

Youth Criminal Justice Outcomes and UK Policy

Secondary Data Project Analysis Plan

Evaluating institution: University of Warwick and CEP, LSE

Principal investigator(s): Nikhil Datta, Rui da Costa



Analysis Plan for YEF Secondary Data Analysis Projects

Project summary

Project title	<i>Youth Criminal Justice Outcomes and UK Policy</i>
Research Team	<i>University of Warwick, Centre for Economic Performance – London School of Economics</i>
Principal investigator	<i>Nikhil Datta</i>
Analysis plan author(s)	<i>Nikhil Datta, Rui da Costa, Matteo Sandi</i>
Overarching research question	<i>How have policy changes in the UK criminal justice policy landscape affected youth outcomes?</i>
Supporting research question(s)	1. What is the relationship between the use of diversion and the likelihood of recidivism among young people who have contact with the criminal justice system? 2. How have structural changes in police forces and the justice system contributed to the use of diversion? What were the consequences for recidivism in the most affected areas? 3. How did the increased use of diversion affect spatial and demographic disparities in criminal and justice outcomes? Have the aforementioned structural changes narrowed or widened these existing inequalities?
Dataset(s) to be used	<i>DfE-MoJ data linkage 2001-21</i>
Population characteristics	<i>Universe of Pupils in State-Maintained schools in England from 2001-21</i>
Years data spans	<i>2001-21</i>
Geographic coverage	<i>England</i>

Primary outcome(s) investigated	<i>Criminal Justice Outcomes</i> <i>Recidivism</i> <i>Court waiting times</i> <i>Commute to court time</i>
Main method(s) to be used or tested	<i>Staggered Difference-in-difference, triple-difference</i>

Analysis plan history

Version	Date	Reason for revision
1.X [<i>latest</i>]		
1.1		
1.0 [<i>original</i>]		<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).....	4
1.3. Hypotheses.....	8
1.4. Key concepts.....	9
2. About the datasets.....	11
2.1. Overview of datasets used	11
2.2. Secondary data source(s).....	12
2.3. Primary data collection.....	13
2.4. Linking datasets.....	21
2.5. Access and data protection.....	23
3. About the data	24
3.1. List of variables	24
3.2. Measurement of key concepts.....	31
3.3. Missing data and attrition.....	34
3.4. Other sources of bias.....	35
4. About the analysis	36
4.1. Overview of analytical approach	36
4.2. Approach to addressing research question(s).....	37
Research question [123]: approach and methods.....	37
5. Project management	40
5.1. Risks and mitigations.....	44
5.2. Timeline	48
6. References.....	49

1. About the project

1.1. Background to the project

In the last decade, England has witnessed a sizeable increase in youth violence, with knife and sharp instrument homicides hitting a record high since 1946. This has stimulated an extensive, mostly non-evidence based, discussion in the media and the political arena. Nevertheless, a comprehensive approach that studies the experience of young people who have contact with the criminal justice system and when, how and why this could lead to a revolving door of repeat youth violence and incarceration is still missing. Developing a better understanding of the effect of the youth justice system on future criminality in England has the potential to help young people who have contact with the criminal justice system avoid recidivism and stay away from prison.

This project brings together academics from the University of Warwick and London School of Economics to study two interconnected research strands related to the youth justice system and its effects on the criminal trajectories of young individuals. The first of these concerns the use of diversion and the impact on recidivism, while the second explores the role of structural change in the criminal justice system. Our contribution to the policy debate in the UK and to the Economics literature will be threefold.

First, we will produce the first piece of rigorous evidence on the effect of youth diversion on the risk of recidivism. While strong evidence exists on the lasting and detrimental impact of arrests and incarceration for the educational and criminal careers of young people (Hjalmarsson, 2008; Mendel, 2011; Aizer and Doyle, 2015; Stevenson, 2017; Mueller-Smith and Schnepel, 2020), very little empirical evidence exists on the potential impact of youth diversion on recidivism, costs of the criminal justice system and the life outcomes of young people. Developing an understanding around this is important for research as well as for policy because diversion has become the standard approach in England for juveniles at low risk of reoffending (Taylor Review, 2016).

Second, studying whether recent courts' and police station closures affected the level of diversion in the youth justice system, the length of commute to local court, and the time elapsed since the offence to completion in youth criminal cases is important because the number of First Time Entrants (FTEs) aged 10-17 to the youth justice system has fallen markedly over the 2010s, especially for FTEs aged 10-17 receiving a caution. As mentioned in the Taylor Review (2016), diversion from the youth justice system of juveniles least likely to reoffend became the norm, with police and youth services seeking to handle the offence informally, while juveniles at greater risk of reoffending continue to enter the youth justice system. During the period of austerity of the early 2010s, the Ministry of Justice (MoJ) and police forces in England experienced budget cuts of more than 25% and 20% respectively.

As a result, by the late 2010s 162 magistrates' courts, where youth courts sit, and 8 crown courts (where more serious offences are tried) were closed. The number of closures of police stations during the same period is even larger with approximately two thirds of the stations closed in England (600 out of 900 police stations). MoJ statistics on waiting times for youth court appointments showed average increases by 40% between 2011 and 2019, increasing up to 491 days in some areas and displaying considerable variation across geography in terms of the impacts. In the context of police station closures, taking London as an example, the average distance to the nearest police station doubled from 1.4 to 3.1 km between 2008 and 2018, affecting response times, ability to solve investigations and clearance rates (Facchetti, 2023), while further evidence suggests proximity to response stations improves clearance rates (Vidal and Kirchmaier, 2018). The combination of these factors could have affected decision-making by the police in terms of arrests and caution patterns, by the CPS in opting away from formal proceedings, and by juveniles by altering their criminal behaviour due to the perceived lack of repercussions.

Third, we plan to examine whether the introduction and changes of sentencing guidelines since 2003 reduced disparity/inequality in sentencing decisions across youth courts and between youths from different socio-economic backgrounds in England. The Criminal Justice Act 2003 gave statutory duty to the Sentencing Guidelines Council (and Sentencing Council which replaced it in 2010). Judges and magistrates must follow the guidelines issued by the Sentencing Council unless under exceptional circumstances, which require written justification. Crime specific guidelines have changed considerably over time, with regular consultations resulting in revisions to guidelines, and these generally happen on a crime-type basis. Between 2010 and 2020 alone 27 new guidelines were published. Geographic variation in sentencing in the UK, controlling for case and individual characteristics, has been historically documented by the Ministry of Justice (e.g., see Mason et al., 2007; Montebruno et al, 2021). Thus, different area crime-type combinations will in turn be differentially exposed to changes in sentencing guidelines depending on the changes in the guidelines, and the average severity of punishments for a particular crime in a particular area. For example, if a guideline becomes narrower, those areas giving more lenient sentences will experience a tighter "floor" in sentencing, while those areas giving more severe sentences will experience a tighter "ceiling" in sentencing. Such variation can be used in conjunction with a staggered continuous triple-difference estimator to explore how sentencing guidelines generate less variance in justice outcomes and give causal estimates on both first and second stage outcomes. The analysis will exploit variation between areas and crime-types, and their interaction, as well as variation across different demographic groups. For example, we can test whether the narrowing of sentencing guidelines changes the disparity in sentencing outcomes for different ethnic groups, which has previously been documented (Hopkins, 2015), or those from different socioeconomic backgrounds. This is of

high importance given the evidence of systematic variation in justice outcomes across regions (MoJ, 2007) as well as ethnicity (Lammy Review, 2017) in Britain. For example, recent Ministry of Justice (MoJ) statistics show between 2010 and 2018 the Black, Asian and Minority Ethnic (BAME) share of youth convictions doubled and as of 2020 more than half the youth custodial population were from a BAME background, despite making up only 18% of the population. Magistrates are typically volunteers who serve in a court in their local community, and no qualification is required to become a magistrate. While they receive training and guidance from a legal advisor it is not a professionalised position. Moreover, recent evidence indicates that even professional judges make systematic mistakes in their pre-trial decisions and their decisions are systematically and unconsciously affected by seemingly unrelated events (Angelova, Dobbie and Yang, 2024). This reflects the incredibly important, yet incredibly difficult task that judges and magistrates are required to perform. Therefore, our analysis will document whether the introduction of the sentencing guidelines may have been significantly helpful for magistrates serving in youth courts to make more balanced decisions within crime-types or whether further modifications to these guidelines might be necessary. The analysis will also be able to examine if crown court judges are subject to similar impacts, despite them typically being long serving legal professionals.

1.2. Research question(s)

Arrested and incarcerated juveniles are less likely to graduate from high school (Hjalmarsson, 2008) and more likely to become recidivists either in youth or in adulthood (Mendel, 2011; Aizer and Doyle, 2015; Stevenson, 2017; Mueller-Smith and Schnepel, 2020). In contrast, youth diversion (youth cautions, out-of-court disposals) can constitute a preferable approach to handle low-level criminality as it might result in reduced recidivism, reduced costs and better outcomes for young people. Although diversion has become the standard approach in England for juveniles at low risk of reoffending (Taylor Review, 2016), little rigorous evidence exists on its impact on serious reoffending and existing spatial inequalities in youth justice outcomes.

This project aims to study how variations in the use of diversion resulting from structural changes, such as court closures, police force closures and sentencing guidelines' introductions, have shaped the criminal trajectories of young people. This research will focus on the severity of recidivism and the heterogeneity in outcomes across English regions, ethnicities and socio-economic backgrounds.

We will investigate the following interrelated research questions:

1. What is the relationship between the use of diversion and the likelihood of recidivism among young people?

2. How have structural changes in police forces and the justice system contributed to the use of diversion? What were the consequences for recidivism in the most affected areas?
3. Have the aforementioned structural changes narrowed or widened existing inequalities? How did the increased use of diversion affect spatial and demographic disparities in criminal and justice outcomes?

Research question 1 relies on variation being induced by the structural change policies (court and police station closures, and changes to sentencing guidelines), while research question 2 is concerned with all policies, and research question 3 is primarily concerned with sentencing guidelines.

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

Question Number ¹	Interim report	Final report
1	<i>Full initial descriptive analysis completed. Preliminary causal analysis.</i> <i>Econometric Modelling: OLS, Difference-in-Difference (with and without Matching), 2 Stage Difference-in-Difference</i>	<i>Final causal analysis & robustness checks.</i> <i>Econometric Modelling: OLS, Difference-in-Difference (with and without Matching), 2 Stage Difference-in-Difference, Event study, Triple-difference, 2 Stage triple-difference</i>
2	<i>Full initial descriptive analysis completed. Preliminary causal analysis.</i> <i>Econometric Modelling: OLS, Difference-in-Difference (with and without Matching)</i>	<i>Final causal analysis & robustness checks.</i> <i>Econometric Modelling: OLS, Difference-in-Difference (with and without Matching), Event study</i>
3	<i>Full initial descriptive analysis completed. Full causal analysis for section a.</i>	<i>Final causal analysis for both parts & robustness checks.</i> <i>Econometric Modelling:</i>

	<i>Econometric Modelling:</i> <i>OLS, Difference-in-Difference (with and without Matching)</i>	<i>OLS, Difference-in-Difference (with and without Matching), Triple difference</i>
--	---	---

1.3. Hypotheses

1. We hypothesise there to be a negative relationship between the use of diversion and the likelihood of recidivism among young people who have contact with the criminal justice system. In other words, we hypothesise that experiencing diversion and thus avoiding a criminal record during youth will be beneficial for the criminal and educational trajectories of the pupils involved. This is because strong evidence exists on the lasting and detrimental impact of arrests and incarceration for the educational and criminal careers of juveniles (Hjalmarsson, 2008; Mendel, 2011; Aizer and Doyle, 2015; Stevenson, 2017; Mueller-Smith and Schnepel, 2020).
2. We hypothesise that recent structural changes in policing and in the justice system, specifically pertaining to court closures and police station closures contributed to the use of diversion and increased waiting time in the criminal justice system due to the reduced capacity of police forces and in courts in England. The effect of these changes on recidivism is not obvious a priori. This is because, on the one hand, as explained above we expect the increased use of diversion to reduce recidivism. On the other hand, these structural changes may have increased the risk of recidivism in the most affected areas due to, among other factors, the increased waiting times in the criminal justice system procedures which can disrupt the reinsertion of young people who have contact with the criminal justice system in society and mainstream schooling. Therefore, which effect will prevail is ultimately an empirical question that will be empirically tested with the DfE-MoJ data.
3. We hypothesise that the increased use of diversion, which is largely discretionary at the level of the deciding actors (police, crown prosecution services and youth offending teams) and by definition it circumvents contact with the criminal justice system, may have increased spatial and demographic disparities in criminal and justice outcomes. This is because we hypothesise that this discretionary measure may have not been used uniformly across different regions and/or demographic groups due to constraints in provision of diversionary routes. Non-uniformity across regions may also occur due to the uneven distribution of provision to support out of court disposals. This is of course just a hypothesis and one may also hypothesise that, if guidance is advising its use and police forces and local partnerships are advocating its use, then one can assume a quasi-mandatory status that should reduce disparities.

This is ultimately an empirical question that will be brought to the data. We also hypothesise that police station and court closures may have widened these existing inequalities as pupils from a low socio-economic background were likely disproportionately exposed to these changes. Finally, we hypothesise that the introduction of sentencing guidelines may have narrowed these existing inequalities because they constituted clear guidelines in the criminal justice system aimed to reduce discretion and increase uniformity across sentencers.

1.4. Key concepts

Table 1.4 Definitions of key concepts

Terms	Definition used
Crime	A crime is a deliberate act that causes physical or psychological harm, damage to or loss of property, and is against the law. In our analysis, we will use the official records of criminal offences from the MoJ's Police National Computer (PNC) database, which includes charges and subsequent convictions and/or cautions.
Youth violence	In our analysis, we will use the definition of youth violence provided by the Royal College of Paediatrics and Child Health (RCPCH), which defines "youth violence" as violence either against or committed by a child or adolescent that can have an impact on individuals, families, communities, and society (RCPCH, 2020). We will also use the YEF definition of violent crime as a " <i>criminal act involving harm against another person that is often more traumatic for the victim (e.g. assault, robbery, homicide).</i> " Within the broader category of "youth violence", we will focus more on violent crimes/offences including rape and sexual assault, robbery, assault and murder as defined in the UK Home Office Crime Classification codes. We will measure these using the official records from the MoJ's PNC database of charges for violent criminal offences with or without injuries for summary and indictable offences, which are more serious offences that must be tried in the Crown Court. ²
Violent Crime/Offence	Violent crime/offence in this report follows the definition used by the DfE and the MoJ and broadly consists of the following categories of offence groups and offence types: indictable-only 'violence against

² An indictable offence usually has more serious punishments (CPS, 2019).

	the person' offences, indictable-only 'robbery offences', and triable either way or indictable-only 'possession of weapons offences'. (DfE, 2023a).
Diversion	Diversion is the legal process in which a person who has contact with the criminal justice system is channelled away from formal judicial proceedings and instead placed into an alternative program or intervention, typically prior to or in lieu of prosecution or sentencing. This process involves suspending or terminating criminal charges on the condition that the person who has contact with the criminal justice system complies with specified requirements or completes a designated program.
Recidivism	Recidivism in the youth criminal context refers to the tendency of a young person who has contact with the criminal justice system to reoffend after having been previously processed through the juvenile justice system. It is typically measured by the rate at which young people who have contact with the criminal justice system are rearrested, reconvicted, or reincarcerated within a specified period following their release or the completion of a diversionary program or sentence. For comparability with official statistics we will use the MOJ definition of proven reoffending although the analysis will not be limited by this definition. MOJ Definition: Proven Reoffending refers to instances where an individual commits a new offense within a specified follow-up period, typically 12 months, after receiving a caution, non-custodial conviction, release from custody, or other formal sanction. For this reoffending to be classified as "proven," it must result in a subsequent conviction, caution, or other formal outcome within an additional period of time, often allowing several months for the new offense to be processed through the criminal justice system.
Local Justice Area	A Local Justice Area (LJA) in the context of the UK justice system is a geographically defined region within which magistrates' courts operate and are responsible for administering justice. Each LJA is established by statutory instruments and determines the jurisdictional boundaries within which magistrates' courts can hear cases, appoint magistrates, and allocate court resources. The concept of LJAs is crucial for organizing the administration of justice at a local level, ensuring that cases are handled by courts that are geographically relevant to the offenses and individuals involved.

Police Force Area	A Police Force Area in the UK is a geographic region defined for the operational jurisdiction of a specific territorial police force. It outlines the boundaries within which the force carries out its law enforcement duties, including crime prevention, investigation, and community engagement. Each area is designed to ensure that policing resources are effectively managed and targeted according to local needs and issues.
Court Closure	Court closure in the UK is the formal administrative action through which a court is officially ceased from conducting judicial proceedings. The process involves a decision by the relevant judicial or administrative authority, such as the Ministry of Justice or a court administrative body, in accordance with statutory provisions and procedural rules. The closure is implemented through a formal order or directive, and the necessary legal procedures are followed to ensure the proper cessation of the court's functions.
Police Station Closure	Police station closure in the UK is the formal administrative action by which a police station is officially ceased from operating. This process involves a decision by the relevant police authority or administrative body, such as the local police force or the Home Office, in accordance with statutory regulations and procedural requirements. The closure is enacted through an official order or directive, and the necessary legal and administrative procedures are followed to ensure the proper termination of the station's operational functions.
Sentencing Guideline	Sentencing guideline in the UK is a formal set of criteria and recommendations issued by the Sentencing Council or other authorized body that provides judges and magistrates with a structured framework for determining appropriate sentences. The guideline includes parameters for assessing the seriousness of offenses, identifying relevant aggravating and mitigating factors, and specifying recommended sentencing ranges or starting points. It is established through legal and procedural processes, ensuring consistency and transparency in sentencing practices across the judicial system.

2. About the datasets

2.1. Overview of datasets used

This project will use the linked dataset from the UK's Department for Education (DfE) and Ministry of Justice (MoJ). This rich dataset enables linking of criminal records of juveniles,

justice system outcomes and information on their educational outcomes for every youth in the UK between 2001 and 2021 inclusive. It contains information on demographics, home address, school exclusions, educational attainment, criminal offences, including type, location, arresting police jurisdiction and co-defendants, and courts proceedings for juveniles, including court location, plea and outcome, from the DfE's National Pupil Database linked at the individual level with the MoJ's Police National Computer and Courts databases for England. It therefore offers enormous potential to follow young people who have contact with the criminal justice system from the date of the offence to the court, and their entire schooling trajectory, thus advancing our understanding of which people who have contact with the criminal justice system enter the justice system and the relationship between the justice system and youth crime, and its interaction with educational outcomes.

In order to implement valid quantitative methods that rely on quasi-experimental variation, the project will merge the individual level datasets described previously with police station and courts closures, local justice area boundaries and offense specific sentencing guidelines.

2.2. Secondary data source(s)

Table 2.2a Dataset Description – *School Census Pupil Level*

Name of dataset	<i>School Census Pupil Level</i>
Data owner(s)	Department for Education
Type of data	Cross-sectional education census
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	Pupil census for all state-maintained schools in England
Years covered or survey waves	2001-2021
Exclusion criteria	Pupils whose education is not funded by the state will not be captured.
Expected population/sample size (following exclusion criteria)	This has information on pupils attending maintained schools from 2001/2 on. In each school year, the universe of pupils in state-maintained secondary schools in England includes approximately 600,000 pupils.

	Therefore, our analysis will include approximately 12 million pupil-year observations.
Documentation	https://www.find-npd-data.education.gov.uk/datasets/775def61-ecd2-4e9a-8ef9-c168c4f51aac

Table 2.2b Dataset Description – *Exclusions Default Data*

Name of dataset	<i>Exclusions Default Data</i>
Data owner(s)	Department for Education
Type of data	Cross-sectional education census
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	All pupil exclusions as collected in the termly School Census (Reason for Exclusion is also included from 2005-06)
Years covered or survey waves	2001-2021
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	This has information on pupil exclusions as collected in the termly School Census. In each school year, the universe of pupils in state-maintained secondary schools in England includes approximately 600,000 pupils, of which approximately 0.5-1% experience permanent exclusion in a school year on average. Therefore, our analysis will include approximately 12 million pupil-year observations and roughly 6,000-7,000 permanent exclusions per year on average.
Documentation	https://www.find-npd-data.education.gov.uk/datasets/78f71e9f-856b-43ee-b0b8-749dd7dd2bb5

Table 2.2c Dataset Description – *Absences Default Data*

Name of dataset	<i>Absences Default Data</i>
-----------------	-------------------------------------

Data owner(s)	Department for Education
Type of data	Cross-sectional education census
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	Absence data for all pupils in state-maintained schools, PRUs and AP academies in England
Years covered or survey waves	2006-2021
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	This has information on pupil absences derived from the termly School Census. In each school year, the universe of pupils in state-maintained secondary schools in England includes approximately 600,000 pupils, of which approximately 20% record multiple unjustified absences from school.
Documentation	https://www.find-npd-data.education.gov.uk/datasets/9cafe398-67af-4dc6-90f3-a9dec511ba92

Table 2.2d Dataset Description – *KS2, KS4 and KS5 Pupil and Exam Tables*

Name of dataset	<i>KS2, KS4 and KS5 Pupil and Exam Tables</i>
Data owner(s)	Department for Education
Type of data	Cross-sectional education census
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	All learners in England who have completed Year 6, Year 11 and post-compulsory education respectively
Years covered or survey waves	2001-2021

Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	Key stage 2, Key Stage 4 and Key Stage 5 attainment data. This has information on the assessment of learners by the end of Key stage 2, Key Stage 4 and Key Stage 5 of schooling.
Documentation	https://www.find-npd-data.education.gov.uk/datasets/d6453111-b401-4420-a68f-7dad865d120f https://www.find-npd-data.education.gov.uk/datasets/d7c2aef7-d051-4b07-86c0-a619bcf94b96 https://www.find-npd-data.education.gov.uk/datasets/82643964-d488-43b2-a50a-0cd4ee3fa2bc

Table 2.2e Dataset Description – *Pupil Referral Unit Census*

Name of dataset	<i>Pupil Referral Unit Census</i>
Data owner(s)	Department for Education
Type of data	Cross-sectional education census
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	Pupil census for all PRUs in England
Years covered or survey waves	2009-2013 (incorporated into the School Census from 2013/14)
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	This has information on all children attending local authority (LA) maintained PRUs. While the sample size of pupils in Pupil Referral Units varies year by year, the count of pupils for the two most recent years for which data are available (i.e., 2019/20 and 2020/21) is respectively 9,602 and 7,665.

Documentation	https://www.find-npd-data.education.gov.uk/datasets/36479c85-5dff-42ec-bdf6-492773eccbae
---------------	---

Table 2.2f Dataset Description – *Alternative Provision Census*

Name of dataset	<i>Alternative Provision Census</i>
Data owner(s)	Department for Education
Type of data	Cross-sectional education census
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	Pupil census for students in AP not maintained by the LA in England
Years covered or survey waves	2007-2021
Exclusion criteria	Pupils whose education is not funded by the local authority will also not be captured, for example, if parents choose to home tutor their child themselves: if this provision is not funded by the local authority, this will not be captured in the AP Census.
Expected population/sample size (following exclusion criteria)	The AP Census includes pupils who attend a school not maintained by a local authority, for whom the authority is paying full tuition fees, or pupils educated other than in schools, pupil referral units, AP academies and AP free schools (from 2013-14) under arrangements made and funded by the authority. While the sample size of pupils in AP varies year by year, the count of pupils for the two most recent years for which data are available (i.e., 2019/20 and 2020/21) is respectively 15,396 and 12,785.
Documentation	https://www.find-npd-data.education.gov.uk/datasets/2f10ee6d-506e-4182-957b-ca88f1a3907c

Table 2.2g Dataset Description – *Police National Computer*

Name of dataset	<i>Police National Computer</i>
Data owner(s)	Ministry of Justice
Type of data	It is used to record convictions, cautions, reprimands and warnings for any offence punishable by imprisonment and any other offence that is specified within the regulations.
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	All linked individuals from Dfe-MoJ dataset
Years covered or survey waves	2001-2021
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	All linked individuals from DfE-MoJ dataset. The Police National Computer contains 13 million person records, and anyone born on the 30 August 1985 or later that ever attended the state-maintained school system in England will appear in our requested data extract.
Documentation	https://www.data.gov.uk/dataset/ab2ef0ee-e741-43c7-b939-d88c19eb69b0/moj-extract-of-police-national-computer

Table 2.2h Dataset Description – *Home Office Court Appearance Statistics (HOCAS)*

Name of dataset	<i>Home Office Court Appearance Statistics (HOCAS)</i>
Data owner(s)	Ministry of Justice
Type of data	The dataset includes Magistrates court data with defendant outcomes including open proceedings
Availability of data	Licence required by the data owners

Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	All individuals from MoJ datasets, including those linked between DfE-MoJ
Years covered or survey waves	2009-2022
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	All individuals from MoJ datasets, including those linked between DfE-MoJ, amounting to approximately 1.8 million individuals.
Documentation	https://datacatalogue.adruk.org/browser/dataset/1131203978465996762/7

Table 2.2i Dataset Description – *Crown Court Defendant Dataset (XHIBIT)*

Name of dataset	<i>Crown Court Defendant Dataset (XHIBIT)</i>
Data owner(s)	Ministry of Justice
Type of data	The Ministry of Justice Data First Crown Court defendant dataset provides data on defendants' appearances in criminal cases before Crown Court in England & Wales, and has been extracted from XHIBIT management information system, used by His Majesty's Courts and Tribunals Service (HMCTS) to manage cases within the Crown Court.
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	All individuals from MoJ datasets, including those linked between DfE-MoJ

Years covered or survey waves	2018-2022
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	All linked individuals from DfE-MoJ dataset
Documentation	https://datacatalogue.adruk.org/browser/dataset/1045635/1

Table 2.2j Dataset Description – *Case management system for crown court cases (CREST)*

Name of dataset	<i>Case management system for crown court cases (CREST)</i>
Data owner(s)	Ministry of Justice
Type of data	Details shared in the extract of this dataset include date and type of offence, the number of people who have contact with the criminal justice system within the case, date of the hearing, and the recorded outcome.
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	All individuals from MoJ datasets, including those linked between DfE-MoJ
Years covered or survey waves	2008-2017
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	All individuals from MoJ datasets, including those linked between DfE-MoJ
Documentation	https://www.gov.uk/guidance/ministry-of-justice-data-first

Table 2.2k Dataset Description – *Offender Assessment System (OASys)*

Name of dataset	<i>Offender Assessment System (OASys)</i>
Data owner(s)	Ministry of Justice
Type of data	The data has been extracted from the Offender Assessment System (OASys), used by His Majesty's Prison & Probation Service (HMPPS) in England to measure the risks and needs of people who have contact with the criminal justice system in custody or under supervision in the community.
Availability of data	Licence required by the data owners
Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	All individuals from MoJ datasets, including those linked between DfE-MoJ
Years covered or survey waves	2011-2022
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	All individuals from MoJ datasets, including those linked between DfE-MoJ
Documentation	https://datacatalogue.adruk.org/browser/dataset/1408722/1

Table 2.2I Dataset Description – *Prison Population, Discharges and Receptions*

Name of dataset	<i>Prison Population, Discharges and Receptions</i>
Data owner(s)	Ministry of Justice
Type of data	Administrative data on people held in custody in prisons and institutions for young people who have contact with the criminal justice system in England, their characteristics, sentence and release.
Availability of data	Licence required by the data owners

Team member(s) who will have access	Nikhil Datta, Rui Costa and Research Assistant
Population/geographic coverage or sampling frame	All individuals from MoJ datasets, including those linked between DfE-MoJ
Years covered or survey waves	2005-2022
Exclusion criteria	N/A
Expected population/sample size (following exclusion criteria)	All individuals from MoJ datasets, including those linked between DfE-MoJ
Documentation	https://datacatalogue.adruk.org/browser/dataset/1045637/1

2.3. Primary data collection

No primary data will be collected

2.4. Linking datasets

The publicly available data on court closures and changes in Local Justice Area boundaries since 2001 have been collected and will be merge with the datasets above described using the exact courts names and locations (COURT_CODE lookup).

The publicly available data on police station closures and addresses since 2008 obtained by FOI for the Metropolitan Police Force Area will be merged with the datasets at the level of police force area identifiers (POLICE_FORCE lookup) and further refined by LSOA of residence of the pupil (LSOAXX_[term][yy]) using the minimum distance between the centroid of the LSOA and the neighbouring police stations. Further data for other major police force areas is being requested via FOI.

The team will collect extensive data on sentencing guidelines for different types of offenses introduced since 2010 following the creation of The Sentencing Council for England and Wales (<https://www.sentencingcouncil.org.uk/offences/>). This data will include:

1. Offense Categories

Seriousness Levels: The guidelines categorize offenses into different levels based on their seriousness. For example, an assault might be classified as "minor," "moderate," or "severe" depending on the harm caused and the culpability of the offender.

Harm and Culpability Factors: These factors help in assessing the seriousness of the offense. Harm refers to the impact on the victim, while culpability refers to the offender's level of responsibility or blameworthiness.

2. Starting Points and Ranges

Starting Point: For each category of offense seriousness, the guidelines provide a starting point for sentencing. This is the sentence that would typically be given for a first-time offender who has been found guilty after a trial.

Sentencing Range: Alongside the starting point, the guidelines provide a range within which the sentence can fall. This range allows for adjustments based on aggravating or mitigating factors.

3. Aggravating and Mitigating Factors

Aggravating Factors: These are circumstances that can increase the severity of the sentence. Examples include previous convictions, use of a weapon, or committing the offense while on bail.

Mitigating Factors: These are circumstances that can reduce the severity of the sentence. Examples include the offender's age, mental health issues, or showing genuine remorse.

4. Guilty Plea Consideration

The guidelines provide for a reduction in sentence if the offender pleads guilty, with the amount of reduction depending on when the plea is entered. The earlier the guilty plea, the greater the reduction, encouraging people who have contact with the criminal justice system to plead guilty at the earliest opportunity.

5. Specific Offense Guidelines

For many offenses, there are detailed guidelines that outline how to assess factors specific to that crime. For example, in cases of burglary, the guidelines might distinguish between domestic and commercial burglary, with different considerations for each.

6. Sentencing Types

The guidelines outline the types of sentences that may be appropriate, such as:

Custodial Sentences: Imprisonment, with options for varying lengths depending on the offense.

Community Orders: Non-custodial sentences that may involve unpaid work, curfews, or rehabilitation requirements.

Fines: Monetary penalties, with the amount usually linked to the seriousness of the offense and the offender's financial situation.

Discharges: Absolute or conditional discharges where no further action is taken or where conditions must be met to avoid further sentencing.

The data on sentencing guidelines will then be merged with the offender level data by date of sentencing (CourtCautionDate) and offense type identifier (CCCJSCode).

The publicly available school-level data on school characteristics and school-level dynamics that we collected will be merged with the DfE-MoJ dataset at the school-level using a school-specific anonymous identifier (URN).

2.5. Access and data protection

The DfE-MoJ dataset will be accessed uniquely via the ONS SRS. Therefore, our use of the data will be subject to the ONS' current regulations in place. We will not need to use any high identifiability data variables (i.e. levels 1 and 2) in our analysis. However, we do need information on the anonymous individual identifier, e.g., the Pupil Matching Reference (PMR) number of pupils in the National Pupil Database (NPD), to be able to merge the different NPD and Ministry of Justice (MoJ) datasets together, e.g., PLASC data with KS4 data and criminal records, at the individual level.

We are aware of the foremost importance of preserving the confidentiality of the data in the analysis and we have extensive experience in working with highly confidential data in the UK and other countries for research purposes. The data will be stored on a secure server and will be accessed by ONS-accredited researchers within the LSE premises, and no attempt will be made to identify young individuals in the DfE-MoJ dataset. At CEP, we fully comply with the LSE Research Laboratory Security Standards for Sensitive Data that are publicly available on the LSE website at the following link:

[LSE Research Laboratory Data Security Policy](#)

LSE also publishes a privacy notice for research subjects that is available at the following link:

[Privacy-Notice-for-Research-v1.2.pdf \(lse.ac.uk\)](#)

Other LSE-wide information on security policies, if required, can be found at the link below:

[Policies and procedures \(lse.ac.uk\)](#)

Should further checks of disclosure and conduct for the procedure be necessary, we would be glad to enclose them.

3. About the data

3.1. List of variables

Table 3.1: Variable definitions

Variable abbreviation	Variable definition	Variable source	Derivation or specification
PupilMatchingRefAnonymous	Character: Unique identifier for a pupil	DfE-MoJ: School Census Pupil Level	Directly provided in the datasets
AgeAtStartOfAcademicYear	Numeric: Age of pupil at start of the academic year (in full years).	DfE-MoJ: School Census Pupil Level	Directly provided in the datasets
EthnicGroup	Categorical: Pupil's ethnic group based on ethnic code.	DfE-MoJ: School Census Pupil Level	Directly provided in the datasets
FSMEligible	Binary	DfE-MoJ: School Census Pupil Level	Directly provided in the datasets
FirstLanguage	Categorical: The language to which the child was exposed during early development and continues to use this language in the home or in the community. If a child acquires	DfE-MoJ: School Census Pupil Level	Directly provided in the datasets

	English after early development, then English is not their first language no matter how proficient in it they become. ENG = English ENB = Not known but believed to be English OTH = Other than English OTB = Not known but believed to be other than English REF = Refused NOT = Information not obtained		
EnrolStatus	C = Current (single registration at this school) G = Guest (pupil not registered at this school but attending some lessons or sessions) M = Current Main (dual registration) S = Current Subsidiary (dual registration) F = FE College (since 2014/15) O = Other	DfE-MoJ: School Census Pupil Level	Directly provided in the datasets

	provider (since 2014/15)		
LSOA01	National Statistics Postcode Directory Lower Layer Super Output Area derived from the pupil's postcode (based on 2001 Census)	DfE-MoJ: School Census Pupil Level	
URN	School unique reference number.	DfE-MoJ: School Census Pupil Level	
HomeLA	LA number based on pupil postcode	DfE-MoJ: School Census Pupil Level	
StartDate	For each exclusion, exclusion start date	DfE-MoJ: Exclusions Data	
PermanentExclusionInd	Binary: Permanent Exclusion Indicator.	DfE-MoJ: Exclusions Data	
Reason	Categorical: For each exclusion, reason for exclusion.	DfE-MoJ: Exclusions Data	
MoJUID	MoJ non-identifiable unique ID	DfE-MoJ: PNC	
CaseID	Identifies individual cases related to each offender. One case may relate to multiple offences.	DfE-MoJ: PNC	

OffenceID	Identifies individual offences for an offender in a case	DfE-MoJ: PNC	
Sex	Gender of the subject.	DfE-MoJ: PNC	
EthnicityCode	An indication of the ethnic appearance of the subject.	DfE-MoJ: PNC	
OffenceStartAge	Age of the offender at the time of the offence.	DfE-MoJ: PNC	
CourtCode	The code identifying the court at which the subject's case was disposed.	DfE-MoJ: PNC	
CourtName	The name and type of a court Note: Used for a non-standard court. This data item may only be used when a non-standard court has to be indicated. If entered the associated court code must be 9998 (for Other).	DfE-MoJ: PNC	Directly provided in the datasets
CourtCautionDate	The date on which the offender was convicted of, or cautioned for, the offence(s).	DfE-MoJ: PNC	Directly provided in the datasets
Cautiontype	This marker indicates the type of Caution received; whether it was adult or juvenile and whether it was conditional or standard.	DfE-MoJ: PNC	Directly provided in the datasets
PNCDisposalCode	The 'type' of the sentence(s) imposed by a court in respect of an offence with which the subject has been charged.	DfE-MoJ: PNC	Directly provided in the datasets

	Identifies the penalty given		
HODisposalCode	The type of sentence imposed by the court, using the Home Office coding scheme.	DfE-MoJ: PNC	Directly provided in the datasets
ACPOCode	A code (ACPO standard) that is unique to the specific type of offence recorded.	DfE-MoJ: PNC	Directly provided in the datasets
CCCJSCode	The CJS offence coding that uniquely describes the offence.	DfE-MoJ: PNC	Directly provided in the datasets
HOffenceCode	An integer used to group the type of offence committed - the full list of over 3,000 offence codes	DfE-MoJ: PNC	Directly provided in the datasets
Offence_group	High level offence group	DfE-MoJ: PNC	Directly provided in the datasets
OffenceStartDate	The first (earliest recorded) date on which the offence was committed.	DfE-MoJ: PNC	Directly provided in the datasets
ProcessForceCode	First two characters of ProcessStationCode, indicating the police force prosecuting the case.	DfE-MoJ: PNC	Directly provided in the datasets
DisposalAmount	This data item indicates the "size" of the sentence imposed by a court (or other authorised agency) in respect of an offence with which the subject has been charged. This variable contains the reported monetary	DfE-MoJ: PNC	Directly provided in the datasets

	values of financial penalties.		
DisposalDuration	This data item indicates the "size" of the sentence imposed by a court (or other authorised agency) in respect of an offence with which the subject has been charged. This variable contains the reported durations of time related penalties.	DfE-MoJ: PNC	Directly provided in the datasets
DisposalDays	DisposalDuration, expressed in days.	DfE-MoJ: PNC	Directly provided in the datasets
IsPrimaryOffence	Indicator of whether this is the main (primary) offence the offender is being tried for	DfE-MoJ: PNC	Directly provided in the datasets
DisposalRank	Ranking of the disposal, in terms of severity, compared to other disposals for that offence.	DfE-MoJ: PNC	Directly provided in the datasets
AdjudicationCode	The recorded result of the court hearing, e.g. guilty/not guilty, for the offence.	DfE-MoJ: PNC	Directly provided in the datasets
PLEA_CODE	1 = guilty / 2 = not guilty / 3 = no plea taken / 6 = guilty by post / 7 = admitted / 8 = denied	DfE-MoJ: HOCAS	Directly provided in the datasets
FINDING	A = not guilty but guilty of another offence / G = Guilty / N = Not Guilty / O = Other	DfE-MoJ: HOCAS	Directly provided in the datasets
date_prel_hearing	Date of the first preliminary hearing at the Crown Court	DfE-MoJ: CREST/XHIB T	Directly provided in the datasets

date_main_hearing	Date of the first main hearing at the Crown Court	DfE-MoJ: CREST/XHIB T	Directly provided in the datasets
COMM_DATE	Date the case is committed to the Crown Court	DfE-MoJ: CREST/XHIB T	Directly provided in the datasets
DATECONV	Date convicted	DfE-MoJ: PRISON DIS	Directly provided in the datasets
DATEREC1	Date of first Reception	DfE-MoJ: PRISON DIS	Directly provided in the datasets
DATEDIS	Date Discharged	DfE-MoJ: PRISON DIS	Directly provided in the datasets
DISCODE	Discharge Code	DfE-MoJ: PRISON DIS	Directly provided in the datasets
EFFLEN	Effective Length of Sentence	DfE-MoJ: PRISON DIS	Directly provided in the datasets
DATEREC1	Date of First Reception	DfE-MoJ: PRISON REC	Directly provided in the datasets
DATESENT	Date Sentenced	DfE-MoJ: PRISON REC	Directly provided in the datasets
SessionsPossible_Spring_ab[yy] AuthorisedAbsence_Spring_ab[yy] AuthorisedAbsenceFlag_Spring_ab[yy] UnauthorisedAbsence_Spring_ab[yy] UnauthorisedAbsenceFlag_Spring_ab[yy]	Number of Spring sessions possible, missed due to authorised absence, missed due to unauthorised absence and flags for persistent absenteeism.	DfE-MoJ: NPD, Absence dataset	Directly provided in the datasets

KS4_PASS_AC	Total number of GCSE/GNVQ qualifications at grades A*-C (GCSE equivalencies).	DfE-MoJ: NPD, KS4 Exam Tables	Directly provided in the datasets
PRU_URN_SPR	PRU's Unique Reference Number	DfE-MoJ: NPD, PRU dataset	Directly provided in the datasets
AP_URN and Aptype	AP's Unique Reference Number and Type of AP (e.g., hospital, out of school, etc).	DfE-MoJ: NPD, AP dataset	Directly provided in the datasets
HIGHLIKERECON	High Likelihood of Reconviction	OASYS	Directly provided in the datasets

3.2. Measurement of key concepts

Table 3.2 Measurement of key concepts

Concept ³	How the concept will be measured and encoded
Diversion	It will be defined a categorical variable that takes value 1 for any case of criminal misconduct which does not include formal prosecution and a court sentence – common examples of diversion include: simple and conditional caution, counselling, educational programs... This variable is possible to construct from the Disposal Code of the offense which gives in detail the penalty received by the person who has contact with the criminal justice system in each case including diversion channels.

³ This should align directly with the names and list of concepts defined in table 1.3

Recidivism	It will be defined as categorical variable that takes value 1 in the case the person who has contact with the criminal justice system appearing in the police and court record is not a first-time offender. This is possible to construct with the data as not only we have indicator of first-time contact with the criminal justice system in the courts data, we have a panel data structure of people who have contact with the criminal justice system which enables us to construct the history of offenses for each individual in our dataset. Additionally, we will define and intensive margin of recidivism which counts the number of subsequent offenses for each individual in our dataset.
Inequality of Outcomes	We will estimate inequality of outcomes measuring the covariate conditional differences at different moments of the distribution (mean, median, variance) of the selected outcomes (diversion, recidivism, severity of sentencing) across the relevant groups (socio-economic status, age, ethnicity, gender).
Sentencing Severity	Sentencing severity will be defined through multiple metrics to capture the intensity of the punishment assigned to people who have contact with the criminal justice system . First, we will use the length of incarceration as a measure, recording the duration of prison sentences, with longer sentences indicating higher severity. We will also differentiate by the type of sentence imposed, with custodial sentences (e.g., prison) considered more severe than non-custodial sentences (e.g., probation). Additionally, sentencing severity will incorporate monetary penalties, measured by the fine or restitution amount relative to the offender's financial status. Expected sentence

	lengths, based on sentencing guidelines, will serve as a benchmark to compare actual sentences and assess the severity of deviations. Lastly, real time-served will be used to capture the actual duration a person who has contact with the criminal justice system spends in custody, accounting for adjustments like early release or parole.
DATECONV	We will use information on the date of the conviction as an outcome to measure how much time passed between the time of the offence and the time of the conviction.
DATEREC1	We will use information on the date of the reception in prison as an outcome to measure how much time passed between the time of the offence and the time of imprisonment.
DATEDIS	We will use information on the date of the reception in prison and discharge from prison to measure how much time an individual spent in prison.
DISCODE	We will use information on the discharge code to conduct heterogeneity analysis between individuals discharged under different circumstances.
EFFLEN	We will use information on the Effective Length of Sentence to measure how much time an individual spent under the sentence.
DATESENT	We will use information on the date of the sentence as an outcome to measure how much time passed between the time of the offence and the time of the sentence.
SessionsPossible_Spring_ab[yy] AuthorisedAbsence_Spring_ab[yy] AuthorisedAbsenceFlag_Spring_ab[yy] UnauthorisedAbsence_Spring_ab[yy] UnauthorisedAbsenceFlag_Spring_ab[yy]	We will use information on justified and unjustified absences as a fraction of all possible sessions to measure truancy and the disruption to “normal” school attendance by a pupil who had contact with the criminal justice system.

KS4_PASS_AC	We will use information on GCSE test scores to measure the effect of diversion on student performance in high-stakes exams.
PRU_URN_SPR	We will use information on the PRU where a pupil is enrolled to measure the likelihood that diversion results in a differential likelihood of enrolment in a PRU.
AP_URN and APtype	We will use information on the AP where a pupil is enrolled to measure the likelihood that diversion results in a differential likelihood of enrolment in a AP institution.
HIGHLIKERECON	We will use information on the risk profile of people who had contact with the criminal justice system to define a binary variable (0/1) that distinguishes high-risk individuals for recidivism from others.

3.3. Missing data and attrition

We anticipate two missing data problems when using the DfE-MoJ linked dataset.

First, the main threat to the quality of our analysis stems from pupils with frequently changing addresses not always being tracked by the NPD. To identify a pupil, the NPD makes use of instant pupil identifiers such as the pupil's name and postcode. However, if a pupil frequently changes addresses over a short span, then the NPD may not accurately track this pupil across different years, until eventually the pupil might disappear from the dataset altogether. Since frequent changes of address are more likely among youth from low-income and broken households, it is therefore important to acknowledge that the pupils who have been able to be matched in both the NPD and the PNC datasets are likely to originate from households with relatively stable socio-economic conditions. This may imply an upward bias in the correlation between diversion and the school trajectory of pupils (i.e., the true correlation might be more negative than what we observe in our data extract).

Second, individuals who were not matched across the DfE and MoJ datasets have very specific characteristics with respect to gender, ethnicity and age. In particular, the ADR UK (2022) finds that 75% of the unmatched cases were male and 75% of the unmatched cases were of White Northern European ethnicity, followed by the general category of "Unknown"

ethnicity⁴ at 11% of all unmatched cases. Individuals of Black, Asian, Middle Eastern, Japanese, Chinese or Southeast Asian ethnicity sum up to a total of 9% of all unmatched cases in the dataset. Finally, unmatched cases were more likely to come from the older (initial) cohorts due to the greater probability of the address listed in the justice data matching the address listed in the education data for the younger cohorts. In this sense, we anticipate the population of white, male and older individuals to be under-represented in the MoJ-DfE dataset. The direction of the bias is ambiguous a priori in this case.

We are aware that the issue of pupils disappearing from the DfE dataset is likely to be biased towards children that may have contact with the criminal justice system. The extent to which this is the case will be tested comparing the rate at which pupils disappear from the dataset whether they appear in the Police National Computer (PNC) or not. This comparison will be made using regression analysis and controlling for other potential determinants of this attrition in the data (e.g., foreign native language). However, it is important to reiterate that we requested access to the list of variables enumerated above from the Police National Computer 2001-2021 and other MoJ datasets for criminal records of individuals at all ages (i.e., for a linked individual while s/he is observed in the DfE data but also after s/he disappears from the DfE data. Therefore, we will be able to observe criminal offences occurred after a linked individual has either reached the compulsory schooling age or s/he has disappeared from the DfE records ahead of time).

Apart from these three shortcomings, we do not anticipate any additional gaps in our data. This is because we requested access to the above NPD extract for all pupils in state-maintained schools, pupil referral units and alternative provision in all school years linked at the individual level with the Police National Computer data and other MoJ datasets from 2001 to 2021. From the MoJ, we requested access to the list of variables enumerated above for records of individuals at all ages. The DfE-MoJ also provides a Match Quality dataset that provides details on how each person who has contact with the criminal justice system was matched to the NPD: this information would allow us to choose the observations for the analysis better, as well as highlight any potential biases in the matching processing. ADR UK (2022) finds that 70% of individuals with a MoJ identifier can be identified to an individual in the DfE data sources.

3.4. Other sources of bias

⁴ It is important to keep in mind that in the Police National Computer (PNC) data, ethnicity of an individual is not self-reported but rather identified by the officer in question. This could potentially explain why the “Unknown” ethnicity category is the second leading category among unmatched cases.

Although our analysis uses administrative data from DfE and MoJ, the data may be biased as some ethnic groups may be over-represented and some others may be under-represented. For instance, regarding criminal activity data, statistics from the Ministry of Justice (MoJ) from 2019 acknowledge that people from BAME ethnic groups (Black, Asian, Mixed, Chinese, and “other”) are over-represented in the UK at every single stage of the UK criminal justice system, be it arrest, prosecution, conviction, or imprisonment (Yasin & Sturge, 2020), and even within this group there is important variation across different ethnicities.

The authors also explain that in the UK criminal system, pleading “guilty” at the sentencing stage often leads to a sentence length discount of one third. However, the authors also highlight that pleading guilty as early as in the sentencing stage is correlated with a greater degree of trust in the criminal justice system, which is something higher among White than among BAME defendants. As a result, while White defendants have a higher rate of “guilty” pleading, the average sentence length for BAME defendants in 2019 was 27.1 months compared to 19.5 months for White defendants (Yasin & Sturge, 2020). Given this sharp discrepancy in trust with respect to the UK criminal justice system, we therefore expect BAME individuals to be over-represented both in terms of offending and reoffending statistics in the datasets. In light of the overrepresentation of some ethnic groups in the British criminal justice system, our analysis will, therefore, take care in interpreting the results of the correlation between offending and ethnicity, so as to avoid stigmatising the overrepresented racial groups.

4. About the analysis

4.1. Overview of analytical approach

As soon as we receive the permission from the data owners, we will start conducting tests of reliability of the linked DfE-MoJ dataset (henceforth, the data), and core dimensions of data quality (completeness, uniqueness, timeliness, validity, accuracy, and consistency) will be assessed. Once we have completed the data quality checks, we will start exploring empirically the relationship between the use of diversion and the likelihood of recidivism among young people who have contact with the criminal justice system. We will do so both through descriptive statistics and regression analysis. To be precise, we will use a combination of graphs, e.g., trees, and tables to visually describe the possible crime trajectories of pupils who experienced diversion. The path from diversion to each terminal node of the tree (e.g., return to school, recidivism, etc) will represent each potential trajectory a pupil may have after diversion. Each node will also contain information on the proportion of pupils who are on that specific trajectory and on the relevant descriptive statistics (e.g., crime rates) for each subsample of pupils.

Subsequently, we will analyse how structural changes in police forces and the justice system may relate to and/or affect the use of diversion. We will also explore the consequences for recidivism in the most affected areas both through descriptive statistics and regression analysis.

Finally, we will investigate whether the aforementioned structural changes narrowed or widened existing inequalities and whether the increased use of diversion affected spatial and demographic disparities in criminal and justice outcomes using regression analysis.

We will primarily use the OLS model (also referred to as the linear probability model when using a binary outcome variable.. The OLS model is a useful econometric tool as it enables us to easily interpret the estimated coefficients. For example, if in an OLS regression for diversion the coefficient for FSM eligibility is 0.02, it means that, for two pupils who are identical in all other factors included in the regression (also known as control variables), the probability that a pupil who is FSM eligible is diverted is 0.02 units (in the dependent variable) higher than for the pupil who is not FSM eligible. Therefore, the OLS model can be helpful in studying the direction of the correlation that different factors may have with our outcomes of interest, and their relative importance.

We will also use propensity score matching to estimate the impact of diversion on youth offending. Using regressions for diversion and offending, we can identify covariates that are associated with both diversion and offending. Holding all other factors constant, by comparing the offending outcomes of otherwise similar pupils (based on other covariates) exposed to different criminal proceedings, we can obtain a better estimate of the effect that experiencing different types of criminal proceedings may have on the probability of offending.

4.2. Approach to addressing research question(s)⁵

Research question [1]: approach and methods

Research question	<i>What is the relationship between the use of diversion and the likelihood of recidivism among young people who have contact with the criminal justice system?</i>
--------------------------	---

⁵ The main methodology remains the staggered difference-in-difference approach from Sun and Abraham (2021). However, the following sections present the methodologies that will be used in the interim report as a partial response to the research questions.

<i>Hypothesis, if relevant</i>	There will be a negative relationship between the use of diversion and the likelihood of recidivism among young people who have contact with the criminal justice system . In other words, we hypothesise that experiencing diversion and thus avoiding the “criminal label” during youth will be beneficial for the criminal and educational trajectories of the pupils involved. This is because strong evidence exists on the lasting and detrimental impact of arrests and incarceration for the educational and criminal careers of juveniles (Hjalmarsson, 2008; Mendel, 2011; Aizer and Doyle, 2015; Stevenson, 2017; Mueller-Smith and Schnepel, 2020).
<i>What will you be able to say by the interim report</i>	<i>By the interim report, we will be able to provide descriptive and correlational results concerning the research question.</i> <i>The final report will also include regression analysis.</i>
<i>Descriptive analysis, if relevant</i>	<i>Using the DfE-MoJ dataset, we will define within each cohort the group of diverted pupils. For each cohort of diverted pupils in our analysis (i.e., the cohorts enumerated in sections 1.2 and 4.1 above), we will check in the data whether they appear in the same or the next academic year in the MoJ data for a subsequent offence. We will exploit information in the MoJ data on the dates of the offence and enrolment in mainstream schooling. This will enable us to examine their journey from diversion to either returning to mainstream school or recidivating.</i> <i>We will express these trajectories using unconditional comparisons of the fractions of pupils who go through one journey or another, e.g., from diversion to recidivism. This will be grounded in what is observed in the data and driven by a thorough knowledge of how to group journeys in a meaningful way. In other words, we will describe the data here and the fractions of pupils who embark on different journeys from their first contact with the justice system. The</i>

	<i>statistics will be a concrete output of our work and they will become visible once access to the actual data is gained.</i>
<i>Models, specifications and statistical techniques used, if relevant</i>	<i>The analysis for the interim report will be descriptive, while the analysis for the final report will include OLS, Difference-in-Difference, 2 Stage Difference-in-Difference, Event study, Triple-difference, 2 Stage triple-difference,</i>
<i>Estimating equation, if relevant</i>	<i>We will regress recidivism and other complementary criminal and justice outcomes on the exposure to structural change, changes in sentencing guidelines, and diversion, with a variety of fixed effects, depending on the specification. This include time, location, and crime-type. The inclusion of time varying controls will be decided carefully due to the endogenous nature of them, and risk of being a “bad control”, however these include local expenditure on policing services, if publicly available.</i>
<i>What does the approach need to succeed (constraints/assumptions)?</i>	<i>We require data on how each criminal offence is handled, available in the MoJ data. We require that these students’ path post-diversion be tracked in the DfE-MoJ dataset to be able to study these questions descriptively as proposed here.</i> <i>In this sense, we require most crime offences that diverted pupils might commit after diversion to be properly recorded by MoJ.</i>
<i>Uncertainty and inference</i>	<i>P-values, t statistics, confidence intervals, F-statistics (when using 2 stage estimates)</i>
<i>Robustness checks</i>	<i>PSM, Synthetic Controls, Event Study</i>
<i>Subgroup you intend to study</i>	<i>Ethnic minorities and pupils diverted from school at different ages.</i>

<i>Changes to the analysis</i>	The analysis will take into account potential not random missing data on outcomes and covariates. Econometric sample selection bounds will be estimated according to level of non-reporting if justified
---------------------------------------	--

Research question [2]: approach and methods

<i>Research question</i>	<i>How have structural changes in police forces and the justice system contributed to the use of diversion? What were the consequences for recidivism in the most affected areas?</i>
<i>Hypothesis, if relevant</i>	Recent structural changes in policing and in the justice system contributed to the use of diversion and increased both waiting time in the criminal justice system due to the reduced capacity of police forces and courts in England.
<i>What will you be able to say by the interim report</i>	<i>By the interim report, we will be able to provide descriptive and correlational results concerning the research question.</i> <i>The final report will also include regression analysis.</i>
<i>Descriptive analysis, if relevant</i>	<i>Using the DfE-MoJ dataset, we will define within each cohort the group of diverted pupils. For each cohort of diverted pupils in our analysis (i.e., the cohorts enumerated in sections 1.2 and 4.1 above), we will check in the data whether police station and court closures correlate with the likelihood to be diverted from the criminal justice system.</i> <i>Similarly to our previous question, we will express these trajectories using unconditional comparisons of the fractions of pupils who go through one journey or another, e.g., pupils in areas more affected by the recent structural changes vs others. For pupils in different regions, we will describe the fractions of pupils taking each potential route. As specified above, these descriptive statistics will describe the data and will be driven by what is observed in the data</i>

	<i>once the DfE-MoJ data become available to us. These descriptive statistics will constitute a valuable output of this research.</i>
<i>Models, specifications and statistical techniques used, if relevant</i>	<i>The analysis for the interim report will be descriptive, while the analysis for the final report will include OLS, Difference-in-Difference, Event study</i>
<i>Estimating equation, if relevant</i>	<i>We will regress recidivism and other complementary criminal and justice outcomes on the exposure to structural change, changes in sentencing guidelines, and diversion, with a variety of fixed effects, depending on the specification. This include time, location, and crime-type. The inclusion of time varying controls will be decided carefully due to the endogenous nature of them, and risk of being a “bad control”, however these include local expenditure on policing services, if publicly available.</i>
<i>What does the approach need to succeed (constraints/assumptions)?</i>	<i>We require data on how each criminal offence is handled, available in the MoJ data. We require that these students’ path pre- and post-diversion be tracked in the DfE-MoJ dataset to be able to study these questions descriptively as proposed here.</i> <i>In this sense, we require most crime offences that diverted pupils might commit after diversion to be properly recorded by MoJ.</i>
<i>Uncertainty and inference</i>	<i>P-values, t statistics, confidence intervals</i>
<i>Robustness checks</i>	<i>PSM, Synthetic Controls, Event Study</i>
<i>Subgroup you intend to study</i>	<i>Ethnic minorities and pupils at different ages. Focus will be on pupils of secondary school age.</i>
<i>Changes to the analysis</i>	<i>The analysis will take into account potential not random missing data on outcomes and covariates. Econometric</i>

	sample selection bounds will be estimated according to level of non-reporting if justified
--	--

Research question [3]: approach and methods

Research question	<i>How did the increased use of diversion affect spatial and demographic disparities in criminal and justice outcomes? Have the aforementioned structural changes narrowed or widened these existing inequalities?</i>
Hypothesis, if relevant	Increased use of diversion increased spatial and demographic disparities in criminal and justice outcomes. Sentencing guidelines reduced spatial and demographic disparities in criminal and justice outcomes.
What will you be able to say by the interim report	<i>By the interim report, we will be able to provide descriptive and correlational results concerning the research question.</i> <i>The final report will also include regression analysis.</i>
Descriptive analysis, if relevant	<i>Using the DfE-MoJ dataset, we will define within each cohort the group of diverted pupils. For each cohort of diverted pupils in our analysis (i.e., the cohorts enumerated in sections 1.2 and 4.1 above), we will focus on youth who have committed a similar offence and check in the data whether the use of diversion correlates with later outcomes in the criminal justice system.</i> <i>We will express these trajectories using unconditional comparisons of the fractions of pupils who go through one journey or another, e.g., pupils who committed a given offence and experienced diversion vs others who committed the same offence and did not experience diversion.</i>

<i>Models, specifications and statistical techniques used, if relevant</i>	<i>The analysis for the interim report will be descriptive, while the analysis for the final report will include OLS, Difference-in-Difference, Triple difference.</i>
<i>Estimating equation, if relevant</i>	<i>We will regress criminal and justice outcomes of the bite of sentencing guideline changes by area x crime type, with location x time, time x crime type, location x crime type fixed effects, in the case of the triple difference approach. Sentencing guideline bite is defined by the proportion of outcomes in the pre-period that would have been adjusted had the sentencing happened in the post-period, similar to the Minimum Wage bite approach (see Datta, Giupponi and Machin, 2019 for an example).</i> <i>This approach can also be changed to exploit only crime type and time variation, or location and time variation in a difference-in-difference approach.</i> <i>Heterogeneity analysis can be carried out by interacting the main right hand side variable with different demographics (e.g. FSM).</i>
<i>What does the approach need to succeed (constraints/assumptions)?</i>	<i>We require data on how each criminal offence is handled, available in the MoJ data. We require that these students' path pre- and post-diversion be tracked in the DfE-MoJ dataset to be able to study these questions descriptively as proposed here.</i> <i>In this sense, we require most crime offences that pupils might commit to be correctly recorded by MoJ.</i> <i>Causality rests on a parallel trends assumption which is testable using an event study.</i>
<i>Uncertainty and inference</i>	<i>P-values, t statistics, confidence intervals</i>

Robustness checks	- PSM, Synthetic Controls, Event Study
Subgroup you intend to study	<i>Ethnic minorities and pupils at different ages.</i>
Changes to the analysis	The analysis will take into account potential not random missing data on outcomes and covariates. Econometric sample selection bounds will be estimated according to level of non-reporting if justified

5. Project management

5.1. Risks and mitigations

Table 5.1 Risks and mitigations

Number	Risk	Likelihood (Low/Medium/ High)	Mitigation
1	<i>Data Reliability</i>	<i>e.g. Low</i>	We have extensive experience of assessing data reliability for DfE as well as numerous police forces in the UK. As a recent example, since 2016 we have had access to National Pupil Database (NPD) data linked with HMRC data on individual tax records and DWP data on individual records of benefits receipts. We are also currently examining the database of the West Midlands Police (WMP) and providing analytical support to WMP's operational agenda. We produced more than 200 pages of descriptive results and presented this in meetings with WMP's data

			analysts and senior officials. Our analysis revealed empirical trends that were not known to WMP before. This analysis also exposed anomalies in the data and led to changes in the production of statistics and data extraction practices by WMP. This reflect our experience of dealing with missing data and it indicates that we would be able to detect whether some groups of population are overrepresented in a pool of observations that may be missing.
2	<i>Identifying individuals from the data</i>	Low	We do not need to use any high identifiability data variables (i.e. levels 1 and 2) in our analysis. In contrast, we need information on the anonymous individual identifier, e.g., the Pupil Matching Reference (PMR) number of pupils in the National Pupil Database (NPD), to be able to merge the different NPD and Ministry of Justice (MoJ) datasets together, e.g., PLASC data with KS4 data and criminal records, at the individual level. Our analysis of the DfE-MoJ data linkage will strictly comply with the regulations in place by the data owners as well as by the ONS. The DfE-MoJ dataset contains de-identified data for each individual, making it

			impossible to identify any particular person within the dataset. Furthermore, as part of our data access agreement, we are subject to strict data disclosure protocols, and any observations below a threshold of 10 will be suppressed and removed from any document that is prepared for publication.
3	<i>Data Confidentiality</i>	<i>Low</i>	We are aware of the foremost importance of preserving the confidentiality of the data in the analysis and we have extensive experience in working with highly confidential data in the UK and other countries for research purposes. No identifiable information will be revealed to anyone of course, and no attempt will be made to identify young individuals in the DfE-MoJ dataset. At CEP, we fully comply with the LSE Research Laboratory Security Standards for Sensitive Data that are publicly available on the LSE website at the following link: <u>LSE Research Laboratory Data Security Policy</u>

			LSE also publishes a privacy notice for research subjects that is available at the following link: <u>Privacy-Notice-for-Research-v1.2.pdf (lse.ac.uk)</u> Other LSE-wide information on security policies, if required, can be found at the link below: <u>Policies and procedures (lse.ac.uk)</u> Should further checks of disclosure and conduct for the procedure be necessary, we would be glad to enclose them.
4	<i>Data Complexity</i>	<i>Low</i>	We have detailed knowledge of the NPD data and we are extremely familiar with its structure. In particular, as we have examined the legislation and the dynamics of behavioural outcomes in England (i.e., school absence as well as lunchtime, temporary and permanent exclusions), we would be able to detect anomalies in the data and thus test its reliability very easily. We do not foresee any difficulties in sharing a metadata dictionary and giving relevant, practical advice in analysing the available data.
5	<i>Data Access</i>	<i>Low</i>	The application for data access has already been submitted.

			The application has received preliminary approval from DfE and it is now waiting for the feedback of the Judicial Office of MOJ. Although we expect some clarifications will be requested by MOJ as it was the case for DfE, we have no reason, based on previous experience and our correspondence with DfE and MoJ until now, to believe the data access will not be approved before the end of 2024.
--	--	--	---

5.2. Timeline

Table 5.2 Timeline

Date	Activity	Staff responsible/leading
Project start	Submit application to ONS for access to DfE-MoJ datasets.	February 2024 – Datta, Costa, Sandi
	Start of hiring process of one or more part-time Research Assistants (RAs) who will be supervised by Datta, Costa and Sandi.	November 2024 – Datta, Costa, Sandi
	Start of descriptive interim report on the evolution of alternative provision in England in the last 20 years, i.e., from the early 2000s.	December 2025 – Datta, Costa, Sandi and RA (to be hired)
Agree study plan	Discussion between YEF and CEP on study plan	September/October 2024 – Datta, Costa, Sandi
Data Access	Submit application to ONS for access to DfE-MoJ datasets.	February 2024 – Datta, Costa, Sandi
	Complete data access obtained.	December 2024 – Datta, Costa, Sandi

Interim report	Descriptive interim report completed.	October 2025 – Datta, Costa, Sandi and RA (to be hired)
Final report	Dissemination of preliminary findings and presentation of the early results of this analysis and collection of feedback from YEF colleagues.	September 2025 – Datta, Costa, Sandi and RA (to be hired)
	Respond to comments from YEF and YEF appointed external peer review	November/December 2025 – Datta, Costa, Sandi and RA (to be hired)
	Submit final report	March 2026 – Datta, Costa, Sandi and RA (to be hired)

6. References

ADR UK (Administrative Data Research UK). (2022, July). Ministry of Justice – Department for Education linked dataset *Feasibility of evaluating early interventions for violence prevention: Data quality report*. Retrieved from: https://www.adruk.org/fileadmin/uploads/adruk/Documents/Feasibility_study_1_MoJ-DfE_linked_dataset_Data_quality_report.pdf

Aizer, A., and J. J. Doyle (2015). Juvenile Incarceration, Human Capital and Future Crime: Evidence from Randomly-Assigned Judges. *Quarterly Journal of Economics*, 130: 759–803.

Angelova, V., Dobbie, W., and Yang, C.S. (2024). Algorithmic Recommendations When the Stakes Are High: Evidence from Judicial Elections. *AEA Papers and Proceedings*, 114: 633–37.

The Crown Prosecution Service (CPS). (2019). *Indictable only cases: sending to the Crown Court*. Retrieved from: <https://www.cps.gov.uk/legal-guidance/indictable-only-cases-sending-crown-court>

Datta, N., Giupponi, G., & Machin, S. (2019). Zero-hours contracts and labour market policy. *Economic Policy*, 34 (99): 369–427. <https://doi.org/10.1093/epolic/eiz008>

Department for Education (2023a, March 30). *Education, children’s social care and offending: Descriptive statistics (technical note)*. Retrieved from <https://explore-education-statistics.service.gov.uk/methodology/education-children-s-social-care-and-offending-descriptive-statistics-technical-note#content-section-0-content-8>

- Fachetti, E. (2024). Police Infrastructure, Police Performance, and Crime: Evidence from Austerity Cuts. IFS Working Paper No. 24/76. Retrieved from: <https://ifs.org.uk/sites/default/files/2024-04/WP202416-Police-infrastructure-police-performance-and-crime-evidence-from-austerity-cuts.pdf>
- Hjalmarsson, R. (2008). Criminal justice involvement and high school completion. *Journal of Urban Economics*, 63(2): 613–630. <https://doi.org/10.1016/j.jue.2007.04.003>
- Hopkins, K. (2015). *Associations between police-recorded ethnic background and being sentenced to prison in England and Wales*. Ministry of Justice (MoJ). Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/479874/analysis-of-ethnicityand-custodial-sentences.pdf
- Lammy, D. (2017). *The Lammy Review: An independent review into the treatment of, and outcomes for, Black, Asian and Minority Ethnic individuals in the Criminal Justice System*. London: Lammy Review.
- Mason, T., de Silva, N., Sharma, N., Brown, D., & Harper, G. (2007). *Local variation in sentencing in England and Wales*. Ministry of Justice, London.
- Mendel, R. A. (2011). *No Place for Kids: The Case for Reducing Juvenile Incarceration*. Annie E. Casey Foundation Technical Report.
- Ministry of Justice (2007). Local Variation in Sentencing in England and Wales. Retrieved from: Local Variation in Sentencing in England and Wales (publishing.service.gov.uk)
- Montebruno, P., Silva, O. and Szumilo, N. (2021). Judge Dread: court severity, repossession risk and demand in mortgage and housing markets. *CEP Discussion Paper No. 1766*, ISSN 2042-2695.
- Mueller-Smith, M., & Schnepel, K. T. (2021). Diversion in the Criminal Justice System. *The Review of Economic Studies*, 88(2), 883–936. <https://doi.org/10.1093/restud/rdaa030>
- Royal College of Paediatrics and Child Health (RCPCH). (2020). *Youth Violence – State of Child Health*. Retrieved from <https://stateofchildhealth.rcpch.ac.uk/evidence/injury-prevention/youth-violence/#:~:text=Youth%20violence%20is%20understood%20as,%2C%20families%2C%20communities%20and%20society.>
- Stevenson, M. (2017). Breaking Bad: Mechanisms of Social Influence and the Path to Criminality in Juvenile Jails. *The Review of Economics and Statistics*, December 2017, 99(5): 824–838.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2), 175-199.

Taylor, C. (2016). *Review of the Youth Justice System in England and Wales*, Ministry of Justice.

Vidal, J. B. I., and Kirchmaier, T. (2018). The Effect of Police Response Time on Crime Clearance Rates. *The Review of Economic Studies*, 85(2 (303)); 855–891.
<https://doi.org/10.1093/restud/rdx044>

Yasin, B. and Sturge, G. (2020, October 02). *Ethnicity and the criminal justice system: What does recent data say on over-representation?* Retrieved from:
<https://commonslibrary.parliament.uk/ethnicity-and-the-criminal-justice-system-what-does-recent-data-say/>