



Data Explained

An investigation of relationships between alternative school settings and youth offending

Ministry of Justice & Department for Education linked dataset - England

Author: Dr Liliana Belkin

Date: September 2026

DOI: <https://doi.org/10.5281/zenodo.22231798>

Summary

This Data Explained summarises experiences and learning from working with the Ministry of Justice & Department for Education linked datasets in the course of producing research into investigating relationships between youth offending and attending Alternative Provision, conceptualising AP as a ‘treatment’ or ‘intervention’ and exploratory analysis with causal inference methods. 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 Data First programme: a ground-breaking data linkage initiative, led by the Ministry of Justice (MoJ) and funded by ADR UK. The data used in this research project comes from the MoJ and Department for Education (DfE) and was accessed through the Secure Research Service (SRS) hosted by the Office for National Statistics (ONS). 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

Alternative provision (AP) is a form of education for pupils who are unable to attend mainstream school, often due to exclusion or a lack of effective support in mainstream settings. AP pupils are disproportionately drawn from groups already known to face elevated risk of youth justice involvement, including those with special educational needs and disabilities, those with social care involvement, and those from disadvantaged backgrounds.

There is strong existing evidence linking disrupted education with youth offending, and pupils who attend AP are known to experience poorer post-16 outcomes than their mainstream peers. However, the relationship between AP itself and offending remains unclear. It is not known whether AP contributes to offending risk, or whether it plays a protective role by offering more supportive environments than mainstream settings can provide for these pupils.

This project used the MoJ-DfE linked dataset to examine patterns of offending among young people born between 1993 and 2000, comparing those who attended AP with those who remained in mainstream education, in order to establish whether attending AP is associated with young people's offending, independent of the broader disadvantage this group tends to experience.

Initial research questions

1. Is exposure to alternative provision associated with young people's offending, compared to young people with similar characteristics in mainstream schools?
2. What factors are associated with offending, and what factors can be considered protective (i.e., reducing the likelihood of offending) for young people exposed to an alternative school intervention? Does this vary based on the ‘type’ of alternative school setting and duration in this setting?
3. Do specific ‘types’ of alternative school settings reduce the likelihood of offending/re-offending for the sample?
4. Is duration in alternative school settings associated with reducing the likelihood of offending/re-offending?

Research methodology

This project used a quasi-experimental approach, comparing a treatment group (young people who attended alternative provision) with a comparison group (young people with similar characteristics who did not attend alternative provision and remained in mainstream schools). This comparison focuses on individuals born between 1st September 1993 and 31st of August 2000 across five school years (SY2008/09-2012/13).

The project was exploratory and utilised causal inference methods, specifically [marginal structural modelling](#) to establish counterfactual populations to test outcomes for children/young people in alternative provision against those in mainstream schools. [Time-to-event analyses](#) was also used to examine time-to-offending for both groups within each school year.

Originally, the project also aimed to build upon previous work to develop a typology of alternative provision settings to classify them based on the type of programme offered (e.g. therapeutic, vocational, online). However, due to unforeseen limitations in the dataset variable descriptions, which will be discussed below, this was not fully developed.

Key variables

The key variables consisted of demographic and social characteristics (age, gender, ethnicity, etc.), school-related characteristics (exclusions, FSM eligibility, absenteeism, SEN status), entry/leaving dates for school settings, and offending dates.

National Pupil Database (NPD)

Datasets used	Variables used	Additional notes
SpringCensus_2009 SpringCensus_2010 SpringCensus_2011 SpringCensus_2012 SpringCensus_2013 APCensus (2007_08-) PRUCensus (2009_10-2012_13)	- academic_year - YearofBirth - Monthofbirth - Age - Gender - FSMeligible - Ethnicity; AP/PRU/Ethnicgroupmajor; AP/PRU/Ethnicgroupminor - Senprovisionmajor - URN/PRU_URN - AP_LA - IDACIdecile - Pru_/IDACIrank - Pru_/IDACIscore - Entrydate - PRU_Leavingdate - AP_Type Description	- PRU Leavingdate was missing - There were no exit/leaving dates within the AP Census for the years used - Ethnicity was collapsed into 6 broad categories (Black, Asian, Other, Mixed, Roma/Traveller, and White). - AP Type Description is labelled as <i>Hospital, Academy, Independent, Non-maintained School</i> and <i>Not a School</i>. - For <i>FSMeligible</i>, <i>Senprovisionmajor</i>, these were calculated as binary variables (Y=1, N=0) for each school year
Absences	- Persistentabsentee_3term - Persistentabsentee15_3term (available in XXX) - Entry_date - Leaving_date - URN_ab - IDACI_S - IDACI_R	
Exclusions	- URN_ex - Startdate_ex - Sessions_ex - Category_ex - IDACI_S - IDACI_R	CiN variables were used to derive two variables; CiN contact history (any CiN contact ever) and CiN contact in academic year. These were binary variables (Y=1, N=0).

CiN (2008_09-)	- CIN_ACADYR - CIN_CINReferralDate - CIN_CINClosureDate - CIN_AgeStartOfCINPeriod - CIN_AgeEndOfCINPeriod	
KS2	- Ks2_acadyr - Ks2_matlev - Ks2_englev - Ks2matpoints - Ks2engpoints - Ks2levxeng - Ks2levxmat	These variables were used to calculate an attainment score using Alice Wickersham's code .

Police National Computer (PNC)

Variables used	Additional notes
- MoJUID - CaseID - OffenceID - OffenceStartDate - CourtCautionDate - DisposalID - DiposalDuration - DisposalDays - DisposalRank - AdjudicationCode	In many instances OffenceStartDate was incorrect and corresponded to the individual's birthdate or was inaccurate as the offence start date recorded was before their birthdate

Summary of comments on specific variables

Variable name(s)	Comments
PRU_ LeavingDate (PRU Census)	This variable had no observations. It was not possible to determine when an individual left a PRU from this dataset.
AP_TypeDescription (AP Census) AP_LA AP_URN not available for (2008/09-2012/13)	This variable provides a minimal description of the AP 'type'. Combined with only the AP_LA and also likely non-existent URNs for AP in this time period, there is very little information about the AP setting or its programmatic details.
Entry_date (Absence) Leaving_date URN_ab	The Entry dates and leaving dates within the Absence dataset were often more reliable than in the Spring Census, however they did not always correspond with the PRU_entrydate and the URN_ab did not always correspond with the Spring Census URN or the PRU URN. There is no entry date in the AP Census, which made it difficult to determine when a pupil began attending an AP setting.
OffenceStartDate (PNC) CourtCautionDate	The date of an offence was critical for the analysis, however there was a fair amount of error in the OffenceStartDates.

How you dealt with data limitations

In order to conduct the analysis and based on unforeseen and unanticipated aspects of the datasets, I have to revise the research questions.

The revised research questions incorporate aspects of the initial research questions:

1. Is exposure to AP a protective factor for vulnerable young people's recent offending (within 1 year)?
2. Among young people who do offend, does the timing of the recorded offence in a school year differ between those in AP and those in mainstream schooling?

Issues with variable availability/missing/incorrect information

Related to both the revised RQ1 and RQ2, as leaving dates were not available in the PRU Census or AP Census, I relied heavily on the information in the Absence dataset to determine school placement. Additionally, if there was any individual in the PRU Census or AP Census within the study period (SY2008/09-SY2012/13), regardless of part-time or short-term placement (which was often difficult to determine), I considered these individuals to have had exposure to AP/treatment within the corresponding academic year.

Using Spring Census, PRU Census, AP Census and Absence datasets, and in some instances the Exclusions dataset, I attempted to trace when an individual was placed in AP/PRU and if or when they exited or returned to mainstream settings. The lack of availability of PRU leaving dates or AP entry/leaving dates for these datasets made it difficult to be precise about entry and exit points and how these related to recorded offences.

Ultimately, I had to exclude some individuals from the dataset for each school year due to lack of information or conflicting information. The conflicts often arose when the Absence dataset stated an entry date, no leaving date and a mainstream URN, but the PRU Census indicated a conflicting or overlapping entry date and a main registration at the PRU setting. Looking at the subsequent academic year, and tracing an individual, it was often possible to determine if they had remained in the PRU or returned to mainstream, however this was extremely time-consuming and there were issues with data wrangling to address the inconsistencies across multiple datasets.

In the case of incorrect offence start dates, I either removed these individuals from the analysis, as I did not have an accurate way to determine when (within a specific school year) an offence occurred or a first offence occurred (i.e, the offence start date was recorded before the individual birth cohort, therefore they would not have been alive). In some cases, I utilised the Court Caution Date instead of the offence start date to determine if an offence occurred in a specific school year. As it was not possible to determine the exact recorded offence date in these cases, I excluded these individuals from any analysis related to when an offence occurred, either before or after entering an AP/PRU (specifically for the time-to-event analyses).

Issues with metadata level of detail and variable specificity

For the initial RQ3, I planned to be able to use the information within the PRU Census and AP Census to determine a 'type' of Alternative Provision. Drawing from prior and existing [alternative education](#) and [AP typologies](#), I was planning to add to or refine these typologies and conduct a more nuanced analysis regarding the programmatic features and differences in offending outcomes for those with AP/PRU exposure. For instance, some APs offer more therapeutic, vocational, or academic programmes, and I wanted to investigate if these programmatic features made any 'difference' or had any influence on offending.

However, there are limitations in the AP Census (for the years of interest), where AP_URN did not exist, meaning it was impossible to know or add in any [GIAS](#) information or Ofsted information about a specific provision. The available 'type description' consisted of 'Hospital,' 'Academy,' 'Independent,' 'Non-maintained School' and 'Not a School'. These descriptors were not included in the metadata file and had to be located in other [DfE publications](#) to clarify what they connote.

Furthermore, 'not a school' can refer to many different types of provision and therefore was not very helpful for determining anything about the programmatic features of the setting. Therefore, the ability to address RQ3 was curtailed. Instead, I used broad 'school type' descriptions in the analyses, for instance 'PRU' and 'AP other' but I was unable to conduct a more nuanced analysis into programmatic features of the range of AP/PRU settings.

Suggested improvements recommended to data owners

Several of the limitations encountered in this project relate to gaps in entry and exit date recording across school-setting datasets. Leaving dates were not available in the PRU Census, and neither entry nor leaving dates were recorded in the AP Census for the years used in this analysis. This made it very difficult to establish precisely when a young person began or ended a placement in alternative provision, and required heavy reliance on the Absence dataset, which did not always align with the entry dates or setting URNs recorded elsewhere. Data owners could consider prioritising the consistent recording of entry and exit dates across all school-setting censuses, including AP and PRU Census, so that placement histories can be reconstructed more reliably without needing to cross-reference multiple, sometimes conflicting, data sources, however this would be retrospective and may not be possible.

The AP Type Description variable, which distinguishes AP settings as Hospital, Academy, Independent, Non-maintained School, or Not a School, offered very limited insight into the nature of a setting's programme. Combined with the unavailability of AP URNs during the years covered by this project, it was not possible to supplement this variable with more detailed information about a setting's provision, for example whether it was therapeutic, vocational, or academic in focus. It would be helpful for data owners to make AP URNs consistently available for any future school year-level data and if possible, to change the way this data is input in the first place, so that future researchers can link AP settings to more detailed provision records, and to include clearer definitions of category labels such as "Not a School" in the metadata.

It would also be helpful for descriptions within the metadata to include information regarding what criteria was used to label different 'type description' of AP provision.

More generally, offence start dates in the PNC were sometimes inaccurate, including instances where the recorded date preceded the individual's birth date. Additional validation of offence dates prior to release, or supplementary guidance on which fields (such as Court Caution Date) are more reliable substitutes, would reduce the need for researchers to exclude individuals from time-sensitive analyses.

Additional data which would help to further develop the research

Consistent, complete entry and leaving dates for all AP and PRU placements, across all available academic years, would substantially improve future work in this area, allowing researchers to determine placement duration and timing relative to offending with much greater precision.

Access to school provision URNs across all years and the ability to easily link this to the Get Information About Schools (GIAS) database, would also be of considerable value. This would allow researchers to draw on more detailed setting-level information, such as programme type, to explore whether particular models of alternative provision (for example, therapeutic, vocational, or academic) are associated with different offending outcomes, an aim this project was ultimately unable to pursue.

Works cited

Aron, L. (2002). *Towards a typology of alternative education programs: A compilation of elements from the literature*. Urban Institute, Washington, D.C. Available at: <https://files.eric.ed.gov/fulltext/ED480992.pdf>

Belkin, L. (2025, July 18). *Alternative provision and offending: Are there any connections?* ADR UK. <https://www.adruk.org/news-publications/news-blogs/article/alternative-provision-and-offending-are-there-any-connections/>

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: survival analysis. *Critical care (London, England)*, 8(5), 389–394. <https://doi.org/10.1186/cc2955>

Children's Commissioner of England (2025). *Special and alternative provision sector report: The Children's Commissioner's School Census*. Available at: <https://assets.childrenscommissioner.gov.uk/wpuploads/2025/12/cc-School-Census-Special-and-alternative-provision.pdf>

Department for Education. (2025). *Alternative provision: Statutory guidance for local authorities, headteachers, school staff and governing bodies* [Statutory guidance]. GOV.UK. Available at: <https://www.gov.uk/government/publications/alternative-provision>

Department for Education. (2026). *Alternative provision census 2027: Business and technical specification 1.0*. GOV.UK Available at: https://assets.publishing.service.gov.uk/media/6a3e4940eae00074150f5fb/Alternative_provision_census_2027_business_and_technical_specification_v1.0.pdf

Department for Education. (2025). *Get Information about Schools (GIAS)*. GOV.UK Available at: <https://get-information-schools.service.gov.uk/>

Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Pennacchia, J., Axler, M., King, S. & Clapham, A. (2025). Choice and diversity in governance in the English alternative provision sector: Implications for educational equity. *British Educational Research Journal*, 51, 1991–2012. <https://doi.org/10.1002/berj.4161>

Rollings, J., Cornish, R., Brennan, I., & Teyhan, A. (2025). Association between school exclusion, suspension, absence, and violent crime. Youth Endowment Fund Secondary Data Analysis Report. Available at: <https://youthendowmentfund.org.uk/wp-content/uploads/2024/12/Association-between-school-exclusion-suspension-absence-and-violent-crime.pdf>

Shinozaki, T., & Suzuki, E. (2020). Understanding Marginal Structural Models for Time-Varying Exposures: Pitfalls and Tips. *Journal of epidemiology*, 30(9), 377–389. <https://doi.org/10.2188/jea.JE20200226>

Disclaimer

This work was produced using deidentified administrative data accessed through the ONS Secure Research Service. The use of data in this work does not imply endorsement by the ONS or the data owners of the interpretation or analysis.

This work uses research datasets which may not exactly reproduce Accredited Official Statistics aggregates. Accredited Official 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/Z503228/1]. Thank you to Dr Arun Sondhi and to the Stakeholder Advisory Group for this project.

About ADR England

ADR England is a portfolio of data linking and research projects, delivered by a range of academic and government partners to provide policy-relevant insights using data held by UK Government departments and public bodies. The geographic scope of ADR England projects varies depending on the data holder, but is typically England-only, England and Wales, Great Britain or UK-wide. ADR England is part of ADR UK, an ESRC investment and part of UKRI.

About ADR 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. By linking together data held by different parts of government and facilitating safe and secure access for accredited researchers to these newly joined-up and de-identified data sets, ADR UK is creating a sustainable body of knowledge about how our society and economy function – tailored to give decision makers the answers they need to solve important policy questions.

Publication contact

Name: Dr Liliana Belkin
Email: liliana.belkin@roehampton.ac.uk

