



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Longitudinal reading measures and genome imputation in the National Child Development Study

**Citation for published version:**

Bridges, EC, Rayner, NW, Mountford, HS, Bates, TC & Luciano, M 2023, 'Longitudinal reading measures and genome imputation in the National Child Development Study: Prospects for future reading research', *Twin Research and Human Genetics*, pp. 1-11. <https://doi.org/10.1017/thg.2023.2>

**Digital Object Identifier (DOI):**

[10.1017/thg.2023.2](https://doi.org/10.1017/thg.2023.2)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Publisher's PDF, also known as Version of record

**Published In:**

Twin Research and Human Genetics

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.






**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



## Article

# Longitudinal Reading Measures and Genome Imputation in the National Child Development Study: Prospects for Future Reading Research

Elinor C. Bridges<sup>1</sup> , N. William Rayner<sup>2,3,4</sup> , Hayley S. Mountford<sup>1</sup> , Timothy C. Bates<sup>1</sup>  and Michelle Luciano<sup>1</sup> 

<sup>1</sup>School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom, <sup>2</sup>Institute of Translational Genomics, Helmholtz Zentrum München — German Research Center for Environmental Health, Neuherberg, Germany, <sup>3</sup>Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom and <sup>4</sup>Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

## Abstract

Reading difficulties are prevalent worldwide, including in economically developed countries, and are associated with low academic achievement and unemployment. Longitudinal studies have identified several early childhood predictors of reading ability, but studies frequently lack genotype data that would enable testing of predictors with heritable influences. The National Child Development Study (NCDS) is a UK birth cohort study containing direct reading skill variables at every data collection wave from age 7 years through to adulthood with a subsample (final  $n = 6431$ ) for whom modern genotype data are available. It is one of the longest running UK cohort studies for which genotyped data are currently available and is a rich dataset with excellent potential for future phenotypic and gene-by-environment interaction studies in reading. Here, we carry out imputation of the genotype data to the Haplotype Reference Panel, an updated reference panel that offers greater imputation quality. Guiding phenotype choice, we report a principal components analysis of nine reading variables, yielding a composite measure of reading ability in the genotyped sample. We include recommendations for use of composite scores and the most reliable variables for use during childhood when conducting longitudinal, genetically sensitive analyses of reading ability.

**Keywords:** Reading; genotype imputation; cohort study; principal components analysis; longitudinal measures

(Received 5 September 2022; revise received 30 December 2022; accepted 17 January 2023)

Poor functional reading skill is a global problem with high prevalence, even in economically developed countries. For example, 21% of the population of the United States have poor English literacy skills (Mamedova & Pawlowski, 2019), and 16.4% of the adult population in England struggle with reading unfamiliar material (National Literacy Trust, ca. 2017). The annual global cost of poor literacy skills is estimated to be £800 billion (World Literacy Foundation, 2018). It is imperative to identify the causes of reading skill variation so that early identification and intervention in cases of potential reading disorder is possible, which can improve both word-reading and comprehension skills (Snowling & Hulme, 2011). This could reduce related negative outcomes, such as lower educational attainment and unemployment (Currie & Thomas, 1999). Cohort studies are particularly valuable for determining the causes of reading ability if they have extensive longitudinal data from birth and linked genetic data. Here, we outline a UK birth cohort study, the National Child Development Study (NCDS), that has a wide range of reading measures in childhood and measures of

functional reading ability in adulthood, along with genotyped data in a large subsample. Data collection began in 1958, and there is a rich variety of environmental variables collected at all age points. We present a series of carefully selected reading measures that we suggest for future studies, along with the protocol for imputation of the genotype data. The imputed data is available for researchers to access from the Centre for Longitudinal Studies.

Longitudinal studies of reading are required to build a complete picture of predictors of reading ability from childhood through to adulthood. The majority of existing longitudinal reading studies have been conducted across short time periods in early childhood (Psyridou et al., 2020). This approach overlooks two reading trajectories: those who read well in early childhood and begin to develop reading problems later, and those who have difficulties in acquiring reading skills in early childhood but go on to read normally (Catts et al., 2012). This provides clear incentive for use of longitudinal reading datasets that extend beyond early childhood. Additionally, several studies have identified that environmental circumstances in early childhood, such as socioeconomic status (SES) and maternal education, are associated with reading outcomes years later (e.g., Russell et al., 2016; Senechal, 2006; Williams & Silva; 1985). While these existing studies have generated insight into early predictors, they do not reach adulthood and

**Author for correspondence:** Michelle Luciano, E-mail: [michelle.luciano@ed.ac.uk](mailto:michelle.luciano@ed.ac.uk)

**Cite this article:** Bridges EC, Rayner NW, Mountford HS, Bates TC, and Luciano M. Longitudinal Reading Measures and Genome Imputation in the National Child Development Study: Prospects for Future Reading Research. *Twin Research and Human Genetics* <https://doi.org/10.1017/thg.2023.2>

© The Author(s), 2023. Published by Cambridge University Press on behalf of International Society for Twin Studies. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



potential lifespan sequelae of reading difficulty. In general, reading fluency appears to stabilize around adolescence (Lohvansuu *et al.*, 2021) with little or no changes during middle age (Reder, 2012). However, small to moderate gains can still be made in adulthood with targeted intervention (Sabatini *et al.*, 2011), and individual differences in rate of change have been identified (Lechner *et al.*, 2021), so it is important to track the trajectory of reading skill over the lifespan. The NCDS contains reading skill data for a large, genotyped subsample up to and including adulthood, allowing longer term longitudinal studies to be conducted.

Comprehensive reading studies should use environmental data to build a full picture of predictors. Some identified predictors of reading ability are physiological in nature (e.g., Leppänen *et al.*, 2010; Lohvansuu *et al.*, 2021), but the majority of early known predictors involve either the Home Literacy Environment (HLE) or home and family circumstances more broadly. For example, association of low SES with poorer reading related skills has replicated across studies and cultures (e.g., Fernald *et al.*, 2012; Fung & Chung, 2019; Molfese *et al.*, 2003; Russell *et al.*, 2016; Williams & Silva, 1985), and literacy activities and experiences gained in the home have repeatedly predicted reading related skills (e.g., Hamilton *et al.*, 2016; Russell *et al.*, 2016; Senechal, 2006). Many predictors of reading ability, including housing tenure, single parenthood, maternal education, and reading to the child (Russell *et al.*, 2016) are present in the NCDS.

While these studies each go some way to untangling the nature of the predictors of reading ability, each fails to control for potential genetic confounding. This prevents us from understanding whether the aforementioned environmental variables are causal (Kendler & Baker, 2006). Researchers have provided clear evidence that genetically informed reading studies are required to complete the picture. For example, one study has shown that maternal reading and language skills are positively correlated with storybook exposure in the home, and once these skills are controlled for, storybook exposure no longer acts as a significant predictor of childhood reading, spelling and language skill (Puglisi *et al.*, 2017). This suggests a heritable component acting on childhood reading skill, demonstrated by the maintained association of the mother's skill level with the skills of the child. The nature of these associations can be clarified by including genotype data for genetically informed reading studies, which can be used to control for genetic confounding.

Twin studies confirm that reading ability has a substantial heritable component (e.g., see Bates *et al.*, 2007; Byrne *et al.*, 2005; Wadsworth *et al.*, 2007) with an estimated heritability of approximately .54–.73 in adolescents (Bates *et al.*, 2004; Wadsworth *et al.*, 2015). In recent years, genomewide association studies (GWAS) have been conducted to identify specific genetic variants that are associated with reading skill (Luciano & Bates, 2019). The largest of these was in ~34,000 individuals and identified one locus of genomewide significance associated with quantitative word-reading, along with a significant SNP-based heritability estimate (Eising *et al.*, 2022). The largest GWAS of dyslexia, which might be considered the low extreme of reading skill (Hulme & Snowling, 2016), has identified 42 associated independent variants (Doust *et al.*, 2022). The NCDS dataset has the potential to contribute to GWAS meta-analysis, increasing sample sizes and allowing for greater power (Panagiotou *et al.*, 2013).

The influence of genetics on reading ability is an area that has been explored using twin cohort studies. Many of these studies point towards stability in the genetics of reading ability in childhood and adolescence. This is a finding that has been replicated with different

methods. For example, use of Cholesky decomposition models have shown high genetic correlations across late childhood and adolescence for reading ability (Betjemann *et al.*, 2007; Wadsworth *et al.*, 2001). Use of DeFries-Fulkner regression has indicated that persisting reading difficulties are due to the same genetic components, and between 60–75% of stability in reading difficulties may be due to genetic influence (Astrom *et al.*, 2007; Wadsworth *et al.*, 2015, 2016). Similarly, latent growth models in twins have shown that genetic influence on reading ability at the age of 6 is related to reading performance through to age 8 (Petrill *et al.*, 2010), and has also indicated that no new genetic factors appear for reading ability between the ages of 6 and 12 (Logan *et al.*, 2013), although a longitudinal genetic model from Erbeli *et al.* (2017) showed a new genetic factor appearing after kindergarten, which influences word reading at first grade and comprehension at seventh grade. Similarly, Ebejer *et al.* (2010) showed that a second genetic factor becomes active on reading ability in Grades 1 and 2, compared to kindergarten. A similar result was demonstrated by Samuelsson *et al.* (2008) in samples from Scandinavia and the United States; however, no additional genetic factor was found in an Australian sample. The authors suggest this may be due to international differences in schooling. These discrepancies show that while we are starting to build a picture of the changing influence of genes on reading over the life-course, most studies have focused on childhood, with a few moving into adolescence, and they have not utilised modern molecular methods.

Data from genotyped individuals can be used to ask and answer more complex questions about prediction of reading ability. For example, the data can be used for polygenic prediction, in which the additive variants of an individual's genome can be combined to indicate their genetic propensity for high reading ability (Luciano, 2017). A polygenic score (PGS) based on dyslexia explained up to 6% of variance in quantitative reading measures in samples both enriched and not enriched for poor reading (Doust *et al.*, 2022). Selzam *et al.* (2017) showed that the proportion of variance in reading ability explained by an educational attainment PGS increased with the age of the child, highlighting the importance of longitudinal data in genetic studies. In another study, PGSs calculated for intelligence, educational attainment, attention deficit hyperactivity disorder (ADHD) and bipolar disorder were found to be significantly correlated with word reading ability (Price *et al.*, 2020). PGS can further be used for gene by environment interaction ( $G \times E$ ) and gene-environment correlation ( $rGE$ ) research over the life course. Whereas generation of PGS for reading ability for gene-environment interplay research has not yet been conducted, the phenotype of educational attainment provides an example; one study has found that environmental factors mediated approximately 40% of the impact of an educational attainment PGS (Allegrini *et al.*, 2020).

In sum, the NCDS 1958 Birth Cohort contains a range of reading and reading related measures throughout childhood, and follow-up variables focusing on functional literacy in adulthood, providing a valuable resource for reading studies. The sizeable subsample for which genetic data is available, along with the richness of other variables collected through the life-course, makes it a potentially valuable resource for genetically sensitive longitudinal reading studies that has been underutilised thus far. The imputed genotyped subsample has, to date, only been available with the 1000 Genomes reference panel (Davies *et al.*, 2015). In this article, we present the NCDS genetic dataset imputed using the more recent Haplotype Reference Consortium r1.1 panel, which offers greater imputation quality (Haplotype Reference Consortium, 2016). Conducting genetic research with NCDS genotype data is

challenging, due to the genotyping of data on multiple chips and a lack of centralised documentation. To aid in future studies, we clearly demonstrate the quality control and processes that took place prior to imputation, so that future researchers may use this updated resource in their own work. In addition to preparation of genetic data, this article presents an overview of reading and reading-related measures available in this dataset from age 7 to age 33. Matched genetic and phenotypic data must be requested from the data holders (Centre for Longitudinal Studies, [n.d.-a](#)). The full sample of NCDS phenotypic data is openly available through the UK Data Service (UK Data Service, [n.d.](#)).

## Material and Methods

### About the NCDS Dataset

The NCDS, also known as the 1958 British Birth Cohort, is a national cohort study that surveyed the parents of babies born during one week in the year 1958 (Power & Elliot, 2005). There have been several follow-ups, which are still ongoing (Centre for Longitudinal Studies, [n.d.-c](#)), and the length of time for which this cohort study has been running makes this an excellent dataset for tracing longitudinal patterns and associations. Data for 17,416 births were collected, with 9137 respondents at the last completed wave at age 55 (Centre for Longitudinal Studies, [n.d.-b](#)). For a sub-sample of the participants, a biomedical survey was also conducted at age 44, which included the collection of DNA samples (Power & Elliot, 2005). This article presents the preparation and imputation of 13,738 overlapping genotyped samples (resulting in a final total of 6431 unique individuals) collected from the NCDS participants on seven different arrays (Table 1). This imputed dataset is available for researchers to access from the Centre for Longitudinal Studies. Access to linked phenotype and genetic data is dependent on research proposal approval.

### Genotyped Data

#### Quality Control Procedure

All quality control was conducted using Plink v1.90b4, R v3.3.2 and RStudio v4.1.2. Code used to carry out these processes, along with further information regarding externally developed scripts and resources, is publicly available as a GitHub repository (Bridges, 2022). Quality control was carried out on each of the seven genetic datasets according to the following steps. SNPs with a call-rate of less than 98% were removed (Turner et al., 2011), and SNPs that were not in Hardy-Weinberg equilibrium were removed at a threshold of  $p < 1 \times 10^{-6}$ . Individuals failing quality control were also removed. Those with a genotyping call-rate of less than 97% were then removed (Wellcome Trust Case Control Consortium, 2007). Individuals showing unexpected levels of heterozygosity ( $\pm 3$  SD from the mean) were removed.

Each dataset was updated to GRCh37 build for consistency, using up-to-date strand files and a series of commands collated in the script Update Build (Robertson, 2012; see Supplementary Information A for further detail). Heterozygous haploid errors were present in six of the seven chips. They were removed by first ensuring that all variants in the pseudo-autosomal region had the correct chromosome code for the Plink format, and any remaining errors were set to missing. Any individuals with discrepant or ambiguous sex data were removed, excluding the Affymetrix 500K chip, for which X chromosome data was not available. A small number ( $n = 8$ ) of related samples (representing 6 individuals) were identified using a relatedness threshold of 0.1875, and

**Table 1.** Breakdown of number of participants sampled on each array in the NCDS after removal of exclusions and duplications, and number of SNPs sequenced in each dataset

Array	N	SNPs
Illumina 1.2M	2908	1157986
Illumina 15k Custom Chip	1475	9803
Illumina Human 660-Quad	871	582892
Infinium HumanHap 550K v1.1	1436	555174
Infinium HumanHap 550K v3	2592	561303
Affymetrix 500k	1477	490032
Affymetrix v6	2979	934967

the sample with the most missing data from each related pair was removed. The computational burden to control for genomic relatedness of six cases was deemed too high to warrant their inclusion, and has the advantage that users will not need to check for relatedness prior to their own analysis.

To identify genetic ancestry outliers, the data was merged with 1000 Genomes reference data (The 1000 Genomes Project Consortium, 2015), and Principal Components Analysis (PCA) was conducted in Plink to identify outliers ( $n = 17$ ). Outliers were removed according to the procedure documented by Meyer (2021a, 2021b), using a theta value of 3.

### Data Preparation and Imputation

The cleaned data were checked against the HRC reference panel r1.1 site list (Haplotype Reference Consortium, 2016) for strand issues, using a specifically developed Perl script named HRC-1000G-check-bim.pl, v.4.3.0 (Rayner, 2020). This script checks for strand inconsistencies between the data and the reference panel, and generates a series of Plink commands that can then be used to remove or update any problematic loci. Perl v5.24.0 was used, followed by Plink to make the aforementioned necessary changes. Sorted VCF files were generated for each chromosome using BCFtools (Danecek et al., 2021). A final set of quality control checks were carried out on the VCF files post-conversion, using a specifically developed Python script (Zhan & Liu, 2016). These checks include identifying duplicated sites, invalid genotypes and NonSNP sites. Python v2.7.10 was used. No issues requiring further action were identified. The data were uploaded to the Michigan Imputation Server for quality control and imputation (Das et al., 2016). Imputation was carried out against the European population of the HRC r1.1 2016 reference panel, using the Minimac4 pipeline. Eagle v2 phasing was used.

### Post Imputation Quality Control

Following imputation, quality control checks were carried out using the 'ic' script developed by Rayner (2016), which incorporates a series of Perl commands to check alternate allele frequencies and imputation quality (Rayner, 2016). This output can be used to evaluate the quality of an imputed dataset, including alternate allele frequency counts and frequency counts of the  $R^2$  imputation quality score by chromosome.

There was considerable overlap of individuals between arrays, with many individuals having been genotyped on multiple arrays. After imputation, we determined the number of unique samples

that should be retained on each array by retaining the highest quality sample for each individual (Table S16). This was determined by listing all samples in order of chip quality, determined by the number of SNPs with an information score  $>0.8$  after imputation (Al-Soufi et al., 2021). The values in Table S16 represent the number of individuals that were retained on each chip after removal of lower quality duplications. It should be noted that duplicates were not removed in the final imputed dataset that is available for researchers.

## Phenotype Data

### Measures

**Age 7.** Three reading measures are available at age 7 in the NCDS. (1) The Southgate Group Reading Test is a 30-item word reading test that assists in the identification of participants with poor reading skills (Shepherd, 2012). The test was distributed in school by the participant's teacher, and one point was awarded for each correct answer (Shepherd, 2012). Of the 30 questions, 16 required the child to correctly circle the word from a list that corresponded to an image, and fourteen required the child to correctly circle the word from a list that corresponded to a word read aloud by their teacher (Shepherd, 2012). Little further information is available about the test; however, a small amount of research on reliability has been conducted. The Southgate Test manual reports parallel reliability of 0.95 (Southgate, 1958, as cited in Tizard et al., 2002). One study has found that the Southgate reading test shows relatively high concurrent validity, as shown by a correlation of .72, with the Concepts About Print test, which measures several early reading-related skills such as book orientation and the relationship between written and oral language (Sultmann et al., 1983). (2) The child's usual teacher was asked to rate the child's reading ability compared to other children of their own age on a 5-point scale (Teacher Rating). The available options were *Avid reader. Reads fluently and widely in relation to his age; Above average ability. Comprehends well what he reads; Average reader; Poor reader. Limited comprehension; Non-reader, or recognises very few words.* (3) The child's usual teacher was asked to identify which level of book the child had reached in a reading scheme (Reading Level). Response options included *Don't know or inapplicable; On prereading activities only; At present on Book 1 or introductory book; At present on Book 2; At present on Book 3; At present on Book 4; Beyond basic reading scheme.* The option *Don't know or inapplicable* was coded as missing (full sample  $N = 148$ , subsample  $n = 66$ ), leaving a 6-point scale. No further information was available on this variable.

**Age 11.** Four reading related measures are available at age 11. (1) A 35-item reading comprehension test was administered by the participant's school, based on the 1947 Watts-Vernon test of reading ability (Shepherd, 2012). For each item, the participant was asked to choose a word from a selection of five in order to complete a sentence, and one point was awarded per correct answer (Shepherd, 2012). (2) The child's teacher was asked to rate the child's use of books compared to other children their age on a 5-point scale (Book Use). Response options included *Exceptional. Reads very widely for pleasure and information; Above average. Turns to books very readily; Average. Skill and comprehension satisfactory for school requirements; Below average. Still learning the skill of reading, not inclined to turn spontaneously to books for pleasure or information; Very poor or non-reader. Recognises few words, very limited use of books because of poor skill.* (3) The participant was asked how often they read books outside of school

work, and was given the option to respond with *often (nearly every day), sometimes, or never or hardly ever* (Reads Books). (4) The participant was asked how often they read magazines, newspapers and comics. The response options were *often (nearly every day), sometimes, and never or hardly ever* (Reads Other).

**Age 16.** There are five reading related measures available at age 16. (1) The same reading comprehension test was administered as at age 11 (Shepherd, 2012). (2) The participant's teacher was asked if the participant could read well enough to cope with everyday needs (Can Cope). This was presented with the options *yes, no, and uncertain*. Uncertain responses were removed from analysis in order to create a binary variable (full sample  $N = 20$ , subsample  $n = 8$ ). (3) The participant was asked whether they often read books outside of school work, and was provided with the categorical response options *often, sometimes, never or hardly ever, and like to but no chance. Like to but no chance* (full sample  $N = 388$ , subsample  $n = 152$ ) was removed from analysis due to its categorical nature, leaving three ordinal categories (Reads Books). (4) The participant's teacher was asked to rate the English ability of the participant from the options *Capable of obtaining an A-level or Higher-grade pass in this subject; Above average. Capable of obtaining O-level or O-grade or CSE grade one; Of average ability in this subject. Capable of obtaining a CSE pass, grades 2–4; Below average. A possible CSE entrant; Little, if any, ability in this subject; Don't know* (English Ability). Those who responded *Don't know* were set to missing to ensure an ordinal set of responses (full sample  $N = 32$ , subsample  $n = 17$ ). (5) The participant was asked to rate their ability in English compared to other people of their age, from the options *never studied, below average, average and above average* (English Rating). *Never studied* responses were removed (full sample  $N = 61$ , subsample  $n = 24$ ).

**Age 23.** At age 23 all participants were asked if they had had problems with reading since they left school (Reading Problems). Possible response options were *yes, no and don't know*. *Don't know* responses were removed from analysis (full sample  $N = 27$ , subsample  $n = 17$ ). The follow-up question was a binary response variable asking whether their problems made things difficult in everyday life. This question was excluded from analysis because data were limited by the screening question. An open text question allowed participants to elaborate on what difficulties they faced. Participants were also asked whether they had attended any courses to improve their skills. This variable was also excluded in this analysis, for two reasons; first, limited data due to the screening question, and second, attendance at a course may not be representative of need as there are multiple practical barriers to attendance, including lack of temporal and financial resources (BSA, 2000, as cited by Melrose, 2014).

**Age 33.** At age 33, participants were again asked whether they had had problems with reading since they left school. If they responded yes, they were asked a series of follow-up questions regarding which common activities they had difficulty with due to their reading problems. Respondents were asked whether they could usually read and understand what is written in a newspaper or magazine; a letter; and paperwork or forms. Respondents were also asked whether they could read aloud to a child from a children's book. Response options included *yes, easily; yes, with difficulty; and no*. Respondents were asked about which aspects of reading they found difficult, and whether they had attended any courses to improve their skills. As for age 23, these follow-up questions were excluded from analysis for the same reasons.

**Age 42 and Beyond.** Reading phenotypes are available beyond age 33 in this dataset. Our analysis did not include variables beyond

age 33 because maximal reading skill is achieved by early adulthood, with research showing that mean literacy scores in population samples are unlikely to change significantly beyond the age of 34 (Lechner et al., 2021). For completeness, we briefly describe the reading variables available. At age 42, participants were asked whether they had done any courses to improve their reading since their last NCDS interview, and, if so, how many. Age 42 reading variables also included several binary response variables, such as *Are you currently on a course to improve reading?*, *Have you ever wanted to improve your reading?*, and *Do/did you feel confident about helping your child(ren) with reading?* In addition to this, participants were asked whether they can *usually read and understand what is written in a magazine or newspaper*, whether they can *read aloud to a child from a child's storybook*, and whether they can *usually read and understand any paperwork or forms you would have to deal with in a job?* For each of these questions, the respondent had the option of responding *yes* or *no*. Each time the respondent answered *yes*, they were asked whether they could *read this easily or with difficulty*. Participants were also asked whether their *ability to read any paperwork or forms you must deal with has improved or got worse over the last 10 years?* Response options included *improved*, *got worse*, and *stayed the same*. Other reading-related variables exist at age 42; however, we chose not to include them here due to lack of specificity, with many of the variables also encompassing writing and mathematics skills. No reading variables were collected with the biomedical subsample at age 44.

Several reading variables were collected at age 46, including how many reading courses, if any, the participant had attended since their last NCDS interview, and a binary response question asking whether they would like to improve their reading skills. Participants were asked how often they read magazines or newspapers for enjoyment, and given six ordinal response options, ranging from *never* to *every day*. Respondents were asked the same question with regard to how often they read books. No reading measures were available at ages 50 or 55.

## Analysis

All phenotypic analysis was carried out in RStudio v4.1.2. Analysis was carried out separately in both the full data sample from UK Data Service ( $N = 18,558$ ), and the quality controlled subsample from the biological survey after removal of duplicated individual samples ( $n = 6431$ ), to allow for comparison. All data for which valid phenotypic values were present were used. Descriptive statistics were generated for selected variables, as described above. Correlations were generated for all continuous, ordinal and binary variables using the *hetcor* function in the *polycor* R package, which is able to determine which pairs of variables are to be correlated using Pearson, polyserial or polychoric methods based on variable type (Fox & Dusa, 2022). All complete pairs were used for analysis. It has been previously stated that correlations across time are a valid method for assessing stability of reading ability (Hulslander et al., 2010). Horn's parallel analysis of principal components was carried out to identify the most appropriate number of components for these variables (Horn, 1965; Franklin et al., 1995). Five thousand iterations were used. We opted to conduct the analysis with three PCs as recommended due to the range of reading variables, as we expected all variables to be related to a broader reading construct, but not necessarily to reading skill. Use of multiple PCs would allow us to assess which variables had high scores on the first PC to ensure that we were, in fact, capturing reading ability. Following this, PCA with oblimin rotation was conducted to

**Table 2.** Number and overall percentage of SNPs that were removed from each array due to failure to pass quality control steps

Array	Low call-rate SNPs		Failed HWE		Total	
	No.	%	No.	%	No.	%
Illumina 1.2M	11893	1.03	0	0.00	11893	1.03
Illumina 15k Custom Chip	220	2.24	0	0.00	220	2.24
Illumina Human 660-Quad	5024	0.86	680	0.12	5704	0.99
Infinium 550k v1.1	15549	2.80	850	0.15	16399	2.95
Infinium 550k v3	2291	0.41	2340	0.42	4631	0.83
Affymetrix 500	21415	4.37	2503	0.51	23918	4.88
Affymetrix v6	12638	1.35	0	0.00	12638	1.35

assess which reading variables could be combined to form general and age-specific composites (Song et al., 2013). PCA was carried out using the principal function in the *psych* R Package (Revelle, 2022), using the correlation matrices generated previously as input. All aforementioned reading variables between ages 7 to 33 were included in the correlations and therefore the PCA.

Following inspection of PCA results, variables were selected for retention in a composite at a threshold of  $>.45$  (Comrey & Lee, 1992, as cited in Finch et al., 2017, p. 1364). The variables retained in the full sample were used to calculate weighted reading composite scores for each individual. This included calculation of an overall composite, comprising of all retained measures, and a composite for use at age 7, 11 and 16. The scores calculated from the full PCA were used in order to weight the individual measures by their contribution to the reading ability component that spanned different measures and times. To achieve this, a  $z$  score was calculated for each data point, and these values were multiplied by the corresponding loading, before being summed to provide the composite (Bridges, 2022). The *polycor* package was then used to correlate the age specific childhood reading composites with adulthood reading variables, in order to assess their validity.

## Results

Tables 2 and 3 show the number and percentage of SNPs and individuals respectively that were removed from each array during the quality control process. The percentage of SNPs removed from each chip was low (maximum 4.88%), as was the percentage of individuals removed, which ranged from 0.95%–9.09%. Imputation was completed successfully for all seven arrays. The number and percentage of variants in each chip with an  $R^2$  score  $> .8$  can be found in Table 4 (Al-Soufi et al., 2021; see Supplementary Material B for full breakdown).

Basic descriptive statistics were generated for each of the key variables in both the full sample and the subsample. Breakdown of all noncontinuous variables is available in Supplementary Information C. The proportion of respondents selecting each binary and ordinal response option was similar in both the full sample and the subsample for all questions. Rates of teacher-reported difficulty in coping with reading were low in both the full sample ( $n = 203$ ) and the subsample ( $n = 40$ ; Figure S7). A similar trend was found in adulthood self-reported reading difficulties; the number of self-reported reading difficulties at age 23 was low in the full sample ( $n = 497$ ) and subsample ( $n = 169$ ). This

**Table 3.** Number and overall percentage of individuals removed from each array due to failure to pass quality control steps

Array	Missing data		Unexpected heterozygosity		Sex errors		Related individuals		Ancestry outliers		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Illumina 1.2M	74	2.54	55	1.89	1	0.10	1	0.03	6	0.21	139	4.78
Illumina 15K Custom Chip	10	0.68	10	0.68	122	8.27	1	0.07	1	0.07	74	9.09
Illumina Human 660-Quad	8	0.92	7	0.80	7	0.80	3	0.34	0	0.00	25	2.87
Infinium 550k v1.1	54	3.79	25	1.74	6	0.42	0	0.00	1	0.07	86	5.99
Infinium 550k v3	5	0.19	26	1.00	3	0.12	0	0.00	2	0.08	36	1.42
Affymetrix 500k	0	0.00	14	0.95	- <sup>a</sup>	- <sup>a</sup>	0	0.00	0	0.00	14	0.95
Affymetrix v6	61	2.05	65	2.18	1	0.03	3	0.10	7	0.23	137	4.60

Note: <sup>a</sup> No sex chromosome data was available for the Affymetrix 500k.

**Table 4.** Percentage and number of SNPs with genome-wide imputation  $R^2$  scores greater than 0.8 for each array

Array	% variants > 0.8	Number of SNPs > 0.8
Illumina 1.2M	35.33	14,276,658
Illumina 15k Custom Chip	0.21	82,702
Illumina Human 660-Quad	29.77	12,026,652
Infinium 550k v1.1	31.28	12,635,492
Infinium 550k v3	33.13	13,386,532
Affymetrix 500K	26.40	10,326,572
Affymetrix v6	31.83	12,858,081

**Table 5.** Descriptive statistics of continuous reading variables in the full sample and matched genotyped subsample

		Mean	Median	Standard deviation	N
Southgate Test	Full sample	23.34	26	7.14	14929
	Subsample	24.20	27	6.48	5821
Comprehension Test Age 11	Full sample	15.98	16	6.29	14130
	Subsample	16.71	17	5.96	5608
Comprehension Test Age 16	Full sample	25.31	27	7.09	11986
	Subsample	26.26	28	6.35	5000

was also the case at age 33 (full sample  $n = 486$ , subsample  $n = 204$ ). A breakdown of the rate of adulthood reading problems is available in Figure S10. Response rates for all variables in the full sample can be found in Supplementary Table S15. All three reading tests showed similar distributions, with slightly higher means and medians in the subsample (Table 5). A chi-square goodness-of-fit test using the Southgate Group Reading test scores showed that the difference in distribution between the full sample (mean = 23.34,  $SD = 7.14$ ) and subsample (mean = 24.20,  $SD = 6.48$ ) was significant ( $p = 6.216e^{-12}$ ,  $df = 30$ ); however, the large sample size meant that very slight differences could be detected. Correlations between all variables ranged from very weak to very strong, with clear

clusters emerging of similar variables (Figure 1). Correlations for variables assessing frequency of book reading at different ages, and English Rating showed the weakest overall correlations with the remaining variables, suggesting that these may not be appropriate for inclusion in the composite. Southgate, Teacher Rating, Reading Level, Book Use, Comprehension at both ages, English Ability, and Can Cope correlated particularly strongly with each other. All of these variables show moderate-to-strong correlations with self-reported reading difficulties at age 23 and 33, confirming the validity of these measures.

Results were similar for the PCA conducted in the full sample and the subsample (Table 6). Horn's parallel analysis indicated that a three PC solution was the most appropriate (Figure 2). Interfactor correlation coefficients were low (Table 7). While loadings differed between the full and subsample, the overall trends were consistent. The full sample results were used to generate composite scores to reduce any selection bias present in the subsample.

PC1 appears to be representative of reading skill level in the participant. In the full sample, PC1 explained 66% of the variance. Ten variables were retained (Table 6). Cronbach's alpha was calculated on the standardized variables weighted by PCA scores that were used to calculate the overall reading composite, and composites at age 7, 11 and 16 (Table 8). All composites were in the acceptable range of  $>0.7$  (Taber, 2017), except for the composite at age 16 comprised of Comprehension, Can Cope and English Ability, which fell below this value. Removal of each variable in turn showed that an acceptable Cronbach's alpha could be achieved by removing Can Cope from the data.

Following this, the three age-specific composites were correlated with each other and with Can Cope, and reading difficulties at age 23 and 33 in both samples (Figure 3). All correlations were moderate to strong, again confirming the validity of these composites and these measures over time. It should be noted that Can Cope had stronger correlations with adulthood reading difficulties (Reading Problems 23 and Reading Problems 33) than the age 16 composite did; however, the age 16 composite showed greater correlations with childhood composites. Correlation coefficients were similar in all cases, suggesting either measure may be used. However, in the subsample, only 40 participants were believed not to be able to read enough to cope, making this a very small sample size for statistical analysis. As a result, we recommend

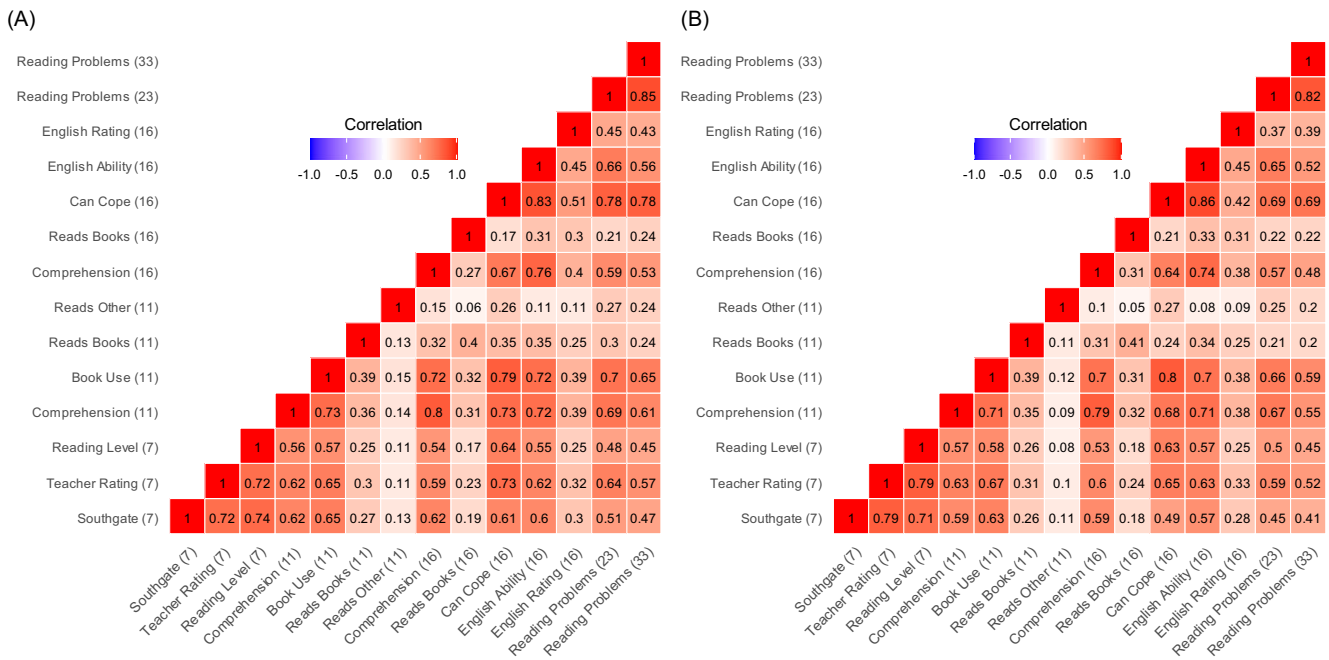


Fig. 1. Correlation Heatmap Between All Reading Variables in the Full Sample (A) and Subsample (B).

Table 6. Principal component loadings and communality from a principal components analysis

Age	Variable	PC1	PC2	PC3	h2
7	Southgate Test	.91*	-.07	-.19	.73
	Teacher Rating	.89*	-.04	-.07	.73
	Book Level	.90*	-.12	-.20	.69
11	Comprehension Test	.76*	.17	.06	.73
	Book Use	.77*	.16	.09	.76
	Reads Books	.09	.72	-.02	.57
	Reads Other	-.10	-.04	.81	.60
16	Comprehension Test	.76*	.15	.01	.68
	Reads Books	-.04	.87	-.09	.72
	Can Cope	.80*	.00	.30	.87
	English Ability	.76*	.19	.05	.73
	English Rating	.21	.39	.31	.43
23	Reading Problems	.64*	.02	.45	.80
33	Reading Problems	.58*	.01	.48	.73

Note: \*These variables passed the 0.45 threshold, and so were included in composite calculation.

Table 7. Interfactor correlations of the three principal components generated by principal components analysis of the full NCDS dataset

	PC1	PC2	PC3
PC1	1.00	.36	.30
PC2	.36	1.00	.18
PC3	.30	.18	1.00

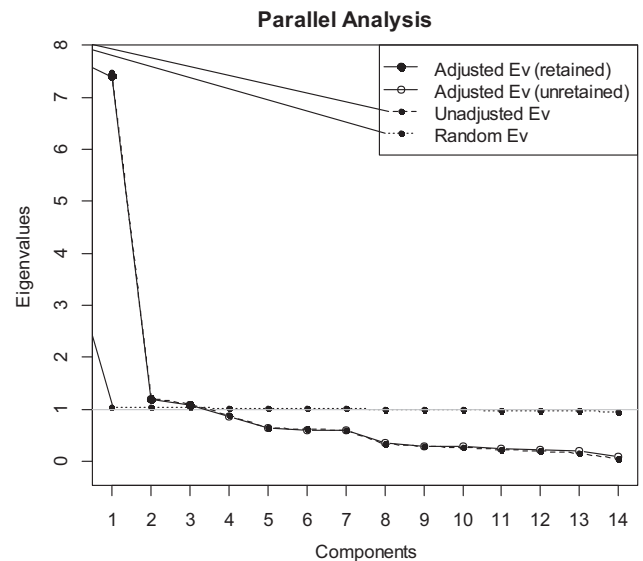


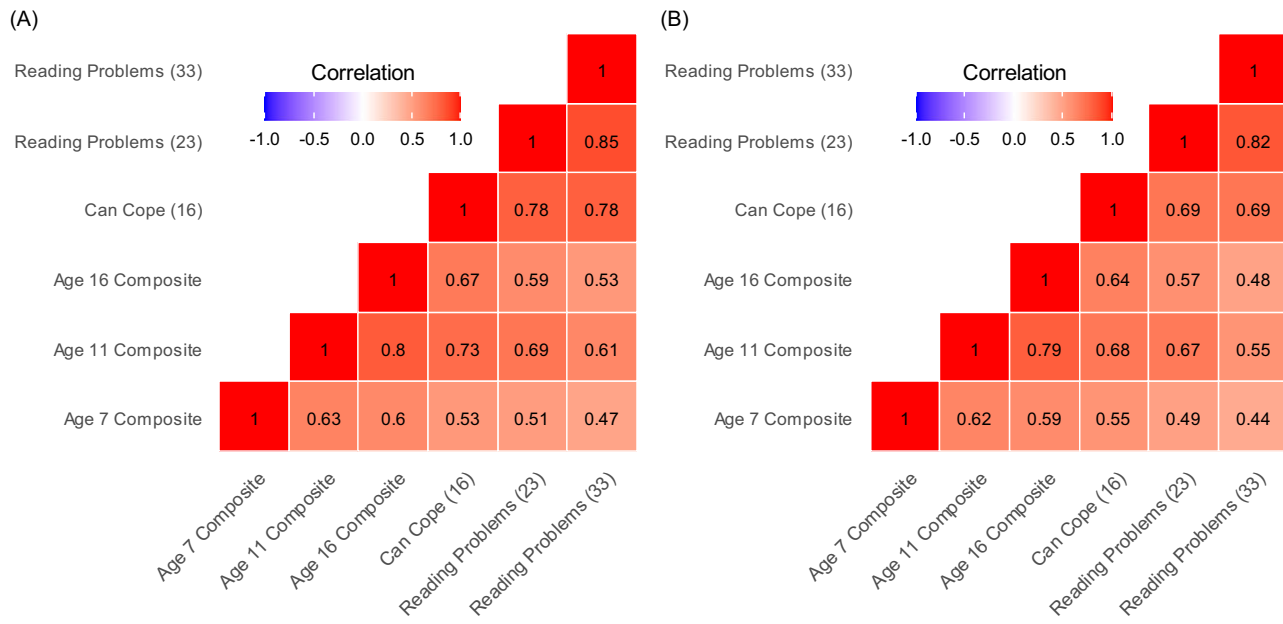
Fig. 2. Eigenvalues generated from A parallel analysis of principal components in the full NCDS sample suggesting that three components should be retained.

carefully weighing these options before choosing which one is right for a particular study.

### Discussion and Conclusion

In this study, we have presented the quality control and data preparation process of seven arrays that form a large genetic dataset, which was matched to a subsample of a UK birth cohort study who have reading skill data in childhood and adulthood. We have imputed missing genotypes for all arrays and carried out post-imputation quality control checks. We have also identified and





**Fig. 3.** Correlation heat map of age-specific reading composites and reading measures in adulthood in the full sample (A) and subsample (B). Note: Age 16 composite does not include Can Cope.

**Table 8.** Cronbach's alpha to show reliability of scale of overall and age-specific reading composites in the full sample and subsample

Composite	Cronbach's alpha
Overall	.82
Age 7	.88
Age 11	.81
Age 16 <sup>a</sup>	.82

Note: <sup>a</sup>This composite does not include Can Cope, as scale reliability was low.

described a large number of phenotypic variables in this sample, which measure reading or reading-related outcomes. Many of these variables are highly correlated, and there is substantial shared variance between them, making this dataset an excellent resource for genetically sensitive studies on reading ability, and allows for such studies to contain a longitudinal element.

Genetic data preparation and imputation were completed successfully; however, imputed data quality is variable across the seven arrays. In particular, there is a very low number of high quality variants for the Illumina 15k chip. It has been shown that a high proportion of unknown variants reduces imputation quality, and so the low number of genotyped loci may be responsible for the small proportion of high quality variants observed (Gao et al., 2011). It should be noted that most of the samples collected on the Illumina 15k chip are also present on other, higher quality chips, meaning that failure to use this chip would only result in the loss of 16 samples (see Supplementary Information D). The remaining six chips show a much higher proportion of high quality variants, which can be used to conduct further association analysis.

The strong relationships between many of the reading and reading-related variables we have investigated offer multiple options for further analysis. The high correlations between almost all variables (Figure S11) suggest that there is a high degree of reliability

between adulthood self-report and childhood teacher report and objective tests when it comes to reading ability. This has two benefits; first, it allows for data reduction when considering reading ability as a composite, avoiding multicollinearity issues and reducing noise due to measurement error; and second, it means that it is feasible to conduct longitudinal reading studies with this data up to adulthood. Additionally, the strength of correlations across the whole studied period provides further evidence that reading ability remains relatively constant as age increases (Reder, 2012). It should be noted that the reading composites we have generated are more informative of poorer reading, given that several of our variables, including the Southgate Test, Can Cope and adulthood self-report measures are intended to capture reading difficulties. While the Southgate Test results show scores from across the distribution, there is a ceiling effect.

The PCA revealed that one PC explained a substantial proportion of variance, again indicating a strong relationship between the different variables across different ages. The genotyped sample, which only included NCDS members still active in the study at age 44, has previously been reported to be less representative of the general population, including fewer poor childhood readers (Atherton et al., 2008). This likely explains the differences we found between PCA results on the full sample and subsample, and for this reason, we recommend projection of the PCA analysis from the full sample, which is more population representative, onto the subsample, rather than using subsample-specific analysis.

The 3PC solution was the most appropriate fit to the data. Upon further inspection, these PCs can be interpreted as three different reading-related constructs. The strongest loadings in PC1 are objective measures of reading ability, teacher ratings, and self-report at age 23 and 33. This suggests that these are the variables that should be included in a measure of reading ability, as they are the strongest indicators of skill. The strongest loadings in PC2 relate to how often the participant reads books at different ages, and this might reflect shared variance related to interest or enthusiasm for reading, a variable in itself that may be of interest

to reading researchers. PC3 highlights variance shared by the reporting of reading difficulties with nonbook reading.

This work has shown that the NCDS is a valuable resource for conducting genetic studies of reading ability, particularly those with a longitudinal element. The imputed arrays can be analyzed separately for GWAS meta-analysis, or alternately, the data can be combined and the array can be used as a covariate for further analysis as a single dataset. The range of reading measures available in this dataset allow for exploration of genetic influence across the full continuum of reading ability in childhood and adolescence, and for reading difficulties over time. The NCDS is currently the longest running UK cohort study for which genotyped data is available, meaning that long-term outcomes of participants, including retirement and life expectancy, can be investigated in this dataset in relation to their phenotypic reading ability and genome. Additionally, the NCDS is an incredibly rich resource in many other domains, including physical and mental health, employment, leisure activities, life in the home and financial information. This matched data provides opportunities for gene-by-environment interplay studies, including from a longitudinal perspective. In particular, this data allows the investigation of potential G × E interactions that provide a more nuanced understanding of how environmental circumstances can influence reading skill across a spectrum of abilities, and across the life course.

To conclude, the NCDS is rich potential resource for future genetic studies of reading ability. An imputed and cleaned, matched genomic dataset is available, and there is a reliable and valid set of reading variables available, as evidenced by the consistency of the different measures of reading ability across types of variable and with increasing age. Finally, the incredible detail contained in the NCDS dataset will help to make it a valuable resource for genomic studies of reading ability in the years to come.

**Supplementary material.** To view supplementary material for this article, please visit <https://doi.org/10.1017/thg.2023.2>

**Data availability.** The full phenotypic dataset is openly available to researchers at the UK Data Service. The genotyped subsample and matched phenotypic data is available from the Centre for Longitudinal Studies at University College London, upon submission of a successful research proposal.

**Acknowledgments.** With thanks to Gail Davies for advising on quality control techniques for pre-imputation genetic data, and Vicki Jackson for providing a pre-imputation protocol that informed part of this work.

**Financial support.** This work was funded by a PhD studentship from the Economic and Social Research Council with the Scottish Graduate School of Social Sciences. Data governance was provided by the METADAC data access committee, funded by ESRC, Wellcome, and MRC (2015-2018: Grant Number MR/N01104X/1 2018-2020: Grant Number ES/S008349/1). This work made use of data and samples generated by the 1958 Birth Cohort (NCDS), which is managed by the Centre for Longitudinal Studies at the UCL Institute of Education, funded by the Economic and Social Research Council (grant number ES/M001660/1). Access to these resources was enabled via the Wellcome Trust and MRC: 58FORWARDS grant [108439/Z/15/Z] (The 1958 Birth Cohort: Fostering new Opportunities for Research via Wider Access to Reliable Data and Samples). Before 2015 biomedical resources were maintained under the Wellcome Trust and Medical Research Council 58READIE Project (grant numbers WT095219MA and G1001799). HSM was supported by the Biotechnology and Biological Sciences Research Council [BB/T000813/1].

**Conflict of interest.** We have no conflict of interest to disclose.

**Ethical statement.** Ethical approval for this project was granted by METADAC.

## References

- Allegrini, A. G., Karhunen, V., Coleman, J. R. I., Selzam, S., Rimfeld, K., von Stumm, S., Pingault, J. B., & Plomin, R. (2020). Multivariable G-E interplay in the prediction of educational achievement. *PLOS Genetics*, *16*, e1009153. <https://doi.org/10.1371/journal.pgen.1009153>
- Al-Soufi, L., Martorell, L., Moltó, M., González-Peñas, J., García-Portilla, M. P., Arrojo, M., Rivero, O., Gutiérrez-Zotes, A., Nacher, J., Muntané, G., Paz, E., Páramo, M., Bobes, J., Arango, C., Sanjuan, J., Vilella, E., & Costas, J. (2021). A polygenic approach to the association between smoking and schizoprenia. *Addiction Biology*, *27*, e13104. <https://doi.org/10.1111/adb.13104>
- Astrom, R. L., Wadsworth, S. J., & DeFries, J. C. (2007). Etiology of the stability of reading difficulties: The Longitudinal Twin Study of Reading Disabilities. *Twin Research and Human Genetics*, *10*, 434–439. <https://doi.org/10.1375/twin.10.3.434>
- Atherton, K., Fuller, E., Shepherd, P., Strachan, D. P., & Power, C. (2008). Loss and representativeness in a biomedical survey at age 45 years: 1958 British birth cohort. *Journal of Epidemiology & Community Health*, *62*, 216–223. <https://doi.org/10.1136/jech.2006.058966>
- Bates, T. C., Castles, A., Coltheart, M., Gillespie, N., Wright, M., & Martin, N. G. (2004). Behaviour genetic analyses of reading and spelling: A component processes approach. *Australian Journal of Psychology*, *56*, 115–126. <https://doi.org/10.1080/00049530410001734847>
- Bates, T. C., Castles, A., Luciano, M., Wright, M. J., Coltheart, M., & Martin, N. G. (2007). Genetic and environmental bases of reading and spelling: A unified genetic dual route model. *Reading and Writing*, *20*, 147–171. <https://doi.org/10.1007/s11145-006-9022-1>
- Bejemann, R. S., Willcutt, E. G., Olson, R. K., Keenan, J. M., DeFries, J. C., & Wadsworth, S. J. (2007). Word reading and reading comprehension: Stability, overlap and independence. *Reading and Writing*, *21*, 539–558. <https://doi.org/10.1007/s11145-007-9076-8>
- Bridges, E. C. (2022). NCDS.imputation. GitHub. <https://github.com/ecbridges/NCDS.imputation>
- Byrne, B., Wadsworth, S., Corley, R., Samuelsson, S., Quain, P., DeFries, J. C., Willcutt, E., & Olson, R. K. (2005). Longitudinal twin study of early literacy development: Preschool and kindergarten phases. *Scientific Studies of Reading*, *9*, 219–235. [https://doi.org/10.1207/s1532799xssr0903\\_3](https://doi.org/10.1207/s1532799xssr0903_3)
- Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology*, *104*, 166–181. <https://doi.org/10.1037/a0025323>
- Centre for Longitudinal Studies. (n.d.-a). *Genetic data and biological samples*. Centre for Longitudinal Studies UCL. <https://cls.ucl.ac.uk/data-access-training/genetic-data-and-biological-samples/>
- Centre for Longitudinal Studies. (n.d.-b). *NCDS Age 55 Sweep*. Centre for Longitudinal Studies UCL. <https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study/ncds-age-55-sweep/>
- Centre for Longitudinal Studies. (n.d.-c). *NCDS Age 62 Sweep*. Centre for Longitudinal Studies UCL. <https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study/ncds-age-62-sweep/>
- Currie, J., & Thomas, D. (1999). *Early test scores, socioeconomic status and future outcomes (Working Paper 6943)*. National Bureau of Economic Research. [https://www.nber.org/system/files/working\\_papers/w6943/w6943.pdf](https://www.nber.org/system/files/working_papers/w6943/w6943.pdf)
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*, giab008. <https://doi.org/10.1093/gigascience/giab008>
- Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., & Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*, 1284–1287. <https://doi.org/10.1038/ng.3656>
- Davies, N. M., Hemani, G., Timpson, N. J., Windmeijer, F., & Davey Smith, G. (2015). The role of common genetic variation in educational attainment and income: Evidence from the National Child Development Study. *Scientific Reports*, *5*, 16509. <https://doi.org/10.1038/srep16509>

- Doust, C., Fontanillas, P., Eising, E., Gordon, S. D., Wang, Z., Alagöz, G., Molz, B., Aslibekyan, S., Auton, A., Babalola, E., Bell, R. K., Bielenberg, J., Bryc, K., Bullis, E., Coker, D., Partida, G. C., Dhamija, D., Das, S., Elson, S. L., & Luciano, M. (2022). Discovery of 42 genome-wide significant loci associated with dyslexia. *Nature Genetics*, *54*, 1621–1629. <https://doi.org/10.1038/s41588-022-01192-y>
- Ebejer, J. L., Coventry, W. L., Byrne, B., Willcutt, E. G., Olson, R. K., Corley, R., & Samuelsson, S. (2010). Genetic and environmental influences on inattention, hyperactivity-impulsivity, and reading: Kindergarten to Grade 2. *Scientific Studies of Reading*, *14*, 293–316. <https://doi.org/10.1080/1088430903150642>
- Eising, E., Mirza-Schreiber, N., de Zeeuw, E. L., Wang, C. A., Truong, D. T., Allegrini, A. G., Shapland, C. Y., Zhu, G., Wigg, K. G., Gerritse, M. L., Molz, B., Alagöz, G., Gialluisi, A., Abbondanza, F., Rimfeld, K., van Donkelaar, M., Liao, Z., Jansen, P. R., Andlauer, T. F. M., Fisher, S. E. (2022). Genome-wide analyses of individual differences in quantitatively assessed reading- and language-related skills in up to 34,000 people. *Proceedings of the National Academy of Sciences*, *119*, e2202764119. <https://doi.org/10.1073/pnas.2202764119>
- Erbeli, F., Hart, S. A., & Taylor, J. (2017). Longitudinal associations among reading-related skills and reading comprehension: A twin study. *Child Development*, *89*, e480–e493. <https://doi.org/10.1111/cdev.12853>
- Fernald, A., Marchman, V. A., & Weisleder, A. (2012). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*, 234–248. <https://doi.org/10.1111/desc.12019>
- Finch, A. P., Brazier, J. E., Mukuria, C., & Bjorner, J. B. (2017). An exploratory study on using principal-component analysis and confirmatory factor analysis to identify bolt-on dimensions: The EQ-5D case study. *Value in Health*, *20*, 1362–1375. <https://doi.org/10.1016/j.jval.2017.06.002>
- Fox, J., & Dusa, A. (2022). *polycor* (0.8-1) [R Package]. <https://cran.r-project.org/web/packages/polycor/polycor.pdf>
- Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: A method for determining significant principal components. *Journal of Vegetation Science*, *6*, 99–106. <https://doi.org/10.2307/3236261>
- Fung, W. K., & Chung, K. K. H. (2019). The role of socioeconomic status in Chinese word reading and writing among Chinese kindergarten children. *Reading and Writing*, *33*, 377–397. <https://doi.org/10.1007/s11145-019-09967-2>
- Gao, G., Limdi, N. A., Liu, N., Zhang, B., Zhang, K., & Zhi, D. (2011). Practical consideration of genotype imputation: Sample size, window size, reference choice, and untyped rate. *Statistics and Its Interface*, *4*, 339–351. <https://doi.org/10.4310/sii.2011.v4.n3.a8>
- Hamilton, L. G., Haiyou-Thomas, M. E., Hulme, C., & Snowling, M. J. (2016). The home literacy environment as a predictor of the early literacy development of children at family-risk of dyslexia. *Scientific Studies of Reading*, *20*, 401–419. <https://doi.org/10.1080/1088438.2016.1213266>
- Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*, 1279–1283. <https://doi.org/10.1038/ng.3643>
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. <https://doi.org/10.1007/bf02289447>
- Hulme, C., & Snowling, M. J. (2016). Reading disorders and dyslexia. *Current Opinion in Pediatrics*, *28*(6), 731–735. <https://doi.org/10.1097/mop.0000000000000411>
- Hulslander, J., Olson, R. K., Willcutt, E. G., & Wadsworth, S. J. (2010). Longitudinal stability of reading-related skills and their prediction of reading development. *Scientific Studies of Reading*, *14*, 111–136. <https://doi.org/10.1080/1088431003604058>
- Kendler, K. S., & Baker, J. H. (2006). Genetic influences on measures of the environment: A systematic review. *Psychological Medicine*, *37*, 615. <https://doi.org/10.1017/s0033291706009524>
- Lechner, C. M., Gauly, B., Miyamoto, A., & Wicht, A. (2021). Stability and change in adults' literacy and numeracy skills: Evidence from two large-scale panel studies. *Personality and Individual Differences*, *180*, 110990. <https://doi.org/10.1016/j.paid.2021.110990>
- Leppänen, P. H., Hämäläinen, J. A., Salminen, H. K., Eklund, K. M., Guttorm, T. K., Lohvansuu, K., Puolakano, A., & Lyytinen, H. (2010). Newborn brain event-related potentials revealing atypical processing of sound frequency and the subsequent association with later literacy skills in children with familial dyslexia. *Cortex*, *46*, 1362–1376. <https://doi.org/10.1016/j.cortex.2010.06.003>
- Logan, J. A. R., Hart, S. A., Cutting, L., Deater-Deckard, K., Schatschneider, C., & Petrill, S. (2013). Reading development in young children: Genetic and environmental influences. *Child Development*, *84*, 2131–2144. <https://doi.org/10.1111/cdev.12104>
- Lohvansuu, K., Torppa, M., Ahonen, T., Eklund, K., Hämäläinen, J. A., Leppänen, P. H. T., & Lyytinen, H. (2021). Unveiling the mysteries of dyslexia — Lessons learned from the prospective Jyväskylä Longitudinal Study of Dyslexia. *Brain Sciences*, *11*, 427. <https://doi.org/10.3390/brainsci11040427>
- Luciano, M. (2017). Making reading easier: How genetic information can help. *Policy Insights from the Behavioral and Brain Sciences*, *4*, 147–154. <https://doi.org/10.1177/2372732217720464>
- Luciano, M., & Bates, T. (2019). Mapping genes involved in reading and language skills. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 533–556). The MIT Press.
- Mamedova, S., & Pawlowski, E. (2019). *Data point: Adult literacy in the United States* (NCES 2019–179). National Center for Education Statistics. <https://nces.ed.gov/pubs2019/2019179.pdf>
- Melrose, K. (2014, August). *Annex A: Encouraging participation and persistence in adult literacy and numeracy*. Behavioural Insights Team. <https://www.publicinformationonline.com/download/201799>
- Meyer, H. (2021a, July 15). *Ancestry estimation based on reference samples of known ethnicities*. plinkQC. <https://meyer-lab-cshl.github.io/plinkQC/articles/AncestryCheck.html>
- Meyer, H. (2021b, July 15). *Processing 1000 Genomes reference data for ancestry estimation*. plinkQC. <https://meyer-lab-cshl.github.io/plinkQC/articles/Genomes1000.html>
- Molfese, V. J., Modglin, A., & Molfese, D. L. (2003). The role of environment in the development of reading skills. *Journal of Learning Disabilities*, *36*, 59–67. <https://doi.org/10.1177/00222194030360010701>
- National Literacy Trust. (ca. 2017). *Adult literacy*. <https://literacytrust.org.uk/parents-and-families/adult-literacy/>
- Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N., & Ioannidis, J. P. (2013). The power of meta-analysis in genome-wide association studies. *Annual Review of Genomics and Human Genetics*, *14*, 441–465. <https://doi.org/10.1146/annurev-genom-091212-153520>
- Petrill, S. A., Hart, S. A., Harlaar, N., Logan, J., Justice, L. M., Schatschneider, C., Thompson, L., DeThorne, L. S., Deater-Deckard, K., & Cutting, L. (2010). Genetic and environmental influences on the growth of early reading skills. *Journal of Child Psychology and Psychiatry*, *51*, 660–667. <https://doi.org/10.1111/j.1469-7610.2009.02204.x>
- Price, K. M., Wigg, K. G., Feng, Y., Blokland, K., Wilkinson, M., He, G., Kerr, E. N., Carter, T., Guger, S. L., Lovett, M. W., Strug, L. J., & Barr, C. L. (2020). Genome-wide association study of word reading: Overlap with risk genes for neurodevelopmental disorders. *Genes, Brain and Behavior*, *19*, e12648. <https://doi.org/10.1111/gbb.12648>
- Power, C., & Elliott, J. (2005). Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*, *35*, 34–41. <https://doi.org/10.1093/ije/dyi183>
- Psyridou, M., Tolvanen, A., Lerkkanen, M. K., Poikkeus, A. M., & Torppa, M. (2020). Longitudinal stability of reading difficulties: Examining the effects of measurement error, Cut-Offs, and buffer zones in identification. *Frontiers in Psychology*, *10*. <https://doi.org/10.3389/fpsyg.2019.02841>
- Puglisi, M. L., Hulme, C., Hamilton, L. G., & Snowling, M. J. (2017). The home literacy environment is a correlate, but perhaps not a cause, of variations in children's language and literacy development. *Scientific Studies of Reading*, *21*, 498–514. <https://doi.org/10.1080/1088438.2017.1346660>
- Rayner, W. (2016). *ic, a post-imputation data checking program*. McCarthy Group Tools. <https://www.well.ox.ac.uk/%7Ewrayner/tools/Post-Imputation.html>
- Rayner, W. (2020). *McCarthy Group Tools*. McCarthy Group Tools. <https://www.well.ox.ac.uk/%7Ewrayner/tools/index.html#Checking>

- Reder, S. (2012). *The longitudinal study of adult learning: Challenging assumptions*. The Centre for Literacy. [https://tcal.org.au/wp-content/uploads/2021/01/Longitudinal-Study-CFLRsrchBrief\\_Chllngng\\_Assmptns.pdf](https://tcal.org.au/wp-content/uploads/2021/01/Longitudinal-Study-CFLRsrchBrief_Chllngng_Assmptns.pdf)
- Revelle, W. (2022). psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 2.2.5. <https://CRAN.R-project.org/package=psych>
- Robertson, N. (2012, January 17). *Update build*. Genotyping Chips Strand and Build Files. [https://www.well.ox.ac.uk/~wrayner/strand/update\\_build.sh](https://www.well.ox.ac.uk/~wrayner/strand/update_build.sh)
- Russell, G., Ukoumunne, O. C., Ryder, D., Golding, J., & Norwich, B. (2016). Predictors of word-reading ability in 7-year-olds: Analysis of data from a U.K. cohort study. *Journal of Research in Reading*, *41*, 58–78. <https://doi.org/10.1111/1467-9817.12087>
- Sabatini, J. P., Shore, J., Holtzman, S., & Scarborough, H. S. (2011). Relative effectiveness of reading intervention programs for adults with low literacy. *Journal of Research on Educational Effectiveness*, *4*, 118–133. <https://doi.org/10.1080/19345747.2011.555290>
- Samuelsson, S., Byrne, B., Olson, R. K., Hulstlander, J., Wadsworth, S., Corley, R., Willcutt, E. G., & DeFries, J. C. (2008). Response to early literacy instruction in the United States, Australia, and Scandinavia: A behavioral-genetic analysis. *Learning and Individual Differences*, *18*, 289–295. <https://doi.org/10.1016/j.lindif.2008.03.004>
- Selzam, S., Dale, P. S., Wagner, R. K., DeFries, J. C., Cederlöf, M., O'Reilly, P. F., Krapohl, E., & Plomin, R. (2017). Genome-wide polygenic scores predict reading performance throughout the school years. *Scientific Studies of Reading*, *21*, 334–349. <https://doi.org/10.1080/10888438.2017.1299152>
- Senechal, M. (2006). Testing the home literacy model: Parent involvement in kindergarten is differentially related to Grade 4 reading comprehension, fluency, spelling, and reading for pleasure. *Scientific Studies of Reading*, *10*, 59–87. [https://doi.org/10.1207/s1532799xssr1001\\_4](https://doi.org/10.1207/s1532799xssr1001_4)
- Shepherd, P. (2012, December). *1958 National Child Development Study user guide: Measures of ability at ages 7 to 16*. Centre for Longitudinal Studies. <https://cls.ucl.ac.uk/wp-content/uploads/2017/07/NCDS-user-guide-NCDS1-3-Measures-of-ability-P-Shepherd-December-2012.pdf>
- Snowling, M. J., & Hulme, C. (2011). Interventions for children's language and literacy difficulties. *International Journal of Language & Communication Disorders*, *47*, 27–34. <https://doi.org/10.1111/j.1460-6984.2011.00081.x>
- Song, M. K., Lin, F. C., Ward, S. E., & Fine, J. P. (2013). Composite variables. *Nursing Research*, *62*, 45–49. <https://doi.org/10.1097/nnr.0b013e3182741948>
- Sultmann, W. F., Elkins, J., Miller, S. F., & Byrne, M. (1983). Linguistic awareness and early reading. *Reading Psychology*, *4*, 327–335. <https://doi.org/10.1080/0270271830040315>
- Taber, K. S. (2017). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*, 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74. <https://doi.org/10.1038/nature15393>
- Tizard, J., Schofield, W. N., & Hewison, J. (2002). Collaboration between teachers and parents in assisting children's reading. In G. Reid, J. Soler, & J. Wearmouth (Eds.), *Addressing difficulties in literacy development: Responses at family, school, pupil, and teacher levels* (pp. 39–57). Routledge.
- Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I. J., Li, R., Ling, H., Manolio, T. A., Matsumoto, M., McCarty, C. A., McDavid, A. N., Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*, *1*, 1–19. <https://doi.org/10.1002/0471142905.hg0119s68>
- UK Data Service. (n.d.). *National Child Development Study*. <https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000032>
- Wadsworth, S. J., Corley, R. P., Hewitt, J. K., & DeFries, J. C. (2001). Stability of genetic and environmental influences on reading performance at 7, 12, and 16 years of age in the Colorado Adoption Project. *Behavior Genetics*, *31*, 353–359. <https://doi.org/10.1023/a:1012218301437>
- Wadsworth, S. J., DeFries, J. C., Olson, R. K., & Willcutt, E. G. (2007). Colorado longitudinal twin study of reading disability. *Annals of Dyslexia*, *57*, 139–160. <https://doi.org/10.1007/s11881-007-0009-7>
- Wadsworth, S. J., DeFries, J. C., Willcutt, E. G., Pennington, B. F., & Olson, R. K. (2015). The Colorado Longitudinal Twin Study of reading difficulties and ADHD: Etiologies of comorbidity and stability. *Twin Research and Human Genetics*, *18*, 755–761. <https://doi.org/10.1017/thg.2015.66>
- Wadsworth, S. J., DeFries, J. C., Willcutt, E. G., Pennington, B. F., & Olson, R. K. (2016). Genetic etiologies of comorbidity and stability for reading difficulties and ADHD: A replication study. *Twin Research and Human Genetics*, *19*(6), 647–651. <https://doi.org/10.1017/thg.2016.80>
- Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*, 661–678. <https://doi.org/10.1038/nature05911>
- Williams, S. M., & Silva, P. A. (1985). Some factors associated with reading ability: a longitudinal study. *Educational Research*, *27*, 159–168. <https://doi.org/10.1080/0013188850270301>
- World Literacy Foundation. (2018, March). *The economic & social cost of illiteracy: A white paper by the World Literacy Foundation*. <https://worldliteracyfoundation.org/wp-content/uploads/2021/07/TheEconomicSocialCostofilliteracy-2.pdf>
- Zhan, X., & Liu, D. (2016). *checkVCF: Sanity check Variant Call Format (VCF) files*. GitHub. <https://github.com/zhanxw/checkVCF>