

Data Explained

Ministry of Justice – Department for Education linked dataset

Exploring the dynamics of school absenteeism and crime

Author: David Buil-Gil

Date: February 2026

This Data Explained report summarises experiences and lessons learned from working with the Ministry of Justice (MoJ) and Department for Education (DfE) linked dataset in the course of conducting research on the relationship between school absences and crime (recorded cautions and convictions). 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 report was made securely available through the ADR UK-funded project “Exploring the dynamics of school absenteeism and crime.” The data used in this research comes from the MoJ-DfE linked dataset and was accessed through the ONS Secure Research Service via the Department of Criminology at the University of Manchester. As this data was not originally collected for research purposes, gaps and inconsistencies in recording are to be expected; several of these are outlined in the sections that follow.

Project details

School attendance is a key protective factor in young people's development, yet absenteeism, and particularly, persistent absenteeism, remains a major concern in England (Long & Roberts, 2025). This project examined how patterns of authorised and unauthorised school absences are associated with later involvement in crime, measured through cautions and convictions recorded in the Police National Computer (PNC). It also explored how these associations vary across different stages of the life course and across social groups, with particular attention to socio-economic disadvantage, and controlling for variables such as ethnicity and sex. Finally, the project analysed the extent to which the effect of socio-economic disadvantage on future crime operates through school absences.

The analysis used the MoJ-DfE linked dataset, which combines detailed education records from the National Pupil Database (including absences and exclusions) with criminal justice records from the PNC. The dataset enables longitudinal tracking of individuals born between the 1990/91 and 1997/98 academic years, from their school years into early adulthood (up to age 32 for some cohorts).

The project investigated whether school absenteeism predicts short-, medium-, and long-term criminal involvement, and whether these effects differ across socio-economic groups. By linking individual-level educational trajectories with later offending outcomes, the project identified for whom and under what conditions absenteeism serves as an early warning signal for future criminal justice contact.

The findings inform early prevention strategies by clarifying whether, when, and how school attendance relates to later offending, and how interventions might be better targeted to reduce inequalities in educational engagement and crime involvement.

Initial research questions

1. What are the effects of school absenteeism on crime at different stages of life?
2. To what extent do the effects of school absenteeism on later offending vary by socio-economic disadvantage?
3. To what effect is the effect of socio-economic disadvantage on later offending mediated by school absences?

Research methodology

This project used longitudinal administrative data from the MoJ-DfE linked dataset for cohorts born between the 1990/91 and 1997/98 academic years in England. School attendance data (authorised and unauthorised absences by term), exclusions, and demographic characteristics were drawn from the National Pupil Database (NPD). These records were linked to criminal justice outcomes from the Police National Computer (PNC), including cautions and convictions recorded from adolescence into early adulthood (up to age 32). Prison population data were used to account for periods of incapacitation.

The analysis began with descriptive and exploratory methods to examine patterns of absenteeism and offending across age, demographic groups, and neighbourhood contexts. This included visualisation of trends over the life-course and group comparisons using summary statistics.

Multivariate regression models (including quasi-Poisson regression models) were used to estimate the association between absenteeism and later offending, controlling for individual and community-level factors. Offending outcomes were classified into short-, medium-, and long-term periods following recorded absences to capture temporal dynamics in the relationship. An article summarising the results is currently undergoing peer review.

Survival analysis was applied to model the timing of first and repeated offending events following periods of absenteeism. Cox proportional hazards models were used to examine how absenteeism relates to the risk of criminal involvement over time (results available in Buil-Gil & Pease, 2025).

Finally, mediation analysis was used to examine whether school absenteeism explains the relationship between socio-economic disadvantage and later crime involvement. This was implemented using quasi-Bayesian simulations with 500 draws. The mediator model was specified as a Poisson regression predicting school absences, while the outcome models were specified as quasi-Poisson regressions predicting cautioned and convicted offences. Mediation analyses were conducted separately for multiple age thresholds to allow the indirect and direct effects to vary across developmental stages. The approach estimated the Average Causal Mediation Effect (ACME), Average Direct Effect (ADE), and Total Effect (TE), as well as the proportion of the total effect of disadvantage that operates through absenteeism. An article summarising the results is currently undergoing peer review.

All analyses were conducted in R software within the ONS Secure Research Service. Only aggregated, disclosure-checked outputs were released, ensuring full compliance with data protection and confidentiality requirements.

Key variables

Dataset	Variables
School census	Pupil identification number, MoJ unique reference number, school establishment number, local authority code, Lower Layer Super Output Area (LSOA) code, year and month of birth, gender, ethnicity group, Income Deprivation Affecting Children Index (IDACI).
Absences (each academic year)	Pupil identification number, MoJ unique reference number, school establishment number, number of sessions possible, number of sessions missed due to authorised absence, number of sessions missed due to unauthorised absence, year and month of birth, gender, ethnicity group, Income Deprivation Affecting Children Index (IDACI).
Exclusions (each academic year)	Pupil identification number, MoJ unique reference number, reason for exclusion, category, number of sessions missed due to exclusions.
PNC	MoJ unique reference number, case ID, offence ID, offence start date, offence start age, offence group, Home Office offence code.
Prison	MoJ unique reference number, date received, sentence in years, sentence in months, date of release.

Summary of comments on specific variables

Variable	Comments
Age (PNC)	A small proportion of pupils with recorded offending data were listed as being under the age of 10, which is below the minimum age of criminal responsibility.
LSOA codes (absences)	LSOA code information, and therefore IDACI scores, was missing for a number of pupils and years.
Sessions missed (absences)	The proportion of pupils with missing information on missed sessions due to absences was higher at age 7 than at later stages. Data on missed sessions at age 7 were also noisier and showed less clear patterns than at other ages, suggesting potential recording issues.
Ethnicity codes (absences)	The categories used to classify ethnic groups vary across years.
Dates (PNC and prison)	Dates follow different formats across datasets, particularly between the PNC and prison datasets.

How you dealt with data limitations

Age (PNC): A very small number of observations (1%) indicated criminal justice outcomes for children below the age of 10. Because individuals under 10 are not criminally responsible in England and Wales, such records are not legally meaningful as cautions or convictions and were likely the result of data misclassification. These observations were excluded from the analytical samples used in regression, survival, and mediation models.

LSOA codes (absences): Where possible, missing pupil-level LSOA codes were supplemented with LSOA details from the closest available point in time. This approach allowed us to recover some contextual data, although it was not feasible in all cases. As a result, LSOA codes for absences (and the corresponding LSOA-level information, including IDACI scores) remained missing for a number of pupils and years.

Sessions missed (absences): The proportion of pupils with missing information on missed sessions due to absences was higher at age 7 than at later stages. Data on missed sessions at age 7 were also noisier and showed less clear patterns than at other ages, suggesting potential recording issues. The variable was retained in the analyses in order to maximise the use of available data. However, we include a note of caution in the article regarding its interpretation.

Ethnicity codes (absences): To ensure consistency over time, we standardised ethnicity classifications using a common set of codes.

Dates (PNC and prison): To ensure consistency, all dates were standardised to a common format.

Suggested improvements recommended to data owners

It would be particularly valuable to consider the following improvements:

- **Clarify records for children under 10:** A small number of cases show cautions or convictions for children below the age of criminal responsibility. These should be investigated and, where invalid, recoded or flagged in the raw data.
- **Harmonise demographic categories:** Key demographic and social variables are coded differently across datasets. Providing integrated categories, or standardised analytic code to harmonise them, would improve consistency and usability.
- **Standardise date formats:** Dates are recorded in different formats across datasets, complicating analysis. Aligning date formats or providing conversion scripts would reduce preprocessing time and errors.
- **Provide synthetic data:** High-quality synthetic data reflecting the structure and key features of the administrative data would support code development, training, and early-stage analysis outside the secure environment.

Additional data which would help to develop the research

Linking the MoJ-DfE administrative records with individual-level Census data within the ONS Secure Research Service would offer substantial additional opportunities for research. Census data contain rich information on household composition, housing conditions, employment, health, and migration, which would allow researchers to better understand the wider social and family contexts associated with school absenteeism and later involvement in crime. Integrating these data sources would support more comprehensive analyses of how educational experiences interact with living conditions, family structures, and socio-economic circumstances across the life course, helping to strengthen both theoretical development and policy relevance.

References

Buil-Gil, D., & Pease, K. (2025). *Data Insight: From school absences to crime involvement*. Retrieved from: <https://www.adruk.org/news-publications/publications-reports/data-insight-from-school-absences-to-crime-involvement/>

Long, R., & Roberts, N. (2025). *School attendance in England*. House of Common Library Research Briefing. Retrieved from: <https://commonslibrary.parliament.uk/research-briefings/cbp-9710/>

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 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 is 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/Z503204/1]

Contact

Name: David Buil-Gil

Email: david.builgil@manchester.ac.uk

