ADR ENGLAND

Data Explained

Ministry of Justice – Department for Education linked dataset

Authors: Janine Boshoff and Dr Matteo Sandi
Date: 31 October 2022

The data discussed in this Data Explained was made securely available by the Ministry of Justice (MoJ) and Department for Education (DfE). This Data Explained summarises experiences and learning from working with the MoJ-DfE linked dataset in the course of producing research into two projects that aim to understand the crime-mitigating impact of education. Namely: i) the impact of basic literacy and numeracy skills on education achievement and criminality, and ii) the effect of in-service teacher training (INSET) days and school breaks on the short-term pattern of juvenile crime.

This Data Explained intends to help guide future researchers using this dataset and feed back into dataset development and documentation. The data used in this research project comes from de-identified administrative data. It was accessed through the ONS Secure research Service from the Centre for Economic Performance at the London School of Economics. The data was not originally collected for research and it is expected that there might be gaps and inconsistencies in its recording.

Initial research questions

Basic skills and crime: the impact of basic literacy and numeracy skills on education achievement and criminality

1. Did the Literacy Hour and Numeracy Hour Programmes raise literacy and numeracy skills in childhood?
2. What is the impact of the literacy and numeracy skills accumulated in childhood on youth crime?

Impact of school attendance on juvenile crime: the effect of INSET days and school breaks on the short-term pattern of juvenile crime

1. Does school attendance influence the day-to-day propensity of young people to commit crime?
2. How do results vary for property crime and violent crime?
3. Are variations in short-term juvenile crime patterns driven by marginal or prolific offenders?

Research methodology

Basic skills and crime

The first step entailed the collection of data on the gradual timing of adoption of the Literacy Hour and Numeracy Hour programmes in English primary schools in the late 1990s. This data was cleaned and formatted in preparation for merging with our variables of interest from the MoJ-DfE linked dataset.

Using the autumn census within the National Pupil Database (NPD), we collected information for children taking the key stage 2 (KS2) exams at the end of primary school (i.e., aged 10) during our timeframe of interest ranging from 1995/96 to 1998/99. This is because, after the 1998/99 school year, all primary schools in England adopted the Literacy and Numeracy Hour Programmes, and so after 1998/99, there is no statistical variation left to exploit for estimation of the causal effects of these reforms. We captured each individual's unique pupil matching reference (PMR) number, their MoJ unique identifying number should they appear in the Police National Computer (PNC), and the establishment number for the school that KS2 takers were attending in the academic years 1995/96 to 1998/99.

From the MoJ dataset, we computed the count of all crimes that were committed within the age range 16-25 by each individual who sat KS2 exams from 1996/97 to 1998/99. We kept the MoJ's unique identifying number for each offender within the PNC as well as the classification of the type of crime committed. The classification was used to create broader crime categories that could be used to determine whether and how basic skills affect different types of crimes differently among young people.

The NPD and MoJ datasets were merged at the individual level to determine if a juvenile committed more than one crime within the age range 16-25. In this analysis, the unit of observation was the individual. Two binary variables are created in the combined datasets that

indicate whether a particular individual was exposed to either the Literacy or the Numeracy Hour Programme. A set of variables was also created to group juvenile crimes among violent crime, property crime, drug crime and other crime.

These categories were defined using the Home Office offence code. "Violent crime" is restricted to indictable offences and it includes Violence against the person and sexual offences. "Property crime" is also restricted to indictable offences and it includes robbery, theft, criminal damage and arson and fraud offences. While "drug crime" is restricted to indictable drug offences, "other crime" includes all the remaining categories in the Home Office offence codes, i.e., possession of weapons, public order offences, miscellaneous crimes against society, summary motoring and non-motoring offences and undefined offences.

Analysis was conducted on the simple count of crime for the full sample. These regressions controlled for a year and school-level fixed effect, as well as for a set of individual level characteristics (such as gender, ethnicity, free-school-meal eligibility status and native language). The analysis tested for anticipation effects in the year leading up to the adoption of either the Literacy or the Numeracy Hour Programmes, and found no evidence that young people enrolled in the schools that adopted these initiatives started to outperform or underperform others prior to the adoption of these initiatives.

Another binary variable was created that would take value zero if no permanent exclusion was received in year 11, and value one if one or more permanent exclusions were received in that year. Again, a set of linear regressions and pseudo-poisson maximum likelihood estimates were produced controlling for the set of fixed effects and control variables mentioned above.

Finally, using the information on student achievement at the end of secondary school, a set of outcomes measuring the performance of students in Key Stage 4 exams were also created for each young person.

Impact of school attendance on juvenile crime

The first step entailed a data-scraping exercise that collected the dates of scheduled INSET days as well as term dates and breaks for secondary schools in England from August 2013 to December 2017. These data were cleaned and formatted in preparation for merging with our variables of interest from the MoJ-DfE dataset.

Using the spring census within the NPD, we collected information for children aged 11 to 16 during our timeframe of interest. We captured each individual's unique pupil matching reference (PMR) number, their MoJ unique identifying number should they appear in the PNC, and the establishment number for the school they were attending for the academic years 2013/2014 to 2017/2018.

From the MoJ dataset, we compiled the dates of all crimes that were committed within the academic calendar years of interest. We kept the MoJ's unique identifying number for each offender within the PNC as well as the classification of the type of crime committed. The

classification was used to create broader crime categories that include property crime and violent crime.

The NPD and MoJ datasets were merged at the individual level, to determine if a young person committed more than one crime within the period under review. The data was then rectangularised to create a dataset with daily frequency that counts the number of crimes committed for each school in the academic years 2013/2014 to 2017/2018. The data was then merged to the school dataset with information on term and INSET dates. It is important to clarify that crime counts sum up daily recorded crime incidents by all pupils in the same school. Therefore, the unit of analysis of this sub-study is school not individuals.

Two binary variables (inset_days and school_off) were created in the combined datasets that indicate whether a particular day coincides with an INSET day or a school break. Three variables were created to reflect the most frequently occurring juvenile crimes: violent crime, property crime and total crime, which is a combination of the other two categories.

These regressions were weighted by the average pupil enrollment across the five academic years and controlled for a date and school-level fixed effect. The analysis tested for differential crime trends in the week leading up to an INSET day or a school break, as well as in the week after an INSET day or a school break.

Another binary variable was created that would take value zero if no crimes were committed on a particular day, and value one if one or more crimes were committed on a particular day. Again, we conducted a linear regression and pseudo-poisson maximum likelihood estimation, weighted by the average pupil enrollment across the five academic years and controlling for a date and school-level fixed effect. The analysis tested again for differential crime trends in the week leading up to an INSET day or a school break, as well as in the week after an INSET day or a school break.

Finally, using the crime count and the average pupil enrollment across the five academic years, a crime rate was created for each of the three crime types (violent crime rate, property crime rate and total crime rate). Again, we conducted a linear regression and pseudo-poisson maximum likelihood estimation weighting by the average pupil enrollment across the five academic years and controlling for a date and school-level fixed effect.

Datasets and variables used

- Autumn and spring census from NPD: pupil identification number, school establishment number, age during academic year, free-school-meal eligibility, gender, ethnicity, native language.
- PNC from MoJ: unique identification number in MoJ, offence start date and the Home Office offence code.

Data limitations encountered

PNC data only recorded offences until December 2017. We had hoped to capture all offences until the end of the 2017/2018 academic year (approximately until the end of July 2018), so our analysis was limited to 4½ academic years instead of the 5 academic years we proposed to analyse initially.[1]

We were unable to compute the count of crimes committed by individuals born prior to August 1985 because the DfE share includes individuals born after August 1985.

How you dealt with data limitations

To study the impact of INSET dates on juvenile crime we needed to collect data ourselves on the INSET dates and schools breaks of each secondary school in England.

Given the detailed nature of the data and the construction of a daily frequency dataset, we believe that conducting the analysis on 4½ academic years instead of 5 academic years should not lead to contrasting results.

Key Stage 4 records are not directly comparable across years and so they need to be standardised within school years.

Suggested improvements recommended to data owners

It would be really helpful to:

- include data on classes and teachers in the NPD
- link the MoJ data with data on earnings, such as the Longitudinal Education Outcomes dataset
- make the MoJ-DfE linked dataset accessible remotely from abroad (i.e., through the use of fingerprints and machines that allow secure access).

Additional data which would help to further develop the research

Linking this data with family linkages and histories of domestic violence would allow us to explore how the effects documented in this study vary according to the family background of an individual.

[1] At the time of writing, the MoJ-DfE data share has been extended up until the end of 2020. However, the text reflects the fact that the empirical analysis was conducted on data requested in 2020 that went up to the end of 2017.

Disclaimer

This work was produced using administrative data accessed through the ONS Secure Research Service. The use of the data in this work does not imply the endorsement of the ONS or the data owners in relation to the interpretation or analysis of the data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

Acknowledgements

The support by the London School of Economics and its staff while conducting this study is gratefully acknowledged. This study is part of the Education Policy and Youth Crime in England project funded by ADR England (Grant number: ES/V01742X/1). We thank the Department for Education and the Ministry of Justice for guidance and enabling access to their linked administrative data extracts.

About ADR England

ADR England is a portfolio of data linking and research projects, delivered by academic and government partners to provide policy-relevant insights using data held by UK Government departments and public bodies. The ADR England portfolio is commissioned and managed by the ADR UK Strategic Hub team embedded within ADR UK's funder, the Economic and Social Research Council. Projects are commissioned in line with the ADR England Strategy, which prioritises research for public good that cuts across traditional policy boundaries.

Contact

Name: Dr Matteo Sandi
Email: m.sandi@lse.ac.uk