher scores indicating more difficulties (see Table 3.1b): • AUT - We will use the kr554a variable • CON - We will use the ku708a variable • HYP - We will use the ku706a variable • PEE - We will use the ku709a variable • PRO - We will reverse score the ku705a variable • BUL-V – We will generate a continuous scale consisting of the following variables f8fp140, f8fp150, f8fp160, f8fp170, f8fp180, f8fp330, f8fp340 , f8fp350 ,f8fp360, which are scored as 0 (no) and 1 (yes). this will result in a derived variable scored - between 0-9 higher scores indicating more instances of bullying victimisations. • BUL-P - We will generate a continuous scale consisting of the following variables f8fp240 , f8fp250 , f8fp260 , f8fp270 , f8fp280, f8fp410, f8fp420 ,f8fp430, which are scored as 0 (no) and 1 (yes).

	this will result in a derived variable scored - between 0-9 higher scores indicating more instances of bullying victimisation.
Emotional Difficulties	We will use the ku707a variable.
Cognitive and Learning Difficulties	These difficulties will captured using the following continuous variables; higher scores indicating fewer difficulties (see Table 3.1b): • NVIQ- we will use the f8ws111variable • REC - We will use the f8sl040 variable • EXP - We will generate a continuous scale consisting of the following variables f8sl051 f8sl052 f8sl053 f8sl054 f8sl055 f8sl056 f8sl057 f8sl058 f8sl059 f8sl060, which are scored as 0 (incorrect) and 1 (correct). This will result in a derived variable scored - between 0-10, higher scores indicating better expressive language. • STM - We will use the f8s1105 variable • REA - We will derive this variable using the following variables f9mw031 f9mw061, both of which are scored on a scale of 0 to 10. The resulting variable will be on a scale of 0 to 20; higher scores indicating higher reading ability. • PRAG - We will use the ku510a variable.

3.3. Missing data and attrition

As with all existing datasets, both datasets have missing data.

ALSPAC. Not all families took part in each wave of data collection. There was selective attrition in ALSPAC. Specifically, families in the following demographic groupings were more likely to drop out: having a single mother, mother with no educational qualifications, mother encountered financial difficulties, child being raised in a large family, mother who smoked, mother who had a poor relationship with the partner, family living in poor housing, mother had been involved in crime and been convicted, or suffered poor mental health during pregnancy (Wolke et al., 2009). Where feasible, we will use multiple-imputation/full information maximum likelihood to deal with missing data and re-run the main statistical models with the imputed data as a form of sensitivity analysis.

DfE-MoJ datasets. Given these datasets are administrative we expect there to be population coverage. The missing data are likely to arise from poor recording of data at the point of data collection.

3.4. Other sources of bias

SEN Identification Bias

SEN identification is a postcode lottery (Hutchinson, 2021). Therefore, children who have been identified as having a SEN may only represent a sub-sample of those who actually

have a SEN. We know that biases exist based on month of birth and possibly due to ethnicity, gender, and socioeconomic status. Known biases like this can be controlled for during our propensity score analysis and also by running the analyses separately for groups of interest (e.g., separately for boys and girls). Additionally, our analysis of children's underlying social, emotional, cognitive, and learning skills (RQ2) to determine the impact of (non) identification will address this SEN identification bias to some extent.

Racial and Socioeconomic Disproportionality

The ALSPAC dataset primarily consists of white families who are more affluent than the average UK family. This presents a problem as it limits the generalisability of the findings of RQ2 to ethnically and socioeconomically diverse families.

Age of Criminal Responsibility

Violent offending will only be apparent in the datasets after the age of 10 years. This is the age of criminal responsibility in England. This is an overarching limitation of this project, and any project relying upon criminal convictions data as it limits the extent to which causal effect can be determined. To some extent it is a chicken and egg situation. There is, however, a difference between aggressive behaviours in childhood, which are likely to be poor emotional or behavioural regulation rather than intentional criminality. Very young children will unlikely have the cognitive capacity to engage in intentional aggressive behaviour. Therefore, this limitation should be considered with this caveat in mind.

Offence Severity

We have narrowed our focus to violent offending. But this is still a broad measures of offending and issues around offence severity still exist. For example, two individuals who have been convicted of a violent offence might receive different sentences because of issues such as culpability, harm caused, and aggravating/mitigating factors. We will not be able to control for these factors in the analysis as they are not available in the datasets.

4. About the analysis

4.1. Overview of analytical approach

We will investigate the relationship between SENs and violent youth offending in England. We will do this by answering a number of research questions focussed on descriptive statistics (RQ1-RQ4), relationship testing (RQ1, RQ3; DfE-MoJ, RQ2, ALSPAC) and causal inference modelling (RQs3;DfE-MoJ).

4.2. Approach to addressing research question(s)

Research question 1: approach and methods

<i>Research question</i>	Are there differences between youth with and without identified SENs based on: a. Age of first violent offence,
--------------------------	--

	b. Number of violent offences (single offence vs multiple offences), c. Caution for first violent offence (no or yes), d. Disposal type (custodial sentence vs other).
<i>Hypothesis, if relevant</i>	The hypotheses for each of the sub-research questions are: a. Youth with SENs will become involved in violent offending earlier than those without SENs, b. Youth with SENs will be more likely to engage with multiple violent offences than those without SENs, c. We expect youth with SENs to be less likely to given a caution than those without SENs, d. We expect youth with SENs to be more likely to receive a custodial sentence than other types
<i>What will you be able to say by the interim report</i>	We will be able to provide a high level summary of the findings.
<i>Descriptive analysis, if relevant</i>	We will produce descriptive statistics for the following variables from Table 3.1a: ● Identified SEN (I-SEN) ● Violent offending (VO) ● Age of violence offence (VO-AGE) ● Number violent offences (NO-VO) ● Caution type (CAUT) ● Disposal type (DISP) We will produce sub-group descriptive statistics for VO, VO-AGE, NO-VO, CAUT, DISP for all of the SEN types (I-SEN, LD-PMLD, SENS, PD, OTH-SEN, ASD, SLCN, SEN-NSA, SPLD, MLD, SEMH-BESD) We will also produce descriptive statistics for the following variables to describe the sample: ● Month of birth ● Ethnicity ● Gender ● Number of years eligible for free school meals ● English as an Additional Language ● Special school (S-SPE) ● Pupil referral unit (S-PRU) ● Alternative provision (S-AP)

	● Mainstream school with SEN unit (S-MSS) ● Mainstream school with resourced provision (S-MSR) ● Statement of SEN (STAT) ● EHCP ● Looked after status (CLA) ● Children In Need (CIN) ● Child Protection Plan (CPP) ● All of the SEN types (SLD-PMLD, SENS, PD, OTH-SEN, ASD, SLCN, SEN-NSA, SPLD, MLD, SEMH-BESD).
<i>Models, specifications and statistical techniques used, if relevant</i>	Logistic regression models will be fitted; one for each of the sub-research questions. Initially, the dependent variable in all models will be violent offending (VO) and the independent variable will be one of the following: ● VO-AGE, I-SEN, VO-AGE x I-SEN interaction ● NO-VO, I-SEN, NO-VO x I-SEN interaction ● CAUT I-SEN, CAUT x I-SEN interaction ● DISP I-SEN, DISP x I-SEN interaction Following this, we will add covariates to the models and re-run them. We expect these will be year of birth, ethnicity, gender, school type, STAT_EHCP, and CLA, but we will also consider others from the list described in the descriptive statistics.
<i>Estimating equation, if relevant</i>	$Y = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \beta_4 C_3 + \beta_5 C_4 + \beta_6 C_5 + \epsilon$ Where: ● Y is the dependent variable ● X is the independent variable ● C₁, C₂, C₃, C₄, and C₅ are the covariates ● β_0 is the intercept ● β_1 is the coefficient for the predictor variable X ● $\beta_2, \beta_3, \beta_4, \beta_5$, and β_6 are the coefficients for the covariates C₁, C₂, C₃, C₄, and C₅ respectively ● ϵ is the error term
<i>What does the approach need to succeed (constraints/assumptions)?</i>	We will test the assumptions for a logistic regression model. The key ones applicable to our analysis are: ● Linearity of logit ● Independent observations ● No multicollinearity

<i>Uncertainty and inference</i>	We will use an alpha level of 0.01 and 95% confidence intervals to infer uncertainty around our estimates and significance. Given that we are working with large datasets and that we will conduct a number of sub-group analyses, we recognise the need to correct for multiple testing to avoid a type 1 error. A family wise correction was not deemed appropriate given the increased risk of type 2 error. We, therefore, will adopt an alpha level of .01 (instead of .05) for all analysis in this project. This is our blanket adjustment for multiple testing.⁴
<i>Robustness checks</i>	We will attempt to impute missing data using multiple-imputation and re-run the models. We recognise that this may be computationally intense and so not possible.
<i>Subgroup you intend to study</i>	We will re-run the models separately for the following groups: • Gender (2 levels) • Ethnicity (5 levels) – if sample size allows, we will use more fine grained levels (e.g., separate Black into Black African and Black Caribbean) • FSM (2 levels) - if sample size allows, we will use 6 levels – quintiles based on number of years eligible plus never) • SLD-PMLD (2 levels) • SENS (2 levels) • PD (2 levels) • OTH-SEN (2 levels) • ASD (2 levels) • SLCN (2 levels) • SEN-NSA (2 levels) • SPLD (2 levels) • MLD (2 levels) • SEMH-BESD (2 levels)
<i>Changes to the analysis</i>	If the data violate assumptions, the models do not converge, or some other analytic problem arises, our first back-up position will be to use a non-parametric approach

⁴ This also applies to research questions 2-4.

	(without covariates) such as a chi-squared test. Failing this, we will stop at producing descriptive statistics for the VO, VO-AGE and DISP variables split by I-SEN.
--	---

Research question 2: approach and methods

<i>Research question</i>	To what extent do social, emotional, cognitive, and learning skills deficits in childhood predict violent behaviour in adolescence?
<i>Hypothesis, if relevant</i>	We expect social, emotional, cognitive, and learning skills deficits to predict violence to a greater or lesser extent.
<i>What will you be able to say by the interim report</i>	We will be able to provide brief summary of findings and summarise results in tables/figures
<i>Descriptive analysis, if relevant</i>	We will produce descriptive statistics for the following variables from Table 3.1b: ● Violent behaviour (VIOB) ● Identified SEN (I-SEN) ● Non-verbal IQ (NVIQ) ● Receptive language (REC) ● Expressive language (EXP) ● Short term memory (STM) ● Reading (REA) ● Pragmatic language (PRAG) ● Autistic characteristic (AUT) ● Emotional problems (EMO) ● Conduct problems (CON) ● Hyperactivity/inattention (HYP) ● Peer problems (PEE) ● Prosocial behaviour (PRO) ● Bullying victimisation (BUL-V) ● Bullying perpetration (BUL-P) We will produce sub-group descriptive statistics for the variables listed above for the two-levels of the following variables ● I-SEN (0=no SEN, 1=SEN). ● C&L (0=no C&L, 1= C&L) ● BESD (0=no C&L, 1= C&L) ● C&I (0=no C&L, 1= C&L) ● S&P (0=no C&L, 1= C&L)

	We will also produce descriptive statistics for the following variables to describe the sample: ● Ethnicity (2 levels) ● Gender (2 levels) ● C&L ● BESD ● C&I ● S&P ● STAT ● S-SPE
<i>Models, specifications and statistical techniques used, if relevant</i>	We will fit a series of linear regression models following a three-step process. The dependent variable in all models will be violent behaviour (VIOB). ● In step 1, we will fit a series linear regression models with one predictor per model only. This will allow us to test the crude associations between deficits in social, emotional, cognitive, and learning skills and violent behaviour. ● In step 2, the predictors that were significant in step 1, will be entered into a single model to test the relationships between deficits in social, emotional, cognitive, and learning skills and violent behaviour after controlling for all other social, emotional, cognitive, and learning skills deficits. ● In step 3, we will repeat the model from step 2 and include the following covariates: ethnicity, gender, FSM, identified SEN, statement of SEN, and special school.
<i>Estimating equation, if relevant</i>	$Y = \beta_0 + \beta_1 X_1 + \dots \beta_n X_n + \beta_{n+1} C_1 + \beta_{n+2} C_2 + \beta_{n+3} C_3 + \beta_{n+4} C_4 + \beta_{n+5} C_5 + \epsilon$ Where: ● Y is the dependent variable ● X_1 is the first independent variable ● X_n is the n^{th} independent variable ● C_1, C_2, C_3, C_4, and C_5 are the covariates ● β_0 is the intercept ● β_1 is the coefficient for the predictor variable X ● β_n is the coefficient for the n^{th} predictor variable ● $\beta_{n+2}, \beta_{n+3}, \beta_{n+4}$, and β_{n+5} are the coefficients for the covariates C_1, C_2, C_3, C_4, and C_5 respectively ● ϵ is the error term

<i>What does the approach need to succeed (constraints/assumptions)?</i>	We will test the assumptions for a linear regression model. They key ones applicable to our analysis are: • Linearity • Independent observations • Homoscedasticity • Normality of errors • No multicollinearity
<i>Uncertainty and inference</i>	We will use an alpha level of 0.01 and 95% confidence intervals to infer uncertainty around our estimates and significance.
<i>Robustness checks</i>	We will attempt to impute missing data using multiple-imputation and re-run the models. We will also consider alternative model specifications to account for the limitations associated with a step-wise regression approach (e.g., a lasso regression)
<i>Subgroup you intend to study</i>	We will re-run the models separately for the following groups: • Gender (2 levels) • Identified SEN (2 levels) • C&L (2 levels) • BESD (2 levels) • C&I (2 levels) • S&P(2 levels)
<i>Changes to the analysis</i>	If the data violate assumptions, the models do not converge, or some other analytic problem arises, our back-up position will be to use a non-parametric approach (without covariates) such as a Mann-Whitney U Test.

Research question 3: approach and methods

<i>Research question</i>	To what extent do the following lead to reduced violent offending for youth with SENs: a. Identification of SEN (identified vs unidentified), b. Timing of SEN identification (relative to the typical time of identification for that SEND need type), c. Consistency of SEN label during school transition (consistent vs inconsistent)
--------------------------	--

	d. Having a statement of SEN/education health and care plan (EHCP versus SEND support), e. Type of school provision: i. special vs any mainstream (with an EHCP), ii. alternative provision vs any mainstream (with an EHCP), iii. mainstream with a SEN unit vs mainstream with neither a SEN unit nor resourced provision (with EHCP) iv. mainstream with a SEN unit vs mainstream with neither a SEN unit nor resourced provision (without EHCP) v. mainstream with resourced provision vs mainstream with neither resourced provision nor a SEN unit/mainstream without SEN unit other (with EHCP) vi. mainstream with resourced provision vs mainstream with neither resourced provision nor a SEN unit (without EHCP).
<i>Hypothesis, if relevant</i>	1. We expect those with identified SENs to be less likely to be involved in violent offending than those with unidentified SEN 2. We expect there to be differences in outcomes depending on need and timing of identification but are unable to predict the direction of these. 3. We expect those who lose their SEN label in the transition from primary to secondary school, or have a different label, will be more likely to be involved in violent offending 4. We expect there to be differences in outcomes for those with and without a statement of SEN/EHCP but the direction of these to vary depending on the type of SEN 5. We expect those attending specialist provision (of any sort) to be less likely to be involved in violent offending.
<i>What will you be able to say by the interim report</i>	No results will be included in interim report

<i>Descriptive analysis, if relevant</i>	We will produce descriptive statistics for the following variables from Table 3.1a: • Violent offending (VO) • Timing of SEN identification (T-SEN) • Consistency of SEN identification (C-SEN) • EHCP/Statement (STAT_EHCP) • Special school (S-SPE) • Mainstream school with SEN unit (S-MSS) • Mainstream school with resourced provision (S-MSR) • Pupil referral unit (PRU) • Alternative provision (AP) • SEN Type (SLD-PMLD, SENS, PD, OTH-SEN, ASD, SLCN, SEN-NSA, SPLD, MLD, SEMH-BESD) We will produce descriptive statistics for the following variables listed above to describe the sample: • Ethnicity (5 levels) • Gender (2 levels) • FSM (6 levels) – quintiles plus never • C&L (2 levels) • BESD (2 levels) • C&I (2 levels) • S&P (2 levels) • STAT (2 levels)
<i>Models, specifications and statistical techniques used, if relevant</i>	We will compute propensity scores and test the hypothesis for our hypothetical intervention and control groups. To calculate propensity scores for the sample, we will fit a series of logistic regression models in which the dependent variables are a series of binary contrasts between: [No special needs] in a mainstream school with neither SEN Unit nor Resourced Provision vs [School Action Plus or School Support] in a mainstream school with neither SEN Unit nor Resourced Provision [School Action Plus or School Support] in a mainstream school with neither SEN Unit nor Resourced Provision vs [School Action Plus or School Support] in a mainstream school with Resourced Provision

	c. [School Action Plus or School Support] in a mainstream school with neither SEN Unit nor Resourced Provision vs [School Action Plus or School Support] in a mainstream school with a SEN Unit d. [School Action Plus or School Support] in a school with a SEN Unit vs [School Action Plus or School Support] in an AP or PRU school e. [School Action Plus or School Support] in a school with Resourced Provision vs [School Action Plus or School Support] in an AP or PRU school f. [Statement or EHCP] in a mainstream school with neither SEN Unit nor Resourced Provision vs [Statement or EHCP] in a school with Resourced Provision g. [Statement or EHCP] in a mainstream school with neither SEN Unit nor Resourced Provision vs [Statement or EHCP] in a school with a SEN Unit h. [Statement or EHCP] in a school with a SEN Unit vs [Statement or EHCP] in a Special School i. [Statement or EHCP] in a school with Resourced Provision vs [Statement or EHCP] in a Special School j. [Statement or EHCP] in a school with a SEN Unit vs [Statement or EHCP] in an AP or PRU school k. [Statement or EHCP] in a school with Resourced Provision vs [Statement or EHCP] in an AP or PRU school l. [School Action Plus or School Support] for < 6 years] in a mainstream school with neither SEN Unit nor Resourced Provision vs [School Action Plus or School Support for >6 years] in a mainstream school with neither SEN Unit nor Resourced Provision m. [School Action Plus or School Support] in year 6 but not in year 8] in a mainstream school with neither SEN Unit nor Resourced Provision vs [School Action Plus or School Support in year 6 and in year 8] in a mainstream school with neither SEN Unit nor Resourced Provision n. [Statement or EHCP in year 6 but not in year 8] in a mainstream school with neither SEN Unit nor Resourced Provision vs [Statement or EHCP in year 6 and in year 8] in a mainstream school with neither SEN Unit nor Resourced Provision
--	---

- o. [School Action Plus or School Support identified late*] in a mainstream school with neither SEN Unit nor Resourced Provision vs **[School Action Plus or School Support identified at typical time point*] in a mainstream school with neither SEN Unit nor Resourced Provision**
- p. [School Action Plus or School Support identified at typical time point*] in a mainstream school with neither SEN Unit nor Resourced Provision vs **[School Action Plus or School Support identified early*] in a mainstream school with neither SEN Unit nor Resourced Provision**
- q. [Statement or EHCP first recorded late*] in a mainstream school with neither SEN Unit nor Resourced Provision vs **[Statement or EHCP first recorded at typical time point*] in a mainstream school with neither SEN Unit nor Resourced Provision**
- r. [Statement or EHCP first recorded at typical time point*] in a mainstream school with neither SEN Unit nor Resourced Provision vs **[Statement or EHCP first recorded early*] in a mainstream school with neither SEN Unit nor Resourced Provision**

* early, typical and late will be defined relative to other children with the same primary type of need at the time of identification or first recorded with EHCP.

We likely use the following co-covariates: birth year, month of birth, gender, ethnicity, English as additional language, reason for child in need status, highest level of social care intervention, free school meals history, income deprivation affecting children index, index of multiple deprivation domains, school attendance, school mobility, suspensions, expulsions, early years foundation stage profile results for each early learning goal, key stage 1 results, key stage 2 results, key stage 1 p scales, and SEN need type for contrasts b-r. These covariates are indicative and will likely change as we start to work with the data.

Once we have calculated propensity scores, we will assess the common support for each contrast a-r by plotting the kernel densities for each pair of treatment (or more treatment) and control (or less treatment). We will make any necessary changes to the scope of the matched groups and discard any treatment contrasts for which there is

	insufficient common support. The next step will be to match intervention young people with controls for each retained contrast. We will begin by testing nearest neighbour matching and assess the balance using standardised mean differences between the resulting pairs of treatment vs control groups. We will then make adjustments to the matching algorithms to address any problems of remaining bias relating to particular covariates from the logistic regressions. When we are satisfied with the balance, we will then use the treatment and control groups to estimate the effects of the treatments on our outcomes. To maximise the robustness of the estimation we will enter the treatment indicator and the covariates from the propensity models into regression models to estimate the effect of each treatment on the outcomes (doubly-robust PSM). We recognise that PSM has some drawback, such omitted variable bias. We will be able to address some of this by including a wide range of covariates that influence both treatment and outcome. We have previously adopted such an approach in this dataset and have found high levels of common support. Therefore, whilst we recognise that this is a concern, we are confident that we will be able to minimise the impact of omitted variable bias through the inclusion of a wide range of covariates. The outcomes we will estimate treatment effects for are: • VO-AGE • NO-VO • CAUT • DISP
<i>Estimating equation, if relevant</i>	$E(Y_1 - Y_0 \mid T=1) = E[Y_1 \mid X, T=1] - E[Y_0 \mid X, T=0] - E[Y_0 \mid X, T=1] + E[Y_0 \mid X, T=0]$ Where: Y = the outcome X = the matching covariate set T = the treatment condition
<i>What does the approach need to succeed (constraints/assumptions)?</i>	Conditional Independence Assumption (i.e. that the assignment to treatment is independent of the outcomes, given the covariate set that influences both). All

	confounding variables must be observed / any unobserved variables play no role in assignment to the treatment condition. Common support exists between the treatment and control groups, i.e. young people exist with the same covariate values (X) in both treatment and control groups.
<i>Uncertainty and inference</i>	Variance of the treatment effect will be estimated through non-parametric bootstrapping to account for the additional variance from estimating p (propensity to be treated). 95% confidence intervals will be applied to estimates and significance tests.
<i>Robustness checks</i>	We will attempt to impute missing data using multiple-imputation and re-run the models. We will use Rosenbaum Bounds to estimate how large a bias (from an unobserved confounding variable) would be required to undermine the inferences from the PSM estimates. As the question of whether there is unobserved bias is untestable, this approach instead informs us of how sensitive any treatment effects are to the possibility of such bias existing.
<i>Subgroup you intend to study</i>	We will run subsample analyses on the following groups for SEN treatment contrasts where significant effects of treatment are found: -Boys -Black young people -White young people -Mixed Black and White Ethnicity young people -Different cohorts by year of birth -Young people with school attendance in the top quartile - for that treatment contrast -Young people with no suspensions from school
<i>Changes to the analysis</i>	Failure of common support will be addressed by limiting to scope of the sample in the estimation, or where this is insufficient, by dropping that particular contrast from the analysis. It is expected that this may occur for some contrasts, but similar analysis in the HOPE Study has found common support for closely paired contrasts and our previous modelling of SEND identification achieved very good classification properties. However, since unexpected common support problems may still occur, we have ensured the contrasts compare

	each treatment group to the next most similar group as common support is most likely to fail when comparing very different groups (e.g. comparing young people in special schools with those with no identified SEND). Model convergence problems are likely to reflect a problem with the specification of the covariates; this is less likely to occur since we have a lot of experience of modelling with these covariates and SEND variables as the dependent variable. However, simplifying the covariate matrix has been the answer to previous convergence issues.
--	---

Research question 4: approach and methods

<i>Research question</i>	What is the impact of cautions on re-offending in youth with identified SENs involved in violent offending. Specifically: - Does being cautioned, after first offence, reduce violent re-offending for youth with SENs? (youth with identified SENs cautioned vs youth with identified SENs charged) - Is there a difference in re-offending rates for youth with SENs depending on the type of caution? (unconditional vs conditional)
<i>Hypothesis, if relevant</i>	The hypotheses for each of the sub-questions are: - We expect youth with SENs who are cautioned to be less likely to reoffend than youth with SENs who are charged with an offence - We expect youth with SENs who are given a conditional caution to be more likely to reoffend compared youths with SENs who are given unconditional cautions.
<i>What will you be able to say by the interim report</i>	No results will be included in interim report
<i>Descriptive analysis, if relevant</i>	We will produce descriptive statistics for the following variables from Table 3.1a: - number of violent offences (NO-VO) - child looked after (CLA) - caution (CAUT) We will also produce descriptive statistics for the following variables to describe the sample: - Ethnicity

	• Gender • Special school (S-SPE) • Mainstream school with SEN unit (S-MSS) • Mainstream school with resourced provision (S-MSR) • Statement of SEN/EHCP (EHCP_STAT) • SEN Type (SPLD, MLD, SLD, PMLD, SEMH-BESD, SLCN, HI, VI, MSI, PD, ASD, OTH, SEN-NSA)
<i>Models, specifications and statistical techniques used, if relevant</i>	We will fit a series of linear regression models following a three-step process. The dependent variables in the models will be: • NO-VO-POST • NO-VO-CLA • In step 1, we will fit a linear regression model with only predictors for CLA-ASB DISP and CAUT. • In step 2, we will fit a series of linear regression models test a single additional predictor at a time for birth year, month of birth, gender, ethnicity, English as additional language, reason for child in need status, highest level of social care intervention, social care referral followed by no action, free school meals history, income deprivation affecting children index, index of multiple deprivation domains, school attendance, school mobility, suspensions, expulsions, early years foundation stage profile results for each early learning goal, key stage 1 results, key stage 2 results, key stage 1 p scales, and SEN need type, measuring these prior to the first offence. • In step 3, for the predictors that were significant in step 2, these will be entered into a single model to test the relationships between the justice and care interventions and violent offending outcomes.
<i>Estimating equation, if relevant</i>	$Y = \beta_0 + \beta_1 X_1 + \dots \beta_n X_n + \beta_{n+1} C_1 + \beta_{n+2} C_2 + \beta_{n+3} C_3 + \beta_{n+4} C_4 + \beta_{n+5} C_5 + \epsilon$ Where: • Y is the dependent variable • X_1 is the first independent variable • X_n is the n^{th} independent variable • C_1, C_2, C_3, C_4, and C_5 are the covariates • β_0 is the intercept • β_1 is the coefficient for the predictor variable X

	• β_n is the coefficient for the n^{th} predictor variable • $\beta_{n+2}, \beta_{n+3}, \beta_{n+4},$ and β_{n+5} are the coefficients for the covariates $C_1, C_2, C_3, C_4,$ and C_5 respectively ϵ is the error term
<i>What does the approach need to succeed (constraints/assumptions)?</i>	We will test the assumptions for a linear regression model. They key ones applicable to our analysis are: • Linearity • Independent observations • Homoscedasticity • Normality of errors No multicollinearity
<i>Uncertainty and inference</i>	We will use an alpha level of 0.01 and 95% confidence intervals to infer uncertainty around our estimates and significance.
<i>Robustness checks</i>	We will attempt to impute missing data using multiple-imputation and re-run the models.
<i>Subgroup you intend to study</i>	We will re-run the models separately for the following groups: • I-SEN (2 levels) • C&L (2 levels) • BESD (2 levels) • C&I (2 levels) • S&P (2 levels) • Gender (2 levels) • Ethnicity (5 levels) – if sample size allows, we will use more fine grained levels (e.g., separate Black into Black African and Black Caribbean) • Different cohorts by year of birth
<i>Changes to the analysis</i>	If the data violate assumptions, the models do not converge, or some other analytic problem arises, our back-up position will be to use a non-parametric approach (without covariates) such as a Mann-Whitney U Test. If the models indicate that there are statistically significant associations (potential treatment effects) between any of CLA-ASB DISP and CAUT and their associated NO-VO outcomes, then these models will be converted into PSM analyses following the model set out for RQ3 to see whether common support and balance can be established to test for treatment effects.

5. Project management

5.1. Risks and mitigations

Table 5.1 Risks and mitigations

Number	Risk	Likelihood (Low/Medium/ High)	Mitigation
1	Delay in data access. For the ALSPAC dataset, we have already gained approval to access. The legal agreements have been signed and the costs have been settled. Data access will be granted imminently. For the MoJ-DfE linked dataset, we have submitted a completed research application; it has been reviewed by the Data Access Group and Data Sharing Approval Panel. We have responded to the comments and had a zoom call with DfE who are happy with our responses. The proposal now needs to be considered by the Data Access Governance Board.	Low	If access to the MoJ-DfE linked dataset is delayed, we can start work on the ALSPAC dataset (RQ2). This will take approximately 3 months. Therefore, data access for DfE-MoJ datasets can be delayed until approximately December 2024 without affecting our ability to meet the requirements of the interim report.

5.2. Timeline

Table 5.2 Timeline

Date	Activity	Staff responsible/leading
w/c 19.09.24	Analysis plan approved by YEF	Umar Toseeb

w/c 19.09.24	Data access approved by DfE-MoJ	Umar Toseeb
w/c 19.09.24	Data access approved by ALSPAC	Umar Toseeb
01.09.24-01.01.25	Determine patterns of missing data	Megan Frith/Emre Deniz
01.09.24-01.03.25	Analysis for RQ1	Megan Frith/Emre Deniz
01.09.24-01.03.25	Analysis for RQ2	Megan Frith/Emre Deniz
01.01.23-01.03.25	Produce interim report	Umar Toseeb
01.03.25-01.07.25	Analysis for RQ3	Megan Frith /Emre Deniz
01.03.25-01.07.25	Analysis for RQ3	Megan Frith/Emre Deniz
01.07.25-01.09.25	Draft academic paper	Whole team
w/c 01.09.25	Finalise academic paper	Umar Toseeb
01.09.25-30.11.25	Revise analysis for RQs1-4	Megan Frith/Emre Deniz
01.09.25-30.11.25	Draft executive summary	Whole team
01.09.25-30.11.25	Draft final report	Whole team
30.11.25	Finalise and submit executive summary and final report	Umar Toseeb

6. References

- Baldry, E., Briggs, D., Goldson, B. and Russell, S. (2017) 'Cruel and unusual punishment': and inter-jurisdictional study of the criminalisation of young people with complex support needs.' *Journal of Youth Studies*, Pp 1-17
- Centre for Justice and Innovation (2024) How is youth diversion working for children with special educational needs and disabilities?, <https://justiceinnovation.org/publications/how-youth-diversion-working-children-special-educational-needs-and-disabilities>
- Coie JD, Dodge KA. Aggression and antisocial behavior. In: Damon W, Eisenberg N, editors. *Handbook of child psychology: Social, emotional, and personality development*. Vol. 3. New York: Wiley; 1998. pp. 779–862.
- Crime Prosecution Service (2022) Cautioning and Diversion – Legal Guidance. <https://www.cps.gov.uk/legal-guidance/cautioning-and-diversion>
- Cronin P, Addo R. Interactions with youth justice and associated costs for young people with speech, language and communication needs. *Int J Lang Commun Disord*. 2021 Jul;56(4):797-811. doi: 10.1111/1460-6984.12628. Epub 2021 May 27. PMID: 34042241.
- Department for Education (2022). Education, children's social care and offending: Descriptive Statistics. https://assets.publishing.service.gov.uk/media/6227a9b58fa8f526dcf89e17/Education_childrens_social_care_and_offending_descriptive_stats_FINAL.pdf
- Department for Education (2022), Suspensions and permanent exclusions data, <https://explore-education-statistics.service.gov.uk/find-statistics/suspensions-and-permanent-exclusions-in-england>
- HM Inspectorate of Probation (2023) Out of court disposals. <https://www.justiceinspectorates.gov.uk/hmiprobation/research/the-evidence-base-youth-offending-services/specific-types-of-delivery/out-of-court-disposals/>
- Hughes N, Ungar M, Fagan A, Murray J, Atilola O, Nichols K, Garcia J, Kinner S. Health determinants of adolescent criminalisation. *Lancet Child Adolesc Health*. 2020 Feb;4(2):151-162. doi: 10.1016/S2352-4642(19)30347-5. Epub 2020 Jan 16. PMID: 31956016.
- Hutchinson, J (2021) Identifying pupils with special educational needs and disabilities. https://epi.org.uk/wp-content/uploads/2021/03/SEND-Identification_2021-EPI.pdf
- Jessiman T, Cameron A. The role of the appropriate adult in supporting vulnerable adults in custody: Comparing the perspectives of service users and service providers. *British Journal of Learning Disabilities*. 2017; 45: 246–252. <https://doi.org/10.1111/bld.12201>
- Senned (2023) Is nobody listening? Speech, language, and communication needs in youth justice. <https://research.senedd.wales/research-articles/is-nobody-listening-speech-language-and-communication-needs-in-youth-justice/>
- Shaw DS, Giliom M, Giovannelli J. Aggressive behavior disorders. In: Zeanah CH Jr, editor. *Handbook of infant mental health*. Vol. 2. New York: Guilford Press; 2000b. pp. 397–398
- Taylor, C (2016) 'Review of the Youth Justice System in England and Wales', [online] London: Ministry of Justice. Pp 18- 80. Available at: Review of the Youth Justice System (publishing.service.gov.uk) Accessed 15 December 2021
- Tremblay RE, Daniel S, Nagin J, Seguin R, Zoccolillo M, Zelazo PD, Boivin M, Pérousse

D, Japel C. Physical aggression during early childhood: Trajectories and predictors. *Pediatrics*. 2004;**114**:e43–e50.

Wolke D, Waylen A, Samara M, Steer C, Goodman R, Ford T, Lamberts K. Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. *Br J Psychiatry*. 2009 Sep;195(3):249-56. doi: 10.1192/bjp.bp.108.053751. PMID: 19721116; PMCID: PMC2802508.