Data Explained

Ministry of Justice – Department for Education linked dataset

School funding, pupil performance and crime: a quasi-experimental study

Author: Dr Will Cook

Date: March 2023

This Data Explained summarises experiences and learning from working with the National Pupil Database (NPD) and Police National Computer (PNC) linked datasets in the course of producing research into the effect of school funding programmes on crime. This publication is intended to help guide future researchers using this data and to provide feedback into future dataset development and documentation.

The administrative data discussed in this Data Explained was made securely available through the Ministry of Justice (MoJ) Data First project funded by ADR UK (Administrative Data Research UK) (Grant number: ES/W002620/1). ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation). The data used in this research project comes from the Department for Education and the Ministry for Justice and was accessed through the Office for National Statistics (ONS) Secure Research Service (SRS). The data was not originally collected for research and it is expected that there are gaps and inconsistencies in its recording, a number of which are detailed in the following.

Project details

This project aimed to assess whether three school funding programmes that have been found to have raised pupil attainment also had the effect of reducing the proportion of pupils committing crime. The three programmes were:

- The Leadership Incentive Grant – A programme from 2003-2006 that provided additional funding to secondary schools with low GCSE pass rates and/or high proportions of pupils eligible for free school meals.
- Pupil Learning Credits – In 2001, secondary schools in mostly urban areas with high proportions of pupils eligible for free school meals were allocated additional per pupil funding. This was to support additional provision for disadvantaged pupils aged 11-14, and was an early forerunner of the pupil premium policy.
- Primary Excellence in Cities – An area-based programme that increased the funding of primary schools in urban areas with high proportions of pupils eligible for free school meals between 2001-2006.

Initial research questions

1. Did school funding programmes that are known to have increased pupil performance also have the effect of reducing crime, both in the short and the long term?

2. What are the potential benefits of this linked dataset for other researchers, policy stakeholders and the public?

Research methodology

The methodology was based around estimating the causal effects of policy on crime.

Both the Leadership Incentive Grant and the Pupil Learning Credits policy were awarded to schools on pre-determined thresholds that related to either their percentage pass rate at GCSE or their percentage of pupils eligible for free school meals (FSM). This feature of the policies facilitates the use of 'regression discontinuity'[1] methods to uncover causal effects. The Primary Excellence in Cities initiative was implemented on specific schools at a particular point in time, allowing for the implementation of 'difference in difference' methods.

[1] A description of this method and that of difference in difference can be found in sections A.2.8. and A.2.7. of the Magenta Book; the HM Treasury guidance on conducting evaluation in public policy https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879418/Magenta_Book_Annex_A_Analytical_methods_for_use_within_an_evaluation.pdf

Key variables

Dataset name	Variable type	Variable name*	Description
NPD	Identifiers	pupilmatchingrefanonymous_xx	Pupil ID to match across NPD datasets
		Laestab_xx	School ID
	Attainment	ks2_englev	Level achieved in KS2 English
		ks2_matlev	Level achieved in KS2 Maths
		ks2_scilev	Level achieved in KS2 Science
		ks2_engtotmrk	Mark for KS2 English
		ks2_mattotmrk	Mark for KS2 Maths
		ks2_scitotmrk	Mark for KS2 Science
		ks4_apmat	Grade in GCSE Maths
		ks4_apeng	Grade in GCSE English Language
	Pupil Characteristics	senstage_xx	Pupil's special educational need category
		freeschoolmeals_xx	Whether a pupil is eligible for free school meals
		gender_xx	Gender of pupil
		ethnicgroup_xx	Ethnic group of pupil

		mothertongue_xx	Whether a pupil was mostly exposed to a language other than English at home
		actualncyeargroup_xx	Year group of pupil (e.g. year 7, year 8, etc)
PNC	Identifier	mojuid	Individual level identifier to link to NPD
	Offence details	hooffencecode	Categorisations of different offence types
		hodisposalcode	Categorisations of different CJS disposals (e.g. community order custodial sentence, etc(
		offencestartdate	Date of offence
		maxadjudicationcode	Outcome for each offence (e.g., guilty, not-guilty, caution, etc)
		offencestartage	Age of individual at time of offence.

* Note: _xx is used to denote that in the School's Census data variables are suffixed with the latter two digits of the year in which the census was taken.

Summary of comments on specific variables

Variable name(s)	Comments
Hooffencecode (PNC)	Match in Home Office (HO) offence descriptions and groups to understand this variable. A significant proportion of offences cannot be matched to HO offence groups, this is because they are either breach offences or were committed outside of England and Wales.
Maxadjudicatiocode/adjudicationcode (PNC)	Consider filtering out 'non-offences' i.e. those categorised as 'not guilty' and 'non-conviction' using this variable. Note that some of these non-offences have sentences attached, this is because they relate to conditions imposed by the court in lieu of prosecution.
ks2_mattotmrk, ks2_engtotmark, ks2_scitotmrk (NPD)	Recognise that pupils who are likely to offend are disproportionately missing from test score data and implement missing data strategies accordingly.
laestab, urn (NPD)	Use the Consistent Schools Database (CSD) to ensure that changes in school governance are not mistaken for changes in school attended.

How you dealt with data limitations

Creating individual level outcome variables

This project aimed to identify whether education policies affected crime outcomes, and users of the NPD-PNC data will typically be using the crime variables as outcomes. Therefore, it is important to understand what outcomes are possible using the dataset and the structure of both the NPD and PNC datasets. The NPD is (mainly) a pupil level records dataset, whereas the PNC is at the individual-offence level. As the PNC data contains multiple records per individual in offence level records, researchers may initially want to prepare the PNC data by creating summary variables for each individual and collapsing the data into person level records in order to match into the NPD.

The simplest approach to do this is to use 'contact with the CJS' as a binary indicator – essentially dividing individuals based on whether they have a record on the PNC. Creating this variable simply involves creating a column of 1s in the PNC dataset and matching into the NPD using the *mojuid* individual level identifier.

Such a variable will include those subsequently found not guilty. These cases may be filtered out using the *maxadjudicatoncode* variable. However, note that such cases may still have custodial sentences recorded against them, in most cases this is because one of the outcomes of the alleged offence is a bindover order, where the court imposes conditions on an individual rather than prosecuting.

Identifying offence groups

Beyond identifying whether an individual has had any contact with the criminal justice system, my project also attempted to analyse whether the funding programmes had any effect on specific types of crime. The version of the PNC used for this research only contained descriptions of offence types by offence code without labels. As such, it was necessary to match in offence descriptions and offence groups (e.g. theft, violence against the person) to understand what the offence types are.

The link between the offence codes and the descriptions was provided in the metadata which needed to be ingested into the ONS SRS in order to be used, and then Home Office (HO) offence group variables were created to categorise the offence codes in the data. Since obtaining the data for this research, the dataset has been updated such that offence group variables are now included, though researchers may still wish to match in more detailed offence descriptions.

When these offence descriptions were matched in for my project it was clear that a large number of offences in the dataset were not matchable to an offence group. On further analysis it was found that the majority of these were breach offences or offences that were committed outside of England and Wales. These types of offences are indicated by a letter prior to the offence code. At first I attempted to identify which offence groups these may relate to using any offences linked to them via the *case_id* variable, but in many cases this was not possible. For a number of these uncategorised offences the offence data is implausible, e.g. >70,000 of these offences are recorded with the year of offence as 1900. The vast majority (>98%) of offences with these implausible dates are those recorded in Scotland (code S in the PNC). As a solution to this it is suggested that offences without a corresponding HO offence description or offence group are coded to an 'other' category within the HO offence group variable. However, users should consider that there is likely to be significant variation as to the type of offence in this 'other' category, and the offence dates are missing or unreliable in a large number of cases for this category.

Controlling for baseline offending

A strong correlate with offending at a given point in time is whether an individual has offended in the past. As part of my project, I was planning to use 'prior offending' as a covariate in my modelling, and to break down the analysis into sub-groups of those with a record of offending prior to policy implementation vs non-offenders. This is comparable to the 'prior attainment' variable used in models with education outcomes.

An issue that I identified however was that while data was available on offences recorded prior to 2000, this only applied to those who had subsequently offended post 2000. This means that a 'prior offending' variable that covers cohorts aged 10 and over prior to 2000 is not a valid measure, as

some individuals will have prior offending records that will not appear in the linked dataset. For many applications this will not be an issue, however much of my analysis concerns pupils who made the transition to secondary school around the turn of the millennium and therefore I had to discard the approach of using prior offending as a variable in my modelling.

Generating comparable metrics across cohorts

In my project I used data from multiple cohorts of pupils. For the difference in difference analysis, it was essential to do this in order to apply the research design. Care needs to be taken when using multiple cohorts of pupils to ensure that crime outcomes are comparable, as older cohorts will naturally have higher crime rates simply because they have longer periods of their life covered in the dataset. In order to make crime outcomes comparable I defined crime outcomes according to offender age – for example, offences committed when aged 21, or under 16. As the dataset used in my research recorded crime up to 2017 (note however that the dataset now goes up to the end of 2020) it was necessary to take extra care that outcome definitions were not affected by this constraint (e.g. by making sure variables defined as crime outcomes committed up to the age of 21 are not measured for cohorts who turned 21 after the year 2017). If using the updated dataset, a similar approach can be taken for 2020.

School identifiers

Over the last thirty years there have been significant changes to the schooling system: conversions to academy status, amalgamations, closures and the opening of new schools. When schools go through these changes, the school identifier will usually change too – this affects the *laestab* and *urn* variables in the dataset. For example, a pupil remaining in a school that converts to an academy would appear as if they had moved schools when the school converts, even though they remain in the same institution. There were two main challenges regarding this in my project.

Firstly, any model that recognises the hierarchical structure of the data when modelling outcomes (e.g. multilevel models, panel models) may not correctly group together pupils from different cohorts that attend the same school, rendering the findings from these models unreliable. Secondly any research that is done that analyses the effect of school level policies may not accurately identify which schools were part of a particular programme.

The solution to this problem is to use a common identifier for schools that change status over time. In my research I have used the school identifier that is present in the dataset in the year of policy implementation. The challenge then is to ensure that any cohorts represented in the modelling that also attended the same school also are linked to this common identifier. A useful aid for this is the 'Consistent Schools Database' that is available on request from CLOSER that provides *laestab* school identifiers for each year and shows how these change for individual schools over time.

Pupil attainment – KS2

Although not the main focus of my project, I used pupil attainment variables to replicate the results of previous studies into the policies I considered and also, in the case of KS2 attainment, as a control

variable. KS2 attainment is available as both the awarded national curriculum level and also as the raw test score for assessments in English, Maths and Science. Test scores need to be used with care as some pupils will have missing values, usually either because they were absent on the day of the test and/or they were not entered for the test. These pupils are twice as likely to have records in the PNC compared to pupils who do have a test mark recorded. It is therefore important not to inadvertently exclude pupils with no KS2 test scores from any analysis that relates to crime outcomes and to implement an appropriate strategy to deal with missingness in the KS2 test score data.

Another issue to be considered regarding KS2 test scores is that the scores are not comparable between years as the mapping of the marks to attainment varies from year to year. So, if multiple cohorts of pupils are to be analysed the scores need to be transformed – one way of doing this is to find the mark threshold for each of the subjects where a pupil achieves national curriculum level 4 (the "expected" level) and zero centre the test score marks around this threshold for each cohort.

Suggested improvements recommended to data owners

Provide a system map explaining how the variable values are created

Using both the NPD and PNC without some degree of knowledge of both the education and criminal justice system is challenging. What would be really useful for researchers is a system map of both the education system and the criminal justice system. This should illustrate how individuals come into contact with each system and how they move through it, identifying where variables in the datasets are generated.

Match in HO offence descriptions and offence groups into the supplied datasets

The offence codes are meaningless without their attached groups or descriptions. This has been partly addressed in the most recent update of the data with the addition of offence groups, but It is likely that many users of the PNC data will want to match in the HO offence descriptions as well. To save needless duplication, these variable should be provided as part of the datasets supplied into the ONS SRS.

Supply the metadata with the raw datasets

In the original data supply the metadata was not supplied with the data. However, it is crucial to work with the metadata in order to understand the data. The metadata should be supplied with the project along with the data.

Provide code where data owners or the ONS have produced their own analysis and/or created derived variables

During the course of the fellowship the DfE and ONS published work using the linked dataset. To avoid duplication, inconsistencies and potential error, the code behind any government publications use should be supplied with the dataset.

Additional data which would help to further develop the research

Police Force Area data reliability
During the period covered by the dataset, there have been a number of concerns raised about the accuracy of data recorded by certain police forces. Where this is likely to have a material impact on any statistics or results produced from the PNC data, this should be flagged in the dataset. This would allow researchers to test whether their results are robust to the exclusion of data from police forces that is not thought to be as reliable as other data in the dataset.

Household identifiers
NPD data contains pupil addresses, though these are rarely made available to researchers. However anonymous household identifiers should be made available in the NPD. That way research could better understand the influences of siblings and household changes in crime outcomes.

Parental crime
Related to the above, the MoJ should consider linking in PNC records that involve parents of children in the NPD. There is growing evidence from the USA on the effect of parental contact with the criminal justice system on child outcomes, however there is almost none in the UK context. Understanding the link between parental crime and child outcomes is important for, for example:

- identifying children most at risk of adverse child outcomes
- the extent of any educational disadvantage to growing up in a household with an offender as a parent
- modelling whether elements of the criminal justice system compound or otherwise reduce other disadvantages.

Archived school level data
There is school level data that is available in the National Archives but is not available within the NPD. For example, this includes historical school funding data, data on class sizes, pupil teacher ratios and pupil composition. As the linked dataset covers cohorts as far back as those born in the mid -1980s, matching in this historical data would aid other projects that are concerned with long term policy evaluation.

Disclaimer

This work was produced using administrative data accessed through the ONS SRS. The use of the data in this work does not imply the endorsement of the SRS or data owners in relation to the interpretation or analysis.

This work uses research datasets which may not exactly reproduce National Statistics aggregates. National Statistics follow consistent statistical conventions over time and cannot be compared to Data First linked datasets.

Acknowledgements

This work was supported by ADR UK (Administrative Data Research UK). ADR UK is a partnership transforming the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives. ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation). Grant number: ES/W002620/1.

Contact

Name: Dr Will Cook
Email: w.cook@mmu.ac.uk