

Data Explained

Ministry of Justice – Department for Education linked dataset

Socio-emotional characteristics in early childhood and offending behaviour in adolescence

Author: Paul Garcia

Date: December 2025

This Data Explained summarises experiences and learning from working with the Ministry of Justice (MoJ) and Department for Education (DfE) linked dataset in the course of producing research into the relationship between early child development and the likelihood of offending in adolescence. 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 project “Socio-emotional characteristics in early childhood and offending behaviour in adolescence” funded by ADR UK. The data used in this research project comes from the MoJ-DfE linked dataset and was accessed through the ONS Secure Research Service via the Institute for Social and Economic Research at the University of Essex. 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

Early developmental characteristics in childhood are key determinants of later life outcomes. This study examines how early developmental difficulties affect the likelihood of being cautioned or convicted for an offence during adolescence. It also explores the pathways through which poor early development shapes offending behaviour, with a particular focus on school outcomes such as attainment, exclusions, and persistent absenteeism.

The analysis uses the MoJ-DfE linked dataset, which brings together child development assessments from the Early Years Foundation Stage Profile (EYFSP) at age 5, educational outcomes up to age 11 (including attainment, exclusions, and absences), and criminal records from ages 11 to 17 recorded in the Police National Computer (PNC).

Children who later offend show lower scores across all EYFSP scales. A factor analysis reveals two underlying dimensions: **cognitive difficulties**, covering challenges in language, literacy, reasoning, and numeracy; and **socio-emotional difficulties**, capturing poor emotional regulation and problems in forming relationships. Cognitive difficulties appear to be largely mediated by educational outcomes, whereas socio-emotional difficulties remain significant in explaining adolescent offending even after accounting for school-related factors.

The study aims to inform early intervention strategies to reduce youth offending by identifying key early childhood difficulties and their pathways to offending. Different types of difficulties require different solutions: cognitive problems may be more effectively addressed through school-based interventions, while socio-emotional difficulties may require family- or community-based support.

Initial research questions

1. What socio-emotional characteristics are exhibited by children in early childhood who engage in offending during adolescence?
2. How do school difficulties (e.g., absenteeism, exclusions) and poor attainment influence the relationship between socio-emotional development and adolescent offending?
3. How do unfavourable characteristics within schools and local authorities interact with socio-emotional development to shape offending behaviour?

Research methodology

This project uses point-scales indicators of child development from the EYFSP for three cohorts of pupils, aged 4–5, from the 2006/2007, 2007/2008, and 2008/2009 school years. These records are matched to later educational outcomes, such as absences, exclusions, and Key Stage 2 assessments, as well as cautions and convictions recorded between ages 11 and 17 in the Police National Computer (PNC).

I use Exploratory Structural Equation Modelling (ESEM), which combines the flexibility of Exploratory Factor Analysis (EFA) with the structure of Confirmatory Factor Analysis (CFA). First, EFA lets the EYFSP data reveal the underlying dimensions of child development. Then, using that structure, CFA provides a more formal model within the SEM framework. ESEM captures the complexity of the EYFSP data by remaining flexible enough to account for indicators shared across domains, while still providing a rigorous structure for testing relationships.

To examine potential pathways, I use mediation analysis to assess whether socio-emotional and cognitive difficulties influence adolescent offending directly, or indirectly through school outcomes such as attainment, exclusions, and persistent absenteeism.

Finally, I account for broader contextual factors, such as schools and local authorities, to examine how individual child development interacts with these environments in shaping offending behaviour in adolescence.

Key variables

Dataset name	
School census:	Pupil identification number, school establishment number, local authority code, year and month of birth, free-school-meal eligibility, gender, ethnicity, native language, number of siblings, whether first born, part-time in school, deciles of Income Deprivation Affecting Children Index (IDACI) and SEN provisions.
EYFSP:	Point-scale indicators of children's development at the end of reception.
CIN:	Procession year, number of previous CPP, referral date, age start of CIN period.
CLA:	Procession year, start date of continuous care, length of care, placement.
Absence:	Establishment, number of sessions possible, number of sessions missed due to authorised absence, number of sessions missed due to unauthorised absence.
Exclusion:	Reason, category, number of fixed exclusions, number of sessions for fixed exclusions, permanent exclusion count.
Key Stage 2:	Test marks (Exam Table)
PNC:	Unique identification number in MoJ, case ID, offence ID, offence start date, offence start age, offence group, primary offence, disposal rank, adjudication code, Home Office offence code.

Summary of comments on specific variables

Variable name(s)	Comments
EYFSP:	I used the 2006/07, 2007/08, and 2008/09 cohorts of pupils in their school reception year. While total scores are available for the full population, disaggregated point-scale data is only available for about 40% of pupils. The documentation indicates that, at the time, local authorities were not required to submit disaggregated data for all pupils, but only for a sample (DfE, 2010).
Key Stage 2:	Choosing the best representation of KS2 assessments is challenging because the system has changed over time, making results not fully comparable across cohorts. For my study cohort, I used whether pupils achieved the expected level, as there were no analogous continuous teacher-assessed measures available across all relevant subjects: Mathematics, Grammar, Punctuation and Spelling, and Reading.
PNC:	A small percentage less than 1% of pupils in the cohorts of interest with recorded offending data were listed as being younger than 10 years old which is the minimum age of criminal responsibility.

How you dealt with data limitations

EYFSP: Because the disaggregated information is only available for a subsample of pupils, there is a concern that the missing data is not random. To address this, I conducted statistical checks and applied an inverse probability weighting procedure, using information on pupils' socio-economic status, geographical location, and demographics. This helps adjust for the fact that we are working with a sample rather than the full population.

Key Stage 2: I requested additional data from the Exam Tables dataset, which includes raw test scores from the national curriculum assessments. This allowed me to construct continuous measures of attainment in the relevant subjects for the cohorts of interest. The data owners responded quickly, as the request fell within the scope of my project, which made it possible to obtain the continuous measures needed for my analysis.

PNC: A small number of records reported children younger than 10 years old as having offending outcomes. Since children under 10 cannot be held legally responsible for an offence in England and Wales, and the number of such cases was negligible, these records were dropped from the sample.

Suggested improvements recommended to data owners

It would be particularly valuable to consider the following improvements:

- **Accuracy of documentation:** While data dictionaries provide a useful overview of available variables and coverage periods, in practice I found several discrepancies between the documentation and the actual data. For instance, this was the case with the EYFSP and KS2 assessments. More consistent alignment between documentation and data content would greatly benefit researchers.
- **Provision of synthetic data:** Providing researchers with high-quality synthetic data that mirrors the structure and key features of the administrative data would be highly beneficial.

Additional data which would help to further develop the research

Although the dataset is already extremely rich, it would be strengthened by additional linkages to complementary sources, most notably the Longitudinal Education Outcomes (LEO) dataset. This would open up the possibility of following these children into later educational and labour market outcomes.

References

Department for Education. (2010, December): Early Years Foundation Stage Profile Attainment by Pupil Characteristics, England 2009/10 [Available at: <https://assets.publishing.service.gov.uk/media/5a7b875c40f0b62826a0429c/sfr28-2010.pdf>, Accessed: 17 December 2024]. Statistical Release.

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/Z502601/1]

Contact

Name: Dr. Paul Garcia

Email: pagarc@essex.ac.uk

