Australian Government
Department of Education

The First Five Years: What makes a difference?

5.1 Methodology

This section details the data and methodology we employed to deliver the findings across the First Five Years project. This includes datasets, key analytical decisions, constructed factors and statistical techniques.

Contents

Data Linkage and Assembly

This project linked what was formerly known as Multi-Agency Data Integration Project (MADIP) data with Australian Early Development Census (AEDC) and Child Care Management System (CCMS) data.

MADIP has since been renamed the Person Level Integrated Data Asset (PLIDA). It is hosted by the Australian Bureau of Statistics (ABS) and is a partnership among Australian Government agencies to develop a secure and enduring approach for combining information on healthcare, education, government payments, personal income tax, and population demographics (including the Census) to create a comprehensive picture of Australia over time (ABS 2021a). This project used a customised MADIP extract that contained data from multiple government datasets (Table 1).

Table 1. First Five Years datasets

Dataset	Abbreviation	Source
Data Over Multiple Individual Occurrences	DOMINO	Department of Social Services
Data Exchange	DEX	Department of Social Services
Medicare Benefits Schedule	MBS	Department of Health
Medicare Enrolments Database	MEDB	Department of Health
Pharmaceutical Benefits Scheme	PBS	Department of Health
Census of Population and Housing		Australian Bureau of Statistics
Personal Income Tax	PIT	Australian Tax Office
Pay As You Go summaries	PAYG	Australian Tax Office
Registries Death Data		Australian Bureau of Statistics
Child Care Management System	CCMS	Department of Education
Australian Early Development Census	AEDC	Department of Education
National Quality Standard	NQS	Australian Children's Education and Care Quality Authority

The ABS provided the infrastructure and support to link these datasets and provided access to de-identified unit record files. Data is made available for analysis through the secure ABS DataLab environment (ABS 2021b).

We measured child developmental vulnerability using the AEDC. We measured child and family attributes that impact development using all available datasets.

Two cohorts were provided for analysis, 2015 AEDC and 2018 AEDC. These datasets were linked to the 2011 Census and 2016 Census respectively. For the First Five Years project we have published analysis on the 2018 cohort. The 2018 cohort was chosen as it is more recent, contains better quality data linkage and is closer to the respective Census. Data and methods described in this paper refer to analyses conducted with the 2018 AEDC cohort.

A custom child-carer relationship dataset was assembled that identified the child and their carer. This dataset identified the type of caring relationship (e.g., parent, grandparent, guardians) and the length of the relationship. The relationship scope dataset for the 2018 cohort used three data sources: Census 2016, DOMINO and CCMS. This dataset linked each child and carer to available MADIP data.

MADIP data was assembled to combine demographic data for individuals and key items including gender, birth, and death information. This combined demographic dataset for the 2018 cohort used five main data sources: MEDB, DOMINO, PIT, Census 2016, and Death Registrations. MADIP data was also assembled to combine location data for individuals. This combined location dataset identified residential location of individuals and the period they resided at each location. The location dataset for the 2018 cohort used four data sources: MEDB, DOMINO, PIT and Census 2016.

Developmental vulnerability

We explored developmental vulnerability in children using the five domains available on the AEDC (AEDC 2019a). The AEDC is nationwide data collection of early childhood development at the time children commence their first year of full-time school. The census involves teachers of children in their first year of full-time school completing a research tool, the Australian version of the Early Development Instrument. The Instrument collects data relating to five key areas of early childhood development domains. The domains and their descriptions are listed below (AEDC 2022):

- **Physical health and wellbeing** - Children's physical readiness for the school day, physical independence and gross and fine motor skills

- **Social competence** - Children's overall social competence, responsibility and respect, approach to learning and readiness to explore new things

- **Emotional maturity** - Children's pro-social and helping behaviours and absence of anxious and fearful behaviour, aggressive behaviour and hyperactivity and inattention

- **Language and cognitive skills (school-based)** - Children's basic literacy, advanced literacy, basic numeracy, and interest in literacy, numeracy and memory

- **Communication skills and general knowledge** - Children's communication skills and general knowledge based on broad developmental competencies and skills

The domain scores were originally developed by the Canadian Early Development Index. An Australian version was developed under license to form the AEDC (AEDC 2019b).

Primarily, analyses focused on the outcome of children being developmentally vulnerable on one or more domains (DV1). Some analyses also investigated developmental vulnerability in individual domains. We calculated the proportion of developmental vulnerability based on children who had a valid indicator. Individual indicators required a valid score in that domain exclusively. In the AEDC, the following criteria need to be met for a valid DV1 score:

- the child had at least five valid AEDC domain scores, or
- the child had one, two, three or four valid AEDC domain scores where at least one of these scores was categorised as developmentally vulnerable.

The proportion of children who were developmentally vulnerable on one or more domains was calculated by taking the number of children with at least one valid AEDC domain score which was categorised as developmentally vulnerable, divided by the number of those with a valid DV1 score (AEDC 2019c).

Another indicator, which focused on children who were developmentally on track on all domains, was also used in this project. The proportion of children who were developmentally on track on all domains was calculated by dividing the number of children who were classified as developmentally on track on all five of the AEDC domains by the total number of children who had valid domain scores in all five AEDC domains.

The definition of being developmentally on track on all domains used in this work has a slightly different definition from both the AEDC OT5 indicator (AEDC 2022) and the Productivity commission's (PC) definition for the Closing the Gap (CtG) dashboard. In 2021 the AEDC introduced a strength-based indicator: OT5 (developmentally on track on all FIVE AEDC domains) and the corresponding OT5flag which indicates whether a child qualifies for the base of the OT5 calculation. In addition to those with 5 valid domain scores, the base for OT5 includes children with less than five valid domain scores where the child has at least 1 valid domain score for which they are not developmentally on track. The PC's CtG target 4 reports the proportion of Aboriginal and Torres Strait Islander children assessed as developmentally on track in all five domains of the AEDC, where the total number of Aboriginal and Torres Strait Islander children in the first year of full-time school are used as the base of calculation (PC 2024).

Analysis cohort

From the full 2018 AEDC dataset we filtered to include children aged 4-6 at the time the instrument was conducted, with no identified special needs and at least one valid score in a domain. Children requiring special assistance because of chronic medical, physical or intellectually disabling conditions based on diagnoses (for example autism, cerebral palsy, Down syndrome) are defined in the AEDC as having 'special needs'. Teachers complete the AEDC for children with special needs but these children are not assigned a domain category (On track, At risk or Vulnerable) nor do they have a valid summary indicator score. Child age was determined by subtracting the age in months from the last month of AEDC data collection. These filters were applied using factors available in the AEDC dataset (Table 2).

Additionally, children who were not linked to MADIP and did not have a parent relationship recorded were filtered from the cohort.

This resulted in a final analytical cohort of 274,123 children (Table 2).

As a simplifying assumption, the only caring relationships analysed within the family unit were between the child and a linked parent(s). Informal or non-primary kinship care relationships were not analysed under the datasets and methodology. Parent relationship was sourced from the relationship scope dataset. Note that parent or carer highest educational attainment was taken directly from the AEDC rather than data associated with the linked parent, see Factors subsection.

This decision to focus on parent relationships was made for multiple reasons, including the difficulty reliably identifying primary caregivers, consistency with existing academic literature and our analyses frequently being centred on parent characteristics (for example household income, maternal age, parental education).

Based on the final analytical cohort of 274,123, the cohort number for individual analyses varies slightly due to the missing values in each domain of interest.

For the predictive modelling and G-computation and Inverse Propensity Weighting (IPW) modelling, multiple factors were analysed simultaneously (see *Predictive modelling* section and the Factors subsection within this *Methodology* section). Children with any factor with missing observations were removed. This resulted in a further reduction of the cohort (see Table 2). This does create a missing value bias that was a limitation of the study.

In particular, children on the AEDC who were unable to be linked to other datasets had higher rates of developmental vulnerability than the general population, and are expected to be less likely to use child care (because additional information collected in the CCMS data should increase the likelihood of linkage). This bias needs to be kept in mind when interpreting the unadjusted results, though is not expected to have the same impact on the modelling (predictive, G-computation or Inverse Propensity Weighting). See the Appendix for further details.

Table 2. Analytical cohort filters

Filter	Population size
Full AEDC cohort	308,873
Valid children (aged 4 to 6, no special needs)	294,605
MADIP-linked children	279,639
Children with parental data (analytical cohort)	274,123
Predictive modelling factors complete data	211,885
G-computation and IPW modelling attendance data	242,313
G-computation and IPW modelling quality data	232,693
G-computation and IPW modelling duration data	232,693

Source: Customised 'First Five Years' extract from the Multi-Agency Data Integration Project, 2021.

Notes: This table compares children from the 2018 cohort of the Australian Early Development Census.

Factors

Based on a literature review of factors known to impact child developmental vulnerability, we developed a list of factors to investigate. We derived these factors from the First Five Years MADIP Custom Extract. This process involved developing business rules to interpret the integrated administrative data and identifying data gaps and limitations of the proxy factors.

Cross-sectional effects

To understand the impact of certain characteristics across the first five years, many longitudinal factors were collapsed into cross-sectional factors.

Cross-sectional factors are time-insensitive – they aggregate all observations across the study period, presenting a summary value. Although longitudinal data was available within MADIP, given the inconsistency of collection across the different data sources and the extra technical complexity added, we aggregated these data to be represented by cross-sectional factors (for example, years in employment).

These cross-sectional factors included binary indicator factors as well as factors that were only measured at a single point in time (e.g. Census and AEDC data).

Child demographics

Gender

We identified children's gender using the AEDC gender factor. Only boys and girls were considered. Any other genders were excluded from analyses because of low sample size.

Language background other than English (LBOTE)

We identified children with LBOTE using the binary AEDC LBOTE factor.

Aboriginal and Torres Strait Islander status

We identified children as Aboriginal and Torres Strait Islander if they were recorded as Aboriginal and Torres Strait Islander in 50% or more of the data sources in which their Aboriginal and Torres Strait Islander status was known. There were four datasets used: AEDC, Census, DEX and CCMS. This method was developed in collaboration with the National Indigenous Australians Agency (NIAA) and follows the principles outlined in the National best practice guidelines (AIHW 2012).

Parent and family characteristics

Maternal age at birth

We identified maternal age at birth using the month and year of birth for both the mother and the child. Birth date information for mothers was sourced from the combined demographics dataset, while birth date information for children was sourced from the AEDC.

Country of parents' birth

We identified country of parents' birth using the country of birth indicator available in the combined demographics dataset. We categorised country of parents' birth into three groups: Australia, other Organisation for Economic Co-operation and Development (OECD), and non-OECD. When a child had parents from multiple categories, we prioritised categorisation in order of non-OECD, other OECD, Australia. This means that children with one parent born in an OECD country and one in Australia were categorised as OECD, while children with one parent born in a non-OECD country and one in an OECD country were classified as non-OECD.

Parent or carer highest educational attainment

We identified the highest level of educational attainment of a children's parents or carers using four factors from the AEDC: *Parent1School*, *Parent1PostSchool*, *Parent2School*, and *Parent2PostSchool*.

These factors recorded the level of high school and post-school education for up to two parents or carers for each child and were prefilled from school records collected when the child was enrolled into school. The order and other information prefilled into the AEDC did not give any identifying information including the gender of the parent or carer, whether they were a primary or secondary caregiver, or whether they were a parent or carer. While the AEDC provided the most current educational status of parents or carers available, it did not individually identify the parents or carers and therefore could not be directly linked to other sources of information about the parents.

For each child, we identified the highest level of education attained by each parent or carer. We then created two factors representing the highest educational attainment of the child's most educated parent or carer and the highest educational attainment of the child's least educated parent or carer.

The AEDC data contains a broad education level of Certificate level I to IV group, by treating four distinct education levels under the Australian Qualifications Framework (AQF) as one group (Department of Education, n.d.). The AQF does not include levels for Senior Secondary Certificate of Education qualifications; however, according to the International Standard Classification of Education 2011 to Australian Standard Classification of Education Concordance, Certificate I-II is below Year 12 (equivalent to Year 10) (DET 2019). The grouping of these levels with Certificate III-IV (higher than Year 12) may bias some results.

Unpaid child care

We identified a child as being exposed to unpaid child care if any parent self-reported providing unpaid child care for their own or other children in the past two weeks in the 2016 Census. Limitations of this definition are that unpaid child care may correspond to a different child and the Census only captures a single point in time.

Single parent household

We identified whether a child lived with one or two parents at the time closest to the completion of the AEDC. We considered the last known address in the period to be the effective address. Children with greater than two parents were out of scope and identified as unknown. To identify whether a child was living with one or two parents we made the following assumptions.

1. If a child and parent were living in the same Statistical Area Level 1 (SA1) they were assumed to be in the same household.
2. If a child only had one parent they were assumed to be part of that parent's household, regardless of whether they shared the same SA1.
3. If a child had exactly two parents and those parents lived in the same SA1 then the child was assumed to be part of that household.

We adjusted the single parent factor if tax data indicated presence of a spouse not otherwise identified. We then made a further adjustment to the factor based on DOMINO welfare parenting payments indicating a single or dual parent arrangement (Services Australia 2021).

Early childhood education and care

Formal child care attendance

We identified formal child care attendance using the Child Care Management System (CCMS) data. The CCMS reports quarterly charged hours of child care enrolment in long daycare centre (LDC), family day care (FDC) and outside school hours care (OSHC, noting this type of care was generally not included as the analysis focused on use before starting school). In our descriptive analysis of developmental vulnerability and rates of being developmentally on track by child care quality, duration or attendance, a child was identified as having attended formal child care if they had a record in the CCMS prior to the year they started school. A limitation of this definition is that it may include children who spent as little as one hour in child care and then never attended again.

In the modelling section, a simpler definition of child care was used whereby a child was identified as having attended child care with any record in the CCMS at all.

Note that the child care attendance here does not include ECEC attendance fully, with the lack of detailed information on government-run pre-school attendance.

Average weekly charged hours

We identified children's average weekly child care hours based on CCMS quarterly charged hours.

In other words, for each child, we found the average quarterly child care hours for all time, converted this value to average annual hours and divided through to find average daily hours. Finally, we multiplied by seven to find average weekly hours. Weekly hours are calculated by summing hours charged per quarter, dividing by the number of quarters the child had attendance data and multiplying by (4×7)/365.25.

The CCMS data included hours charged rather than hours attended. Business knowledge suggests that children attend for roughly 70% of the hours that families are charged. Therefore, it is likely that our *average weekly hours* factor is an overestimation of the amount of child care a child attended.

Quality

We identified child care quality using the Australian Children's Education & Care Quality Authority (ACECQA) National Quality Standards (NQS) (ACECQA 2021). The NQS ratings apply to all education and care services under the NQF and are not just a child care rating. The NQS data included quality ratings for seven quality areas (Education program and practice, Children's health and safety, Physical environment, Staffing arrangements, Relationship with children, Collaborative partnerships with families and communities, and Governance and leadership), in addition to an overall quality rating, for each child's most common care service provider per quarter. We created a quality factor that further simplified the overall quality rating into four categories:

1. Not yet at standard: quality ratings that did not meet the National Quality Standard. This included *Significant Improvement Required* and *Working Towards NQS ratings*.
2. At standard: quality ratings that met the National Quality Standard. This included the *Meeting NQS* rating.
3. Above standard: quality ratings that exceeded the National Quality Standard. This included *Exceeding NQS*, and *Excellent* ratings.

4. Provisional: not yet assessed under the NQF. At any one time a small proportion of services will only recently have been approved and may not have started operating or may have only been operating for a short period of time. In general, state and territory regulatory authorities will assess and rate newly approved services within 9-18 months of operations commencing.

From all the child care service providers with a non-provisional rating attended by a child, we identified the most frequently occurring quality rating for a child. Not all child care service providers had quality rating information. Our method for calculating modal quality is limited by CCMS and NQS data. NQS data for quality are recorded quarterly per child for only their most-attended child care provider. By linking the NQS data back to CCMS data we can also tell how many providers a child had per quarter and how many hours they attended in total. In the case that a child attended more than one provider, we cannot tell how the hours were divided between them.

Preschool attendance

Preschool school program is delivered to children in the year before they start full-time school. Different states and territories have different names for preschool services. In NSW, ACT and NT, the name preschool is used. In Victoria, Queensland, Tasmania and Western Australia, they are known as kindergartens, while in South Australia both preschool and kindergartens are used.

We defined a preschool attendance as having attended preschool as recorded by the teacher in the AEDC dataset or if the child had at least 600 hours of *Long Day Care* (LDC) in the CCMS in the year before school. Attendance for 600 hours at LDC in the year before school corresponds to the Universal Access National Partnership aim from both federal and state and territory governments to guarantee 600 hours of preschool or high-quality child care to all children (DESE 2021).

The CCMS tracks attendance hours on a quarterly basis. For the 2018 AEDC cohort, we summed the 2017 LDC quarterly hours in the CCMS. If a child had at least 600 hours of LDC in 2017 they were identified as having attended preschool.

The CCMS data included hours charged rather than hours attended, so this definition may overestimate the number of children who attend preschool through a centre-based day care provider.

Health and mental health

Child and parental mental ill-health

We defined child and parental mental ill-health using methodology advised by the Department of Health following the Australian Institute of Health and Welfare (AIHW) mental health report (AIHW 2021). This method used Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Scheme (PBS) data for items known to be associated with mental ill-health. An individual was identified as having mental ill-health if they accessed any mental health service or prescription in the relevant period (see below). If an individual had MBS or PBS data available but did not access specific items related to mental ill-health they were identified as not having mental ill-health. If an individual did not appear on MBS nor PBS they were identified as missing data.

The use of linked MBS and PBS data as a proxy for mental ill-health has limitations. A person accessing a mental health services or medication does not mean they have a diagnosed mental health condition. Not all individuals with mental health conditions will use government-supported services and no diagnostic information is captured in either dataset. Not all government-supported services are included, such as State and Territory or Primary Health Network services. While this is the best available measure we can derive from the available data, it is likely a measure of their access to services as well as whether people have mental ill-health. The relationship between having a mental ill-health and using MBS or PBS mental health services is likely to be different for different population groups: for example, people with lower incomes with high distress tend to be less likely to access MBS mental health services (Mental Health Australia 2023).

For MBS, we used Medicare items for:

- Psychiatric services;
- General Practitioner mental health services, such as a mental health plan;
- Clinical psychology services;
- Other psychology services; and
- Other mental health professional services.

For PBS, we used the Anatomical Therapeutic Classification codes:

- NO5A Antipsychotics;
- N05B Anxiolytics;
- NO5C Hypnotics and sedatives;
- N06A Antidepressants;
- N06B Psychostimulants and nootropics.

We identified a child as having mental ill-health if they accessed the above services between birth and 2018. We defined a child as having experienced parental mental ill-health if any of their parents had accessed the above services from one year before childbirth to 2018.

Years with parental mental ill-health

We identified long-term mental ill-health in parents by counting the number of years a parent had accessed mental health services as a proxy. If the parent accessed a mental health service in a given calendar year between one year before childbirth and 2018, this was identified as a year of access. The total years of access were then counted. For children with multiple parents, the parent with the longest mental ill-health period was taken as the relevant experience period for that child.

Age at mental ill-health onset

We defined a child's age-at-onset using their earliest recorded quarter of mental ill-health since birth. Parental mental ill-health age-at-onset was also calculated with respect to the child's age. Birth date information for children was sourced from the AEDC. For children with multiple parents, the parent with the earliest mental ill-health was taken as the age-at-onset.

Child and parent chronic ill-health

We identified child and parental chronic physical ill-health using a method developed by the Department of Health. This method used Medicare Benefits Schedule (MBS) data for items specific to chronic disease management. An individual was identified as having a chronic physical ill-health if they accessed any chronic disease management item (see below) during the relevant period. If an individual had MBS or PBS data available but did not access specific items related to chronic ill-health they were identified as not having ill-health. However, this factor should not be interpreted literally, as many (if not most) chronic health services would be delivered under other items, such as standard general practice consultations. If an individual did not appear on MBS nor PBS they were identified as missing data.

For MBS, we used Medicare items for:

- Chronic Disease Management;

- Allied Health.

Only MBS items for chronic ill-health related to physical ill-health were counted in the chronic health measure. There were a small number of cross-disciplinary MBS items that referred to both physical and mental ill-health and appeared in both derived factors. There are limitations with identifying chronic ill-health based only on use of MBS services for chronic disease management and allied health, as this approach may only capture a subset of those with chronic ill-health. Future work could explore using a rules-based approach for inferring conditions such as RX-Risk (Pratt et al. 2018), which uses PBS data.

We identified a child as having a chronic ill-health if they accessed the above services between birth and 2018. We defined a child as being exposed to parental chronic ill-health if any of their parents had accessed the above services from one year prior to childbirth to 2018.

Income and socio-economic status

Household income

Household or family income for each child was identified based on the sum of disposable parental income equivalised to the size of the household per financial year. Only parents identified in the relationship and location datasets who lived with the child that financial year were counted. We did not attempt to include income from siblings, other carers, or the children themselves in the household income.

The sum of the parents' income was aggregated from individual parent disposable incomes each financial year.

Disposable household income was used for consistency with external publications, including the ABS, Household, Income and Labour Dynamics in Australia (HILDA), OECD and the Smith family children's charity (ABS 2019; Wilkins et al. 2019; OECD 2021; Australian Council of Social Service 2018).

To calculate individual disposable income, we aggregated income from PIT, PAYG and DOMINO welfare for each individual parent closely following ABS methodology (ABS 2020; see *Appendix* for Individual income rules). Parents with no data across all three input sources were coded as missing.

For a minority of cases where we could only identify one parent's income and spousal income was present, we included the spouse income to the household total. For cases where two parents' income was reported we did not count spousal income, as we assumed the second income was the spouse. The Australian Tax Office (ATO) defines a spouse as a domestic partner residing together (ATO 2020) and based on this we included spouse dollars to the household total.

Spouse income data from PIT only had gross income factors available and did not include any net tax information so disposable income was not directly calculable. To estimate the tax paid, we took the average tax paid as a proportion of gross income for an individual in each tax bracket and applied that proportion to the spouse's gross income amount, to make an approximate net tax. We then subtracted the approximate net tax from the spouse's gross income to determine disposable spouse income.

Incomes are reported in June 2020 dollars, based on the Consumer Price Index (CPI) (ABS 2021c).

To identify parents that lived with the child we used shared SA1s for each financial year from the combined location dataset, as per the single-parent household derivation. The combined location dataset included time-sensitive, parent and child address data sourced from MEDB, DOMINO, PIT and Census 2016.

Household size was defined based on the count of parents (up to two parents) and children in the house each financial year. This number potentially changed each financial year.

We counted parents as per the relationship and shared SA1 data, as described above. When only one parent was identified we added a further parent if identified via spouse income. We made a further adjustment to the number of parents if DOMINO welfare parental payment data indicated this parent was in fact a single or couple parent (Services Australia 2021).

To count the number of children, we used the number of dependent children as per the PIT data each financial year. For multiple-parent families, the maximum value for the number of dependent children was taken as the number of children in the house. If this number was zero, we added one as the child in the AEDC cohort had to be counted.

To calculate household income, we summed the total disposable incomes of parents who lived with the child for that financial year, adding disposable spouse dollars for single parents. We then equivalised the total income by dividing it by the square root of household size (i.e. number of parents plus number of children).

The square root method for equivalising the household income was preferred as we did not have age data available for siblings, a requirement of other methods (OECD n.d.).

Household income decile

We identified the household income decile using household income across financial years from birth to completion of the AEDC. The household income decile was taken for each child for each financial year excluding the top 1 per cent and bottom 2 per cent of household incomes as per the top-tail ABS methodology (ABS 2019). To make a lifetime measure, we then took the most commonly occurring household income decile per child across all available financial years as our measure of household income decile.

Years with an employed parent

We identified children with employed parents for each financial year and created a lifetime factor that counted number of financial years a child had any employed parents. A child was deemed to have employed parents if any parent earned income by exertion (that is, excluding passive income streams) and lived with the child in a given financial year. Income by exertion was calculated using PIT and PAYG data following ABS methodology guidelines (ABS 2020; see *Appendix* for income rules). The total number of financial years the child had an employed parent was then counted between birth and the completion of the AEDC.

Neighbourhood SES

We identified the socio-economic status (SES) of the child's neighbourhood of residence using Socio-Economic Indexes for Areas (SEIFA) Index of Relative Advantage and Disadvantage (IRSAD). Developed by the ABS, IRSAD ranks geographical areas into 10 deciles based on relative social and economic advantage and disadvantage (ABS 2018a). We created an SES factor based on the Statistical Area Level 1 (SA1) of the child's residential address (ABS 2018b). When a child had no address but the parent(s) did, we used the SA1 of the parent's address. We defined neighbourhood SES as the child's lifetime minimum IRSAD decile, as it indicated at least some disadvantage during the child's life.

Welfare payments

We identified children who had parents who received welfare payments using the DOMINO dataset. As advised by Department of Social Services, welfare was categorised into non-exclusive payment groups based on receiving the following benefits:

- **Welfare receipt:** based on any benefit received.
- **Income Support:** Abstudy (Secondary/Tertiary), Age Pension, Austudy, Carer Payment, Disability Support Pension, Jobseeker Allowance, Newstart Allowance, Partner Allowance (historical), Parenting Payment Partnered, Parenting Payment Single, Partner Allowance, Sickness Allowance, Special Benefit, Sole Parent Pension, Widow Allowance, Wife Pension Age, Wife Pension DSP, Widow B Pension, Youth Allowance (apprentice), Youth Allowance (other), Youth Allowance (student), Youth Training Allowance.
- **Non-Income Support:** based on any benefit received that was not Income Support (above).
- **Rent Assistance:** Rent Assistance Family, Rent Assistance Parenting, Rent Assistance Newstart, Rent Assistance Pension, Rent Assistance Abstudy.
- **Disability Support Pension:** Disability Support Pension, Sickness Allowance.
- **Carer Payment:** Carer Payment, Carer Allowance.
- **Family Support:** Dad and Partner Pay, Double Orphan Pension, Assistance for Isolated Children, Family Tax Benefit Part A, Family Tax Benefit Part B, Parental Leave Pay, Parenting Payment Partnered, Parenting Payment Single, Sole Parent Pension.
- **Unemployment Payment:** Jobseeker Allowance, Newstart Allowance, Partner Allowance (historical), Partner Allowance, Special Benefit, Youth Allowance (other).
- **Student Benefits:** Abstudy (Schooling Applicant), Abstudy (Secondary/Tertiary), Austudy, Youth Allowance (apprentice), Youth Allowance (student), Youth Training Allowance.

- **Age Pension:** Age Pension, Widow Allowance, Wife Pension Age, Wife Pension DSP, Widow B Pension.

A child was identified as receiving a payment type if any parent received a relevant payment between the birth month of the child and the month of completion of the AEDC. Birth and AEDC completion date information was sourced from the AEDC.

Special child care benefit (SCCB)

We identified children whose parents received a SCCB using the following five SCCB identifiers in the CCMS, which indicated whether a child ever received any of the following:

- **Special Child Care Benefit:** A general factor for any child receiving special benefits. If a child had any of the following three benefits, then they also had this one too.
- **At-Risk Child Care Benefit:** A subset of the Special Child Care Benefit awarded to children who were identified as being at risk of negligence or abuse, as recognised by their child care provider.
- **Financial Hardship Child Care Benefit**: A subset of the Special Child Care Benefit awarded to children whose families were undergoing financial hardship, as recognised by their child care provider.
- **Grandparent Child Care Benefit:** A subset of the Special Child Care Benefit awarded to children who were cared for by their grandparents instead of their parents.
- **Jobs Education and Training (JET) Child Care fee assistance:** JET Child Care fee assistance provides extra help with child care costs for parents on income support while looking for work, studying or starting a job.

A note on exclusion of factors related to home learning

For our predictive and statistical modelling, and in any of our analysis, we opted to not include AEDC factors related to home learning environment because of potential post-treatment bias. While we understand there to be significant correlation between developmental vulnerability and other data items collected in the AEDC that do not directly contribute to the calculation of developmental vulnerability, we have chosen to exclude these factors from our analyses to avoid bias. We made this decision on advice from our academic partners at the Life Course Centre (personal communication, 5 July 2020):

> *The desirability of including these predictors is understandable. An extensive body of knowledge demonstrates the home learning environment, including activities such as reading to the child, and active engagement with schools is strongly associated with development of language, cognition, and social and emotional regulation. Thus, they are sources of intervention to improve child developmental outcomes and important confounders to consider when evaluating the effects of other covariates, such as childcare or socioeconomic status. The main reason we would advocate against using these measures, however, is that they are collected at the same time as the Australian Early Developmental Census (AEDC), whilst your covariates of interest are*

> *before the AEDC. Thus, they are post-treatment variables. Montgomery et al. (2018) provide detailed justifications for why this post-treatment adjustment should not be undertaken as it can bias estimates of causal (including descriptive or correlational) effects.*

The Vulnerable and Disadvantaged Children Project (Department of Education 2023), undertaken concurrently with this project, reported a strong positive correlation between developmental vulnerability and factors from Sections D and E of the AEDC, in particular, responses to the questions *E6 – has parent(s)/caregiver(s) who are actively engaged with the school in supporting their child's learning* and *E7 – is regularly read to/encouraged in his/her reading at home as far as you can tell*. Although the inclusion of these factors improved the performance of our models, we have excluded them for the reasons discussed above.

Analysis methodology

To investigate the factors influencing child development, we used descriptive analysis, predictive modelling, and statistical inference modelling. The differences between these techniques can be understood by the types of questions they are trying to answer.

Descriptive analysis seeks to understand how factors may be related to one another, for example *Are there differences in developmental vulnerability for children with different mental health status?*

Predictive modelling is used to achieve the best possible accuracy when trying to predict an outcome based on several factors, for example *How well can we predict developmental vulnerability?*

Statistical inference modelling answers 'what if?' questions, for example *What if we gave all children formal child care? How would this affect their development?*

Models used to assess the relative importance and relative effect sizes of our factors included logistic regression, random forest and causal inference. Our researchers used R (R Core Team 2020) to generate our results. RStudio (RStudio Team 2020) was used for R programming. Packages from these programs are also referenced throughout this document.

Descriptive analyses

Descriptive analysis was used to show trends between developmental vulnerability and other factors. This analysis compared frequencies and prevalence of developmental vulnerability between groups.

Predictive modelling using logistic regression and random forest

A stepwise logistic regression with Bayesian Information Criterion was used to select a reduced number of predictor variables. Using a both-direction regression, the algorithm combines both forward and backward steps, optimizing the model by adding significant variables and removing insignificant ones. This analysis used all available variables as listed in the methodology section, requiring all incomplete observations to be removed. As such, the sample size was reduced. This analysis was performed with R, using the *stats* package (R Core Team 2020).

The variables selected are used to generate two factors, namely relative importance of variables and the average marginal effects (AMEs), for their predictive power to model the developmental vulnerability of children.

A random forest model (Breiman 2001) was used to rank the importance of factors (selected from the stepwise logistic regress) for predicting developmental vulnerability for each gender (James et al. 2013). Permutation importance was calculated for each factor. To be *important* a factor must have a positive effect on the prediction accuracy. It involves calculating the drop in the predictive accuracy when the feature's values are replaced with randomly permuted values. This analysis was performed with R using the *caret* (Kuhn 2008) and *ranger* (Wright and Ziegler 2017) packages.

Average marginal effects (AMEs) for the indicators selected from the stepwise logistic regression average the effects of each indicator on developmental vulnerability across all children in the study. For example, the AME for 'preschool attendance' was calculated by taking the average of the effects of preschool across all children for all mental health, chronic health and other covariates to give an average percentage change in the likelihood of developmental vulnerability associated with this factor. This gave a broad overview of how each of the covariates impacted the likelihood of a child being developmentally vulnerable.

Estimate effect of child care using statistical inference techniques

The most robust approach to measure the effect of a treatment is to conduct a Randomised Controlled Trial (RCT), where people are randomly assigned to a treatment or control group and their outcomes measured afterwards. Due to the randomisation of treatment, any difference in outcome can be attributed to the different assignment rather than other differences in the populations (termed confounders), provided the sample size is large enough and the samples are truly random.

To establish a causal effect of child care, or any treatment, from observational data, additional analysis is needed which also requires the data to meet several identification criteria (Hernán and Robins 2020). Specifically, we need to meet:

- Exchangeability (no unmeasured confounding). This occurs when predictors of the outcome are equally distributed across treatment groups, as is assumed to occur in a randomised control trial. For observational studies this cannot be empirically tested, so researchers often try to approximate with enough confounders as informed by existing studies (Hernán 2012).
- Positivity: all individuals have a positive probability of receiving each treatment at every level of the confounders (Cole and Hernán 2008).
- Consistency: different ways of assigning a treatment have the same outcome (Cole and Frangakis 2009).
- No interference: a person receiving the treatment does not alter someone else's chance of receiving the treatment (Hernán 2012).

Once these criteria are met, an appropriate statistical analysis adjusts for confounding and is used to estimate the causal effect (Chatton et al. 2020). For example, regression modelling with inverse propensity weighting or G-computation.

However, statistical adjustment can only consider variables that are present in the data, unlike randomisation in an RCT, which should account for all other factors. Child characteristics, such as innate cognitive capabilities and parental characteristics such as quality of care provided, are difficult to measure or are not covered by available data. Thus, we would be unlikely to adjust for, or confirm the non-importance of, key confounding variables (violating exchangeability) and we cannot conclude causal effects for child care on developmental vulnerability.

Nonetheless, statistical methods are used to estimate the effect of child care on AEDC developmental vulnerabilities after adjusting for imbalance in confounding covariates. The results establish a better estimate of the effect of child care on developmental vulnerability; however, they require more research to be confirmed before the estimates could be considered causal.

We use two methods reducing confounding and compare their results: G-computation and Inverse Propensity Score Weighting (Tartaglia and Knapp (unpublished)). G-computation aims to compare treatments by estimating the outcome if every member of the study population were assigned to all the possible treatments available. This is achieved by first fitting a model to the data, then making copies of that data with each observation having a duplicate set of every different treatment. Predictions are then made as to the outcome for the copied data using the model fitted on the original data. Thus, for each person we get a result as though they were treated with every possible treatment. The difference between the treated/untreated outcome can then be used to estimate the average treatment effect and the risk ratio if everyone in the population is treated compared to a scenario where everyone is untreated. Inverse Propensity Weighting aims to use weights to adjust for the impact of covariates by upweighting or downweighting observations based on how common other factors are associated with the treatment group they are in. For example, if children from a high SES decile are more likely to attend child care they will be down-weighted proportionately, thereby adjusting for the effect of being in a high SES decile.

The two methods generally produce comparable results. In most cases in our report, only G-computation results were displayed. Studies on the effectiveness of inference techniques suggest that G-computation can be the best performing of the methods at discovering underlying causal relationships (Chatton et al. 2020). As such, G-computation results were given preference over Inverse Propensity Weighting results, noting the results were generally in agreement.

Selecting an appropriate set of confounders for which to control is critical for reliable effect estimation. This was based on available literature and in consultation with expert stakeholders. The covariates that were adjusted for in both causal inference techniques include whether a parent was a single parent or not, remoteness, child and parent mental health condition, child and parent chronic physical health condition, OECD country of origin of parents, Aboriginal or Torres Strait Islander status, highest education of parents, language background other than English, age of the mother at birth, household income, preschool attendance, years with an employed parent, socio-economic status of area of residence and gender.

References

ABS (Australian Bureau of Statistics) (2018a) *2033.0.55.001 – Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016*, ABS, accessed 5 January 2021.

ABS (2018b) *1270.0.55.005 – Australian Statistical Geography Standard (ASGS): Volume 5 - Remoteness Structure, July 2016*, ABS website, accessed 8 January 2021.

ABS (2019) *Household Income and Wealth, Australia methodology, 2017–18 financial year*, ABS website, accessed 11 February 2021.

ABS (2020) *Personal Income in Australia methodology, 2011–12 to 2017–18*, ABS website, accessed 23 February 2021.

ABS (2021a) *Multi–Agency Data Integration Project (MADIP)*, ABS website, accessed 21 December 2021.

ABS (2021b) *DataLab*, ABS website, accessed 21 December 2021.

ABS (2021c) *Consumer Price Index, Australia*, ABS website, accessed 23 February 2021.

ACECQA (Australian Children's Education and Care Quality Authority) (2021) *National Quality Standard*, ACECQA, accessed 10 December2021.

ACOSS (Australian Council of Social Service) (2018) *Inequality in Australia 2018 HTML*, ACOSS accessed 23 February 2021.

AEDC (Australian Early Development Census) (2019a) *About the AEDC domains*, AEDC website, accessed 18 February 2021.

AEDC (2019b*) Development of the AEDC*, AEDC website, accessed 23 February 2021.

AEDC (2019c) *FAQs for researchers*, AEDC website, accessed 23 February 2021.

AEDC (2022) *AEDC 2021 National Report,* AEDC website, accessed 3 September 2024.

AIHW (Australian Institute of Health and Welfare) (2012) *National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people*, catalogue number IHW 74, AIHW, Australian Government, accessed 3 September 2024.

AIHW (2021) *Mental health services in Australia*, AIHW website, accessed 23 February 2021.

ATO (Australian Taxation Office) (2020) *Spouse details – married or de facto 2020*, ATO website, accessed 23 February 2021.

Breiman L (2001) '*Random forests*', *Machine Learning*, 45(1):5–32, doi:10.1023/A:1010933404324.

Chatton A, Le Borgne F, Leyrat C, Gillaizeau F, Rousseau C, Barbin L, Laplaud D, Léger M, Giraudeau B and Foucher Y (2020) 'G–computation, propensity score–based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study', *Scientific Reports*, 10(1):9219, doi:10.1038/s41598–020–65917–x.

Cole SR and Frangakis CE (2009) 'The consistency statement in causal inference: a definition or an assumption?', *Epidemiology*, *20*(1), 3–5, doi:10.1097/EDE.0b013e31818ef366.

Cole SR, and Hernán MA (2008) 'Constructing inverse probability weights for marginal structural models' *American journal of epidemiology*, *168*(6), 656–664, doi:10.1093/aje/kwn164.

DET (Department of Education and Training) (2019) *International Standard Classification of Education 2011 (ISCED 2011) to Australian Standard Classification of Education (ASCED) Concordance*, accessed January 2024.

Department of Education (n.d.) *AQF Qualifications*, Australian Qualifications Framework website, accessed January 2024.

Department of Education (2023) Measuring vulnerability and disadvantage in early childhood data collections, Department of Education.

Hernán M (2012) ‘Beyond exchangeability: the other conditions for causal inference in medical research’, *Statistical Methods in Medical Research*, *21*(1), 3–5, doi:10.1177/0962280211398037.

Hernán M and Robins J (2020) *Causal inference: What if*. Chapman & Hall/CRC, Boca Raton.

James G, Witten D, Hastie T and Tibshirani R (2013) *An introduction to statistical learning,* Springer, New York.

Kuhn M (2008) *‘Building predictive models in R using the caret package’*, *Journal of Statistical Software*, 28(5):1-26, doi:10.18637/jss.v028.i05.

Mental Health Australia (2023) *Mapping mental health care*, Mental Health Australia, accessed 3 September 2024.

Montgomery JM, Nyhan B and Torres M (2018) ‘How conditioning on post treatment variables can ruin your experiment and what to do about it’, *American Journal of Political Science*, 62(3):760-775, doi:10.1111/ajps.12357.

OECD (Organisation for Economic Co-operation and Development) (n.d.) “Household income”, in Society at a Glance 2024: OECD Social Indicators, OECD Publishing, doi:10.1787/c8568e7f-en.

OECD (2021) *Household disposable income (indicator)*, OECD website, accessed 23 February 2021, doi:10.1787/dd50eddd-en.

Pratt NL, Kerr M, Barratt JD, Kemp-Casey A, Ellett LMK, Ramsay E, Roughead EE (2018) ‘The validity of the Rx-Risk Comorbidity Index using medicines mapped to the Anatomical Therapeutic Chemical (ATC) Classification System’, BMJ Open 8:e021122, doi:10.1136/bmjopen-2017-021122.

Productivity Commission (PC) (2024) Socio-economic outcome area 4: Aboriginal and Torres Strait Islander children thrive in their early years, accessed 1 October 2024.

R Core Team (2020) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team (2020) *RStudio: Integrated Development for R*, RStudio, PBC, Boston, MA.

Services Australia (2021) *Parenting Payment – who can get it*, Services Australia website, accessed 23 February 2021.

Tartaglia E and Knapp S (unpublished) ‘Causal Inference in Education Data’, *Australian Government Department of Education, Skills and Employment (DESE)*.

Wilkins R, Laß I, Butterworth P and Vera-Toscano E (2019) ‘The Household, Income and Labour Dynamics in Australia Survey: Selected Findings from Waves 1 to 17’, *Melbourne Institute: Applied Economic & Social Research*, University of Melbourne, accessed 23 February 2021.

Wright MN and Ziegler A (2017) ‘Ranger: A fast implementation of random forests for high dimensional data in C++ and R’, *Journal of Statistical Software*, 77(1):1–17, doi:10.18637/jss.v077.i01.