Psychiatry Research 354 (2025) 116826

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Original article

Combining genetic and behavioral predictors of 11-year language outcome

Loretta Gasparini [a,b,*], Daisy A. Shepherd [a,b], Katherine Lange [a,b], Jing Wang [a,b], Ellen Verhoef [c], Edith L. Bavin [d,a], Sheena Reilly [a,b,e], Beate St. Pourcain [c,f,g], Melissa Wake [a,b,h,1], Angela T. Morgan [a,i,j,1]

[a] Murdoch Children's Research Institute, Royal Children's Hospital, 50 Flemington Rd, Parkville 3052, VIC, Australia
[b] Department of Paediatrics, The University of Melbourne, Parkville 3052, VIC, Australia
[c] Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[d] School of Psychology and Public Health, La Trobe University, Bundoora, Australia
[e] Health Group, Griffith University, Gold Coast, Australia
[f] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
[g] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands
[h] Liggins Institute, The University of Auckland, Grafton, Auckland 1023, New Zealand
[i] Department of Audiology and Speech Pathology, The University of Melbourne, Parkville 3052, VIC, Australia
[j] Royal Children's Hospital Melbourne, 50 Flemington Rd, Parkville 3052, VIC, Australia

ARTICLE INFO

Keywords:
Language development
Language disorders
Genetics
Polygenic score
Sensitivity and specificity
Longitudinal studies
Machine learning

ABSTRACT

Background: Rapid population-level identification of language disorders could help provide care to young children to improve their outcomes. Two previous studies identified and replicated up to six parent-reported items that predicted 11-year language outcome with ≥71 % sensitivity and specificity. Here, we assess whether including genetic propensity for toddlerhood vocabulary improves predictive accuracy.
Method: The Early Language in Victoria Study (ELVS) recruited 1910 8-month-olds in Melbourne in 2003–2004. The Longitudinal Study of Australian Children (LSAC) recruited 5107 0–1-year-olds across Australia in 2004. Both collected parent-reported items at 2–3 years, a comparable 11-year language outcome: the Clinical Evaluation of Language Fundamentals (CELF-4) Core Language score or Recalling Sentences subtest, and biospecimens for genotyping. We derived polygenic scores capturing participants' genetic propensity for parent-reported 24–38-month vocabulary. We calculated univariate associations with continuous language outcomes. We used ensemble method SuperLearner to estimate how accurately the parent-reported predictors and polygenic scores predict low 11-year language outcome (>1.5 standard deviations below the mean) in each cohort.
Results: Language outcome was available for 839 ELVS and 1441 LSAC participants. Polygenic scores accounted for little variance in continuous language outcomes (R^2 < 1.3 %). Adding polygenic scores to the predictor sets increased accuracy of predicting language outcome by up to 7 %, but inconsistently between analyses.
Conclusions: Polygenic scores derived for toddlerhood vocabulary did not meaningfully improve predictive accuracy of individuals' language outcome when added to the phenotypic predictor set. Presently, parent-reported measures or clinician observation appear best for predicting language outcome at this age.

Abbreviations: ALSPAC, Avon Longitudinal Study of Parents and Children; AUC, Area Under the receiving operating characteristic Curve; CDI–III, MacArthur–Bates Communicative Development Inventories, Third Edition; CELF–4, Clinical Evaluation of Language Fundamentals 4th Edition, Australian Standardized Edition; CI, confidence interval; DNA, deoxyribonucleic acid; EAGLE, Early Genetics and Life Course Epidemiology; ELVS, Early Language in Victoria Study; GWAS, genome–wide association study; LSAC, Growing Up in Australia: The Longitudinal Study of Australian Children; OSF, Open Science Framework; PEDS, Parents' Evaluation of Developmental Status; PGS, polygenic score; PGS–ex, polygenic score derived for expressive vocabulary using all available SNPs after clumping; PGS–ex–bayes, polygenic score derived for expressive vocabulary using a Bayesian–based approach to SNP filtering; PGS–rec, polygenic score derived for receptive vocabulary using all available SNPs after clumping; PGS–rec–bayes, polygenic score derived for receptive vocabulary using a Bayesian–based approach to SNP filtering; Pheno, set of phenotypic predictors only; SD, standard deviation; SNP, single nucleotide polymorphism; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology.

* Corresponding author.
E-mail address: loretta.gasparini@mcri.edu.au (L. Gasparini).
[1] Joint senior authors

https://doi.org/10.1016/j.psychres.2025.116826
Received 1 September 2025; Received in revised form 23 October 2025; Accepted 6 November 2025
Available online 7 November 2025

1. Introduction

About 7–10 % of children have a lifelong language disorder, defined as language difficulties impacting everyday functioning (Bishop et al., 2017; Calder et al., 2022; Norbury et al., 2016). Language disorders can impact socio-behavioral, academic, employment and quality of life outcomes (Dubois et al., 2020; Van Barreveld et al., 2025; Ziegenfusz et al., 2022) and are unlikely to resolve, although intervention can help improve functioning (Bishop et al., 2017). Language disorders are largely unknown to the community and under-researched (McGregor, 2020).

We don't yet have rapid, accurate, population-level ways to identify language disorders in the preschool years (Feltner et al., 2024). This hinders providing early support to young children with language disorders, leaving them underdiagnosed and critically underserved (McGregor, 2020). A previous study investigated all the individual parent-reported survey items (approximately 2000) collected in the Early Language in Victoria Study (ELVS) from 8–36 months of age, spanning the family, socioeconomics, early communication, language and other development. This identified a set of six parent-reported survey items administered at 2–3 years that predicted 11-year language with 75 % sensitivity and 81 % specificity in ELVS (Gasparini et al., 2023). In a replication using a national-based cohort (Longitudinal Study of Australian Children; LSAC), these same predictors yielded 78 % sensitivity and 71 % specificity for predicting 11–12-year language (Gasparini et al., 2024b). This set of six predictors with replicated results has similar or better sensitivity and specificity for predicting late childhood language outcome compared to other studies that have looked at the accuracy of early predictors (e.g. Armstrong et al., 2018; and for a review see Feltner et al., 2024). Although 70–80 % sensitivity and specificity has been cited as potentially acceptable for developmental screening (Council on Children With Disabilities et al., 2006; Wallace et al., 2015), it still misclassifies many children. If used for screening, this would prevent children misclassified as not having a language disorder from receiving the support they need. It would waste time and money and cause undue stress for families whose child is misclassified as having a language disorder. Improving predictive accuracy would bring such tools closer to suitability for population-wide screening (Feltner et al., 2024).

Current approaches to identifying preschool children likely to have persisting language difficulties are limited to observable traits of the child or their environment (e.g. Armstrong et al., 2018; Borovsky et al., 2021; Feltner et al., 2024; Law et al., 2023; McKean et al., 2016). However, there is evidence that both genetic and environmental factors influence language and related skills (Mountford et al., 2022). The human genome is present at birth and consistent throughout life, so if accurate measures representing genetic propensity for later language outcome become available, they could be collected at birth or in early life to aid early identification (Plomin and Von Stumm, 2022).

Previous studies examining genetic measures of language propensity are scarce and have yielded mixed results. In one study, genetic propensity for language disorder using genome-wide association studies (GWAS) from a cohort of 700 individuals explained 3.8 % of genetic variance for having a language disorder (Schiavon et al., 2024). In another, genetic propensity for 8-year expressive vocabulary derived using GWAS from a cohort of 2718 children and a constrained set of genes was found to account for only 0.18 % of variance in the phenotype for which it was derived (8-year expressive vocabulary) (Newbury et al., 2019). In that same study, genetic propensity scores for 15-month vocabulary accounted for 0.11 % variance of that phenotype, and for all other genetic propensity scores, the variance explained of the phenotypes for which they were derived was <0.1 %. The approach for deriving genetic measures can greatly affect the effect size (Pain et al., 2021). Genetic propensity scores for toddlerhood vocabulary calculated using larger GWAS samples than Newbury et al. (2019) and Schiavon et al. (2024) have been found to be associated with later phenotypes including listening comprehension, literacy, intelligence, and Attention-Deficit/Hyperactivity Disorder (Verhoef et al., 2020, 2023). Extending these findings, our study is, to our knowledge, the first to examine the relations between genome-wide genetic propensity scores based on a large GWAS sample ($n > 6000$) and a late-childhood global language phenotype that spans receptive and productive vocabulary and grammar. Furthermore, our study uniquely investigates whether genetic propensity scores can help identify children with language difficulties at the individual level. Combining genetic propensity scores with behavioral predictors may improve classification models beyond relying on only one approach, as has been found in risk profiling for various psychiatric and neurodevelopmental conditions (Bourque et al., 2025; Dybdahl Krebs et al., 2024; Murray et al., 2021), but not yet for predicting later language outcomes.

1.1. Aim

In the current study, we derived polygenic scores (PGSs), reflecting part of an individual's genetic propensity for toddlerhood expressive and receptive vocabulary. We added these PGSs to an established, replicated set of parent-reported language predictors (Gasparini et al., 2023; Gasparini et al., 2024b). We assessed whether this strategy improves accuracy for predicting individuals' 11-year language outcome.

2. Methods

We preregistered the methods in a protocol and reported protocol amendments. These and all Supporting information files and statistical analysis code are available on Open Science Framework (OSF) at https://osf.io/mrxdg/?view_only=90878858c980441d8e0190a1ce23cc54 under '*Files*'. For ELVS and LSAC, ethical approval was granted, guardians provided written, informed consent, and permission was granted to use the data for the current study.

We follow the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines (von Elm et al., 2007). See Supporting information 1 for additional methodological details and Supporting information 2 for explanations of key terms used in this paper.

2.1. Study design and participants

In this study, we used two cohorts, ELVS (Reilly et al., 2018) and LSAC (Clifford et al., 2019; Edwards, 2014; Sanson and Johnstone, 2004), to assess the predictive accuracy of our genetic and parent-reported predictors. We refer to these as the "analysis datasets". To derive the genetic predictors, we drew data from a meta-GWAS across multiple cohorts of children (Verhoef et al., 2023), which we refer to as the "GWAS dataset". In this study, the GWAS and analysis datasets are independent from each other, so no individual person participated in more than one. We describe the analysis datasets below, and describe the GWAS dataset and procedures for deriving the genetic predictors in 2.2.2.

Both ELVS and LSAC are Australian prospective longitudinal population-based cohorts that recruited infants in 2003–2004 (at recruitment, ELVS: $n = 1910$; LSAC: $n = 5107$) and collected data on children's families, environment and development at least biennially from infancy until adolescence (at 11 years, ELVS had $n = 851$; at 11–12 years LSAC had $n = 1874$ participants).

ELVS recruited in six metropolitan areas of Melbourne, Victoria. Participants were excluded from joining ELVS if at recruitment (8 months of age) the Maternal and Child Health Nurse identified them as having a disability likely to interfere with their language development. This included developmental delay, cerebral palsy, a syndrome, intellectual or physical disability, cleft lip or palate, vision or hearing impairment. Children were also excluded if their parents did not understand English sufficiently to complete questionnaires designed for grade 6 reading level.

LSAC recruited across Australia using a clustered design and children were excluded from recruitment if they lived in some remote parts of Australia, were not enrolled in Medicare or had no fixed address at the time of sampling. See Supporting information 1 for further cohort details.

2.2. Variables

2.2.1. Parent-reported predictors

ELVS and LSAC participants' parents completed surveys relating to their child, the family and home environment. The current study uses six parent survey items (all administered in English; listed in Table 1) from ELVS at ages 24 and 36 months, and from LSAC at 2–3 years. We used this set of predictors in this study as it has been found to have ≥71 % sensitivity and specificity for predicting 11-year language outcome, replicated across ELVS and LSAC (Gasparini et al., 2023; Gasparini et al., 2024b). Its predictive accuracy is comparable to any of the other predictor sets investigated in those two studies, which spanned thousands of predictors across various domains of child development and their environment (always maximum 70–85 % sensitivity and specificity) (Gasparini et al., 2023; Gasparini et al., 2024b). However, in the current primary analysis we excluded the variable about child behavior as it had 57 % missing data amongst LSAC children with an available outcome, while the other predictors all had <20 % missingness (Gasparini et al., 2024b). Excluding this variable, the remaining five predictors still had satisfactory predictive accuracy (Area Under the receiving operating characteristic Curve, AUC>71 %, see Supporting information 4).

2.2.2. Genetic predictors

Participants provided biospecimens for genotyping at 7 years in ELVS (Richards et al., 1993) and at 11–12 years in LSAC. Samples with 300 ng DNA (deoxyribonucleic acid) were genotyped. We imputed further genotypes and cleaned the data, excluding single nucleotide polymorphisms (SNPs, genetic variants at a single position in the genetic code) and participants with low-quality data based on our preregistered criteria. See Supporting information 1 for further details on genotyping, imputation, data cleaning, quality checks and PGS calculation.

GWAS summary statistics estimate the association of each SNP with a phenotype of interest. In this case, we obtained GWAS summary statistics from the EAGLE (Early Genetics and Life Course Epidemiology) Consortium (Verhoef et al., 2023). The analyses from the EAGLE Consortium comprised seven cohorts (Avon Longitudinal Study Parents and Children [ALSPAC], Barwon Infant Study, Copenhagen Prospective Studies on Asthma in Childhood, Generation R, Raine Study, Twins Early Development Study, and LSAC) amounting to 17,298 children of European descent. Within each cohort participating in the meta-GWAS, expressive vocabulary size between 24–38 months of age was estimated using age- and language-specific, parent-reported word lists adapted from the MacArthur-Bates Communicative Development Inventories (CDI, Fenson et al., 1993) or the Language Development Survey (Rescorla, 1989). Additionally, at 38 months in ALSPAC, receptive vocabulary size was estimated using the CDI. Genetic samples were also collected from participants from each cohort. Verhoef et al. (2023) then meta-analyzed seven cohort-level GWAS summary statistics (reflecting associations of each SNP with 24–38-month vocabulary size) to yield meta-GWAS summary statistics for toddlerhood expressive (n = 16,615) and receptive (n = 6291) vocabulary.

Verhoef and colleagues' (2023) meta-GWAS for expressive vocabulary originally included the LSAC cohort. So, for the current study, we reran their meta-GWAS for expressive vocabulary excluding LSAC, to create independence between the GWAS dataset and our analysis datasets (ELVS and LSAC). Thus, our PGS for expressive vocabulary was derived from GWAS summary statistics for expressive vocabulary size from the remaining six EAGLE cohorts (n = 15,481). Verhoef and colleagues' (2023) GWAS summary statistics for receptive vocabulary were derived using participants from ALSPAC only (n = 6291), thus they were already independent from our analysis datasets, and we used these summary statistics in the current study.

For each ELVS and LSAC participant (from whom a genotyped sample was available and who met our quality control criteria), each of their genetic variants (SNPs that met our quality control criteria) were weighted by their relative effect size obtained from the toddlerhood vocabulary GWAS summary statistics (Verhoef et al., 2023). We then summed these weighted values across all matched SNPs to compute a PGS for expressive vocabulary at 24–38 months (denoted 'PGS-ex') and a PGS for receptive vocabulary at 38 months ('PGS-rec') for each ELVS and LSAC individual (Qi et al., 2012). We inspected the PGS distributions for normality and transformed each PGS to internal z-scores, where a higher PGS indicates greater genetic propensity for larger vocabulary size at 24–38 months.

2.2.3. Outcome variable

At 11 years the ELVS participants completed the Clinical Evaluation of Language Fundamentals, Fourth Edition, Australian Standardized Edition (CELF-4) in English (Pearson Education, 2008; Semel et al., 2006). In this study, the ELVS 11-year language outcome was the CELF-4 Core Language score (Semel et al., 2006). The CELF-4 is a standardized tool used widely in clinical practice that yields a Core Language score with a mean of 100 and standard deviation (*SD*) of 15, where a higher score corresponds to better performance.

At 11–12 years, the LSAC participants completed the Recalling Sentences subtest of the CELF-4 in English (Semel et al., 2006), which is the LSAC 11-year language outcome in this study. Participants heard audio recordings of sentences via audio file on an iPad which they were asked to repeat verbatim (Wang et al., 2018). This differs from the ELVS outcome measure because LSAC did not administer the whole CELF-4, only this subtest. Based on previous studies showing sentence repetition to strongly agree with global language skills (Archibald and Joanisse, 2009; Botting et al., 2001; M. Klem et al., 2015; Redmond et al., 2019; Smith et al., 2019), we consider this subtest a suitable language outcome measure for our purposes, but we discuss the limitations of potentially misclassifying some LSAC children in 4.1. The Recalling Sentences subtest age-related scaled scores range from 1–18 with a mean of 10 and *SD* of 3.

Language skills exist on a continuum, with no clear cut-off distinguishing typical language and language disorders (Bishop et al., 2016). Thus, we examined univariate associations between continuous

Table 1
Six questions asked at 24 and 36 months that yielded 75 % sensitivity and 81 % specificity for predicting 11-year low language ability in the ELVS cohort and 78 % sensitivity and 71 % specificity in the LSAC cohort.

Question	Options	Source
Mark the sentence that sounds most like the way your child talks at the moment. If your child is saying sentences even longer or more complicated than the two provided, mark the second one.	This dolly big This dolly big and this dolly little	CDI-III
"Circle"	Says Not yet	CDI-III
"Accident"	Says Not yet	CDI-III
"Kangaroo"	Says Not yet	CDI-III
"Forget/forgot"	Says Not yet	CDI-III
Do you have any concerns about how your child behaves?*	No Yes A little	PEDS

CDI-III: MacArthur-Bates Communicative Development Inventories, Third Edition (Dale, 2007); PEDS: Parents' Evaluation of Developmental Status (Glascoe, 1999).

* Excluded from our primary analysis due to high data missingness. Included in a sensitivity analysis.

polygenic scores and continuous CELF-4 Core Language or Recalling Sentences scores. However, as we are interested in identifying children likelier to have later language difficulties, we used a dichotomous language outcome when evaluating predictive performance. The CELF-4 Core Language score has been found to have 100 % sensitivity and 89 % specificity for diagnosing Language Disorder at a cut-off of 1.5*SD* below the population mean (Pearson Education, 2008). Thus, in our primary analysis of predictive performance, we dichotomized the outcome as "low" language at this cut-off of >1.5*SD* below the population mean (Core Language score <77.5 or Recalling Sentences scaled score <5.5), or as "typical" language (Core Language score ≥77.5 or Recalling Sentences scaled score ≥5.5). In sensitivity analyses we used 1.*25SD* and 2*SD* cut-offs (see 2.3.1.2). We consider a low CELF-4 Core Language or Recalling Sentences score as a potential indicator, but not a diagnosis, of Language Disorder.

2.3. Statistical methods

We conducted statistical analyses in RStudio (R Core Team, 2020; RStudio Team, 2020), with self-developed R code provided on the OSF repository.

2.3.1. Primary analysis

2.3.1.1. Univariate associations. Separately for ELVS and LSAC, we calculated Pearson's correlations with 95 % confidence intervals (CIs) between (a) PGS-ex and PGS-rec, and (b) each of PGS-ex and PGS-rec and the continuous language outcomes. We squared the Pearson's correlations values to obtain R^2 estimates, which represent the percentage of variance one variable accounts for in the other in a univariate association.

2.3.1.2. Predictive performance. We used SuperLearner (Polley et al., 2021; van der Laan et al., 2007), an ensemble approach that uses various machine learning algorithms to estimate the predictive accuracy (AUC) of the PGSs, the parent-reported predictor sets, and combinations thereof. We applied 10-fold cross-validation and included various individual algorithms in our SuperLearner algorithm (see Supporting information 1 for details). We ran models using complete case data (individuals with no missing variables included in the given model) (Dashti et al., 2021) and report for each model the total number of participants with complete case data included.

For ELVS and LSAC separately, we ran SuperLearner with the following predictor sets: (i) five parent-reported predictors (denoted as 'Pheno'); (ii) PGS-ex, (iii), PGS-rec, (iv) PGS-ex & PGS-rec, (v) Pheno & PGS-ex; (vi) Pheno & PGS-rec; (vii) Pheno & PGS-ex & PGS-rec. For these seven sets of predictors, we calculated the AUC of the predictor sets with corresponding 95 % CIs. We compared the AUC of the different predictor sets, to understand how much the predictive accuracy of the set of phenotypic predictors improves when adding a genetic predictor, namely PGSs for toddlerhood vocabulary size, replicated across two cohorts.

Additionally, we combined the ELVS and LSAC cohorts and ran SuperLearner with the five parent-reported predictors ((i) Pheno, as above) to improve precision (we could not combine the PGSs between cohorts due to technical barriers, see Supporting information 1 for an explanation).

When inspecting predictor set accuracy, we consider AUC, sensitivity or specificity of <70 % as "unsatisfactory", 70–79 % as "fair", 80–89 % as "good" and greater than 90 % as "excellent". We consider the 95 % confidence intervals (CIs) to reflect the precision and level of uncertainty of our results.

2.3.2. Additional analyses

We ran various sensitivity analyses to support our primary analyses. See the full methods and results in Supporting information 1 and 4 respectively. Briefly, (1) we meta-analyzed the correlation coefficients of the PGSs and language outcome. (2) We included in all models only the participants for whom the five phenotypic predictors, the PGSs and a language outcome were all available (rather than all participants with complete cases for each given model). (3) We included in the models only participants of European genetic ancestry. (4) We derived the PGSs using the same method as Verhoef et al. (2023) using a Bayesian-based approach developed by Ge et al. (2019). (5) We used alternative language outcome cut-offs of 1.25 and 2*SD*, instead of 1.5*SD*, below the mean. (6) We used the CELF-4 Recalling Sentences subtest as the outcome in ELVS. (7) We added the PGSs to the six-item predictor set reported in two previous studies that we excluded in our main analysis as it resulted in greater data loss. Finally, (8) we estimated the predictive accuracy of a predictor set replacing the vocabulary item 'kangaroo' with 'today', to create an alternative that may be more suitable for international users. We report in the Results any results that do not accord with the primary results.

3. Results

3.1. Participant characteristics

A language outcome was available for 839 (43.9 %) ELVS and 1441 (28.2 %) LSAC participants recruited at baseline. In both cohorts, the group of participants for whom an outcome was available lived in more socioeconomically advantaged neighborhoods, had parents with higher education levels, had fewer Aboriginal and Torres Strait Islander families, and fewer families from non-English-speaking backgrounds than those for whom an outcome was not available. In Supporting information 3 we report participant characteristics and summarize the missingness levels and proportions or means and *SDs* of all the variables used in our analyses.

3.2. Primary analysis

3.2.1. Univariate associations

Linear associations between PGS-ex, PGS-rec and the 11-year language outcome were similar in the two cohorts (Fig. 1). There was a modest overlap in variance between the expressive and receptive PGSs (14–16 % variance; Table 2). There was very small linear association between either PGS and language outcome in both cohorts ($r < 0.114$, corresponding to the PGSs accounting for <1.3 % variance in continuous language outcomes). The ELVS 95 % CIs suggest some uncertainty (the upper 95 % CI value suggests up to 3.8 % variance accounted for by PGS-rec), but the LSAC 95 % CI values all indicate very small associations (upper 95 % CI values suggesting maximum 0.9 % variance accounted for by PGS-ex).

3.2.2. Predictive performance

SuperLearner prediction model results are presented in Table 3. The five parent-reported phenotypic predictors (Pheno) resulted in an estimated AUC of 72–86 %, similar to results in the studies that developed this predictor set (Gasparini et al., 2023; Gasparini et al., 2024b). When adding the PGS derived from receptive vocabulary to the five phenotypic predictors (Pheno & PGS-rec), the AUC increased by about 1 % in ELVS and 4 % in LSAC, compared to Pheno (see Fig. 2). In both cohorts, this predictor set (Pheno & PGS-rec) yielded the highest predictive accuracy (AUC). However, 95 % CIs largely overlapped with those obtained for the phenotypic predictors alone. Adding PGS-ex instead of PGS-rec, or both, to the phenotypic predictors yielded similar results, suggesting only marginal improvement by adding genetic propensity scores for toddlerhood vocabulary. Both PGS-ex and PGS-rec had low predictive accuracy when studied without phenotypic predictors, with AUC values close to chance level at 50 %.

Next, we combined ELVS and LSAC data and calculated the

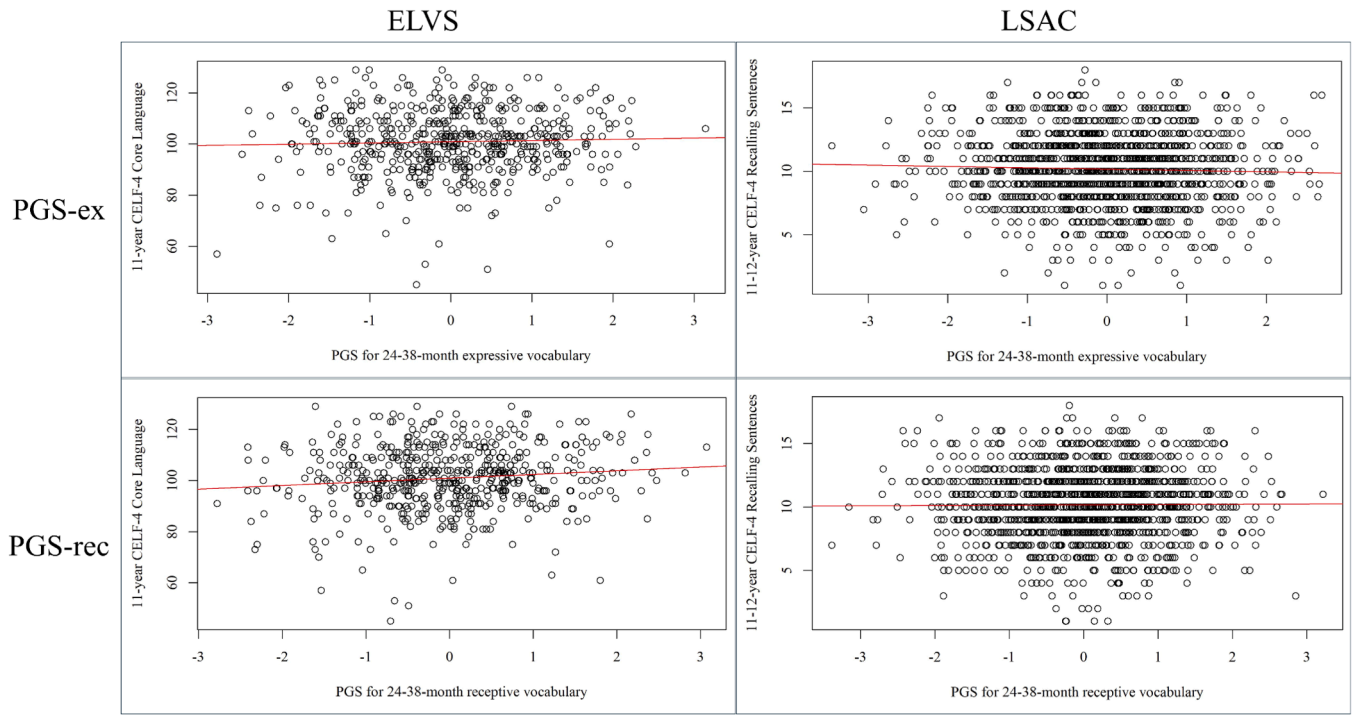

Fig. 1. Scatterplots of the relationships between z-scaled PGSs reflecting genetic propensity for 24–38-month expressive (PGS-ex, top) and receptive (PGS-rec, bottom) vocabulary (x-axis) and 11-year language outcomes (y-axis). On the left are the results for ELVS and on the right LSAC. Regression lines are included.

Table 2
Pairwise correlation coefficients (r) and % variance accounted for ($R^2 \times 100$) with 95 % CIs between PGSs reflecting genetic propensity for expressive vocabulary (PGS-ex), PGSs for receptive vocabulary (PGS-rec), and 11-year language outcome (ELVS: CELF-4 Core Language score, LSAC: CELF-4 Recalling Sentences scaled score), assuming a linear association.

	ELVS		LSAC	
	r (95 % CI)	$R^2 \times 100$ (95 % CI)	*r* (95 % CI)	$R^2 \times 100$ (95 % CI)
PGS-ex ~ PGS-rec	0.385 (0.324, 0.444)	14.840 (10.470, 19.686)	0.396 (0.354, 0.436)	15.668 (12.521, 19.029)
11-year-language ~ PGS-ex	0.038 (−0.046, 0.122)	0.146 (0.000, 1.480)	−0.038 (−0.093, 0.017)	0.146 (0.000, 0.872)
11-year-language ~ PGS-rec	0.114 (0.030, 0.196)	1.293 (0.090, 3.831)	0.008 (−0.047, 0.064)	0.007 (0.000, 0.407)

predictive accuracy of the five parent-reported phenotypic predictors (Table 4). As expected, this shows predictive accuracy values in between those of ELVS and LSAC when examined alone, with increased precision as reflected by the slightly narrower 95 % CIs (compare with 'Pheno' in Table 3). However, the sensitivity estimates remain rather imprecise, with 95 % CIs spanning a 20 % range, reflecting the low prevalence of low language outcome.

3.3. Additional analyses

Our sensitivity analyses methods and results are described in Supporting information 1 and 4 respectively. Meta-analyzing the pairwise correlation coefficients yielded similar point estimates but wider 95 % CIs than the single cohorts (11-year-language~PGS-ex: R^2=0.005 % [0.000 %, 20.259 %]; 11-year-language~PGS-rec: R^2=0.317 % [0.000 %, 38.440 %]) suggesting uncertainty in these estimates. In our sensitivity analyses where we used an alternative subsample of participants, method for deriving PGSs, language outcome, or phenotypic predictor set, adding PGSs to the phenotypic predictor set sometimes improved predictive accuracy by up to 7 % (see Supporting information 4 Table S4.12), but sometimes not at all (see Supporting information 4 Tables S4.9, S4.11), and never consistently between different analyses, PGS type (expressive or receptive), or cohorts (ELVS or LSAC). In all cases across our primary and sensitivity analyses, when comparing the predictive accuracy of phenotypic predictors only and the addition of a PGS, 95 % CI values overlapped substantially.

4. Discussion

Genetic propensity scores (PGSs) derived for toddlerhood vocabulary had very small linear relationships with 11-year language outcomes in both ELVS and LSAC, accounting for a maximum of 1.3 % variance. The amount of variance explained is up to almost 10 times larger than a previous study of PGSs derived for various childhood language phenotypes using only a constrained set of genes, who found at most 0.18 % variance explained by their polygenic scores (Newbury et al., 2019). However, we note that our 95 % CIs include values as low as 0.00 % variance, indicating low precision in our point estimates.

Previous studies have suggested that genetic propensity for toddlerhood vocabulary may lay a foundation for later literacy and cognitive skills (e.g., genetic correlation of 0.79 with spelling outcomes in Verhoef et al., 2023). Our results provide limited evidence for such genetic links with late childhood global language performance, although this may reflect low statistical power.

Given that PGSs capture only part of (genetic) contributions to a trait, it is expected that our PGSs alone could not predict language outcome at the individual level (Bourque et al., 2025; Murray et al., 2021). But neither did our PGSs improve the prediction of individuals' language outcomes when combined with parent-reported predictors, suggesting they capture little unique information. Adding PGSs to the phenotypic predictor set improved predictive accuracy (AUC) by up to 7 % in some cases, but sometimes not at all. The 95 % CIs always overlapped, and findings were not consistent between cohorts, PGS type

Table 3
Area under the curve (AUC) of SuperLearner models with different combinations of predictors for each cohort ELVS (left) and LSAC (right). 'Typical' and 'Low' show the number of participants included in each group after removing missing data. Bold indicates the models with the highest AUC for each cohort.

	ELVS				LSAC			
	n		Model*	AUC (95 % CI)	n		Model	AUC (95 % CI)
	Typical	Low			Typical	Low		
Five phenotypic predictors (Pheno)	750	25	SuperLearner	0.861 (0.783, 0.939)	1134	48	SuperLearner	0.729 (0.645, 0.812)
			Discrete	0.845 (0.759, 0.930)			Discrete	0.742 (0.660, 0.823)
Expressive vocab PGS (PGS-ex)	525	21	SuperLearner	0.486 (0.338, 0.634)	1188	68	SuperLearner	0.447 (0.376, 0.517)
			Discrete	0.431 (0.266, 0.595)			Discrete	0.461 (0.389, 0.534)
Receptive vocab PGS (PGS-rec)	525	21	SuperLearner	0.560 (0.409, 0.711)	1188	68	SuperLearner	0.419 (0.346, 0.492)
			Discrete	0.448 (0.284, 0.611)			Discrete	0.425 (0.351, 0.499)
PGS-ex & PGS-rec	525	21	SuperLearner	0.554 (0.397, 0.710)	1188	68	SuperLearner	0.453 (0.381, 0.525)
			Discrete	0.612 (0.471, 0.754)			Discrete	0.442 (0.374, 0.509)
Pheno & PGS-ex	499	16	SuperLearner	0.862 (0.770, 0.954)	994	41	SuperLearner	0.759 (0.683, 0.835)
			Discrete	0.798 (0.684, 0.912)			Discrete	0.770 (0.696, 0.844)
Pheno & PGS-rec	499	16	SuperLearner	0.851 (0.747, 0.955)	994	41	**SuperLearner**	**0.785 (0.720, 0.850)**
			Discrete	**0.873 (0.798, 0.947)**			Discrete	0.781 (0.710, 0.851)
Pheno & PGS-ex & PGS-rec	499	16	SuperLearner	0.851 (0.748, 0.954)	994	41	SuperLearner	0.780 (0.713, 0.847)
			Discrete	0.834 (0.720, 0.947)			Discrete	0.752 (0.674, 0.830)

* 'SuperLearner' is ensemble-based using multiple algorithms (listed in Supporting information 1). 'Discrete' is the single best performing algorithm. We select the best performing of the two for a given predictor set.

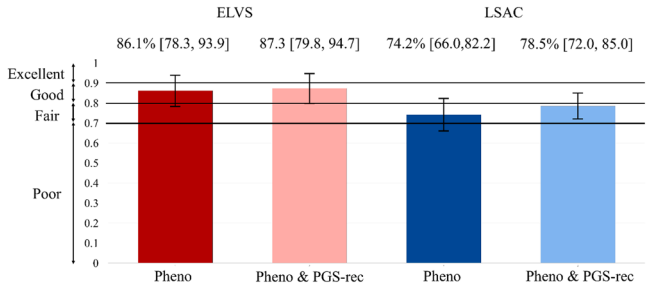

Fig. 2. AUC [95 % CIs] of (a) the five parent-reported phenotypic predictors only, (Pheno, dark) and (b) that set of phenotypic predictors plus the PGS derived for receptive vocabulary PGS-rec (Pheno & PGS-rec, light) in ELVS (red, left) and LSAC (blue, right).

Table 4
AUC, sensitivity and specificity of SuperLearner models with the set of five parent-reported predictors, combining ELVS and LSAC. 'Typical' and 'Low' show the number of participants included in each group after removing missing data.

	n		Model*	AUC (95 % CI)	Sensitivity (95 % CI)	Specificity (95 % CI)
	Typical	Low				
Pheno	1884	73	SuperLearner	0.78 (0.71, 0.84)	0.73 (0.61, 0.82)	0.76 (0.74, 0.78)
			Discrete	0.80 (0.74, 0.85)	0.74 (0.62, 0.84)	0.74 (0.72, 0.76)

* 'SuperLearner' is ensemble-based using multiple algorithms (listed in Supporting information 1). 'Discrete' is the single best performing algorithm. We select the best performing of the two for a given predictor set.

(expressive or receptive vocabulary), or sensitivity analyses, limiting our confidence in any improvement afforded by the PGSs.

This small and inconsistent relationship between the PGSs and 11-year language outcome may be because expressive and receptive vocabulary fluctuate greatly throughout toddler- and childhood (Reilly et al., 2014). A GWAS derived for alternative language phenotypes may be more effective at capturing relevant genetic effects. Some methodological explanations for our findings may be due to our approach to calculating PGSs. Although we did follow the approach that yielded the best results in a previous study examining body mass index (Lange et al., 2022). And in a sensitivity analysis, we used the approach used by the EAGLE Consortium (Ge et al., 2019; Verhoef et al., 2023), which was found to be on average one of the best methods for deriving PGSs in a methodological study of different derivation approaches and various outcomes (Pain et al., 2021). Additionally, restricting SNPs to those that could be imputed in both cohorts may have resulted in more consistent results across cohorts, albeit the smaller set of SNPs would have resulted in lower precision, so we opted against this. Our current sample size may have also been underpowered to detect a substantial contribution of the PGSs, as shown by wide 95 % CIs in the meta-analysis of correlation coefficients.

4.1. *Strengths and limitations*

Our study has many strengths. The GWAS from which we derived our genetic measures is the largest to date that examined genetic associations with language-related phenotypes (PGS-ex: $n = 15{,}481$, PGS-rec: $n = 6291$ participants) (Verhoef et al., 2023) and we exploited two different PGS approaches that yielded similar results. Both ELVS and LSAC recruited across whole communities with broad inclusion criteria. Repeating our results across these two cohorts allowed us to observe whether our results replicated in two Australian population-based cohort studies. Collecting a standardized language outcome in late childhood allowed us to identify children with persisting language difficulties (a characteristic of Language Disorder) (McKean et al., 2017). Our ELVS language outcome is a widely used measure for diagnosing Language Disorder, giving us confidence that the ELVS children were classified appropriately as having low or typical language at 11 years (see next paragraph for a discussion on limitations of the LSAC outcome). Using the ensemble method SuperLearner minimizes the effects of arbitrary decisions by running multiple algorithms and parameters, it is able to flexibly observe non-linear relationships and interactions between variables, and it uses cross-validation to reduce overfitting (Lever et al., 2016).

Regarding limitations, our complete case analyses likely introduced selection bias, as the groups of ELVS and LSAC participants from whom an outcome was available are more socioeconomically advantaged, have fewer Aboriginal and Torres Strait Islander families and families from non-English-speaking backgrounds than their counterparts for whom an outcome was not available. A limitation of our language outcome was that LSAC only administered the CELF-4 Recalling Sentences subtest, which has been found to have good to excellent, but not perfect, agreement with global language across several studies (Archibald and Joanisse, 2009; Botting et al., 2001; M. Klem et al., 2015; Redmond

et al., 2019; Smith et al., 2019). By extrapolating the level of agreement between CELF-4 Core Language Score and Recalling Sentences in ELVS, we estimate we may have misclassified about a quarter of LSAC children we classified as having low language outcome, or about 2 % of all LSAC children (see Supporting information 4, Section 6). While this may somewhat lower one's confidence in our LSAC results, it is worth noting that our interpretations are consistent between both ELVS, who used a robust outcome, and LSAC. That is: across both cohorts, the parent-reported predictor set has fair predictive accuracy, and adding PGS does not consistently improve predictive accuracy.

4.2. Future directions

The currently available PGSs derived for parent-reported toddlerhood vocabulary did not consistently correlate with late-childhood language outcomes across two Australian cohorts. To estimate individuals' genetic propensity for persisting language skills, future research could derive PGSs for direct language outcome measures collected at later ages, namely from 4 years to around 11 years and at multiple timepoints, to better characterize children's persisting language skills (McKean et al., 2017). These outcome measures should be ones that have been validated as representing overall language skills (e.g. CELF Core Language score, or sentence repetition which has been found to be a good proxy: Archibald and Joanisse, 2009; Botting et al., 2001; M. Klem et al., 2015; Redmond et al., 2019; Smith et al., 2019). While Newbury et al. (2019) did include a variety of later language phenotypes and did not find any large genetic effects, this may be due to their smaller GWAS sample and the fact that they only included a constrained set of genes to derive polygenic scores (they did this to increase power, due to their small sample). Meanwhile Schiavon et al. (2024) estimated a 3.8 % genetic effect of language disorder using GWAS with a later language phenotype, albeit their GWAS sample size was also small ($n = 700$). Thus, deriving GWAS summary statistics and PGSs each from much larger, as well as more diverse, samples may yield more precise and generalizable results. A future study could also systematically compare different PGS calculation methods (e.g. as per Lange et al., 2022; Pain et al., 2021) to determine the optimal methods for deriving language-related PGSs.

4.3. Practical implications

Adding a PGS for toddlerhood receptive vocabulary to the phenotypic predictors resulted in the highest AUC in both cohorts in our primary analysis. However, as the improvement was only marginal and inconsistent, the current findings do not justify collecting biospecimens to individually predict language outcomes. For now, parent-reported measures or assessment of language skills by a trained examiner appear most promising for predicting in the preschool years later language outcome (Feltner et al., 2024).

Our five-item phenotypic predictor set could be suitable for recruiting 2–3-year-old participants into research studies at large scale where more accurate but lengthier instruments (e.g. Feltner et al., 2024) are not feasible. As such, we have made code available for researchers to input the results of parents' surveys to yield estimated probabilities of participants having persisting language difficulties (Gasparini et al., 2024a). The predictor set in its current form is inappropriate for clinical screening for language disorder due to the proportion of children it would misclassify (about 25 %). Instead, clinicians may use longer, more accurate screening tools or a longitudinal surveillance approach, following best practice (Bishop et al., 2016; Eadie et al., 2022; Feltner et al., 2024).

5. Conclusion

We have reported a set of five parent-reported variables that can be asked when children are aged 2–3 years with "fair" accuracy (74 % sensitivity and specificity) for predicting which children will have low language skills at 11–12 years. Currently available PGSs derived for toddlerhood vocabulary did not meaningfully improve predictive accuracy of individuals' language outcome when added to the phenotypic predictor set. This reflects the fact that PGSs only account for part of an individual's genetic propensity for a trait, and genetically complex phenotypes like language likely involve a strong interplay between genetics and environment. Different approaches to phenotyping and deriving genetic measures, as well as larger samples may be needed to develop genetic predictors of language for individual-level prediction. For now, parent report or assessment by a trained examiner, perhaps using a longitudinal surveillance approach (Eadie et al., 2022), appear most promising for predicting later language outcome in the preschool years.

Funding/Support

LG was supported by an Australian Government Research Training Program (RTP) Scholarship and MCRI PhD Top Up Scholarship. DS was supported by the Lorenzo and Pamela Galli Medical Trust. JW was supported by the Jack Brockhoff Foundation Early Career Medical Research Grant and MCRI Generation Victoria (GenV) postdoc fellowship. MW was supported by a Senior Research Fellowships (1046,518). AM was supported by an NHMRC Investigator grant (1195,955). ELVS was supported by the National Health and Medical Research Council (NHMRC) of Australia (grant numbers 237,106, 436,958, 1041,947) and small grants from the Murdoch Children's Research Institute (MCRI) and La Trobe University. EV and BSTP were funded by the Max Planck Society. The Child Health CheckPoint has been supported to date by the National Health and Medical Research Council (NHMRC) of Australia (Project Grants 1041,352, 1109,355), The Royal Children's Hospital Foundation (2014–241), MCRI, The University of Melbourne, National Heart Foundation of Australia (100,660), Financial Markets Foundation for Children (2014–055) and Victorian Deaf Education Institute. Research at the MCRI was supported by the Victorian Government's Operational Infrastructure Support Program.

Role of funders

DSS played a role in the study design of LSAC; however, no other funding bodies had a role in the study design and conduct; data collection, management, analysis, and interpretation; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The MCRI administered the research grants for the study and provided infrastructural support (IT and biospecimen management) to its staff and the study but played no role in the conduct or analysis of the study.

Ethics statement

LSAC has been approved by The Royal Children's Hospital Melbourne Human Research Ethics Committee (33225D) and The Australian Institute of Family Studies Ethics Committee (14–26). ELVS has been approved by the Royal Children's Hospital (23018, 27078, 94330) and La Trobe University Human Ethics Committee (03–32). A parent/guardian provided written informed consent for their child's participation.

Data sharing statement: ELVS data can be made accessible upon submitting a proposal to the Investigator team (Tricia Eadie: peadie@unimelb.edu.au) and receiving Ethics approval. LSAC data can be made accessible upon submitting an application to the Australian Data Archive (https://ada.edu.au/).

CRediT authorship contribution statement

Loretta Gasparini: Writing – review & editing, Writing – original

draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Daisy A. Shepherd:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Katherine Lange:** Writing – review & editing, Software, Methodology, Formal analysis. **Jing Wang:** Writing – review & editing, Methodology. **Ellen Verhoef:** Writing – review & editing, Resources, Methodology. **Edith L. Bavin:** Writing – review & editing, Resources, Methodology, Investigation. **Sheena Reilly:** Writing – review & editing, Resources, Methodology, Investigation. **Beate St. Pourcain:** Writing – review & editing, Resources, Methodology. **Melissa Wake:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization. **Angela T. Morgan:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

LSAC was conducted in partnership between the Department of Social Services (DSS), the Australian Institute of Family Studies (AIFS), and the Australian Bureau of Statistics (ABS). The iPad Apps to administer the Recalling Sentences subtest in LSAC were developed with support from the Centre of Research Excellence in Child Language (Australian National Health & Medical Research Council grant 1023493). We thank the ELVS, LSAC, CheckPoint, ALSPAC, BIS, COPSAC, GenR, Raine Study, and TEDS study participants, caregivers, staff and students for their contributions. We thank the Victorian Maternal and Child Health nurses who supported ELVS recruitment, Tricia Eadie, Bibi Gerner, Diana Hermith-Ramirez, and other ELVS researchers, and Vicki Jackson for providing the ELVS genotype data. For insightful contributions, thank you to Richard Saffery, Marta Schiavon, and attendees of the XVI$_{th}$ International Congress for the Study of Child Language 2024.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2025.116826.

References

Archibald, L.M.D., Joanisse, M.F., 2009. On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. J. Speech. Lang. Hear. Res. 52 (4), 899–914. https://doi.org/10.1044/1092-4388(2009/08-0099).

Armstrong, R., Symons, M., Scott, J.G., Arnott, W.L., Copland, D.A., McMahon, K.L., Whitehouse, A.J.O, 2018. Predicting language difficulties in middle childhood from early developmental milestones: a comparison of traditional regression and machine learning techniques. J. Speech. Lang. Hear. Res. 61 (8), 1926–1944. https://doi.org/10.1044/2018_JSLHR-L-17-0210.

Bishop, D.V.M., Snowling, M.J., Thompson, P.A., Greenhalgh, T., the CATALISE-2 consortium, 2017. Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: terminology. J. Child Psychol. Psychiatry 58 (10), 1068–1080. https://doi.org/10.1111/jcpp.12721.

Bishop, D.V.M., Snowling, M.J., Thompson, P.A., Greenhalgh, T., CATALISE consortium, 2016. CATALISE: a multinational and multidisciplinary Delphi Consensus study. Identifying language impairments in children. PLoS. One 11 (7), e0158753. https://doi.org/10.1371/journal.pone.0158753.

Borovsky, A., Thal, D., Leonard, L.B., 2021. Moving towards accurate and early prediction of language delay with network science and machine learning approaches. Sci. Rep. 11 (1), 8136. https://doi.org/10.1038/s41598-021-85982-0.

Botting, N., Conti-Ramsden, G., Faragher, B., 2001. Psycholinguistic markers for specific language impairment (SLI). J. Child Psychol. Psychiatry Allied Discip. 42 (6), 741–748. https://doi.org/10.1017/S0021963001007600. Cambridge Core.

Bourque, V.-R., Schmilovich, Z., Huguet, G., England, J., Okewole, A., Poulain, C., Renne, T., Jean-Louis, M., Saci, Z., Zhang, X., Rolland, T., Labbé, A., Vorstman, J., Rouleau, G.A., Baron-Cohen, S., Mottron, L., Bethlehem, R.A.I., Warrier, V., Jacquemont, S., 2025. Genomic and developmental models to predict cognitive and adaptive outcomes in autistic children. JAMa Pediatr. 179 (6), 655. https://doi.org/10.1001/jamapediatrics.2025.0205.

Calder, S.D., Brennan-Jones, C.G., Robinson, M., Whitehouse, A., Hill, E., 2022. The prevalence of and potential risk factors for developmental language disorder at 10 years in the raine study. J. Paediatr. Child Health. https://doi.org/10.1111/jpc.16149 jpc.16149.

Clifford, S.A., Davies, S., Wake, M., 2019. Child Health CheckPoint: cohort summary and methodology of a physical health and biospecimen module for the Longitudinal Study of Australian Children. BMJ Open. 9 (Suppl 3), 3–22. https://doi.org/10.1136/bmjopen-2017-020261.

Council on Children With Disabilities, 2006. Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. Pediatrics 118 (1), 405–420. https://doi.org/10.1542/peds.2006-1231.

Dale, P.S., 2007. MacArthur-Bates Communicative Development Inventory—III, 3rd edition. Paul H. Brookes Publishing Company.

Dashti, S.G., Lee, K.J., Simpson, J.A., White, I.R., Carlin, J.B., Moreno-Betancur, M., 2021. Handling Missing Data When Estimating Causal Effects with Targeted Maximum Likelihood Estimation. https://doi.org/10.48550/arXiv.2112.05274 arXiv.

Dubois, P., St-Pierre, M.-C., Desmarais, C., Guay, F., 2020. Young adults with developmental language disorder: a systematic review of education, employment, and independent living outcomes. J. Speech. Lang. Hear. Res. 63 (11), 3786–3800. https://doi.org/10.1044/2020_JSLHR-20-00127.

Dybdahl Krebs, M., Georgii Hellberg, K.-L., Lundberg, M., Appadurai, V., Ohlsson, H., Pedersen, E., Steinbach, J., Matthews, J., Border, R., LaBianca, S., Calle, X., Meijsen, J.J., Ingason, A., Buil, A., Vilhjálmsson, B.J., Flint, J., Bacanu, S.-A., Cai, N., Dahl, A., Schork, A.J., 2024. Genetic liability estimated from large-scale family data improves genetic prediction, risk score profiling, and gene mapping for major depression. Am. J. Hum. Genet. 111 (11), 2494–2509. https://doi.org/10.1016/j.ajhg.2024.09.009.

Eadie, P., Levickis, P., McKean, C., Westrupp, E., Bavin, E.L., Ware, R.S., Gerner, B., Reilly, S., 2022. Developing preschool language surveillance models—cumulative and clustering patterns of early life factors in the early language in Victoria study cohort. Front. Pediatr. 10, 826817. https://doi.org/10.3389/fped.2022.826817.

Edwards, B., 2014. Growing up in Australia: the Longitudinal Study of Australian Children: entering adolescence and becoming a young adult. Fam. Matters. 95, 5–14.

Feltner, C., Wallace, I.F., Nowell, S.W., Orr, C.J., Raffa, B., Middleton, J.C., Vaughan, J., Baker, C., Chou, R., Kahwati, L., 2024. Screening for speech and language delay and disorders in children 5 years or younger: evidence report and systematic review for the US preventive services task force. JAMa 331 (4), 335. https://doi.org/10.1001/jama.2023.24647.

Fenson, L., Dale, P.S., Reznick, S., Thal, D., Bates, E., Hartung, J., Pethick, S., Reilly, J., 1993. MacArthur Communicative Development Inventories: User's Guide and Manual. Singular Publishing Group.

Gasparini, L., Shepherd, D.A., Bavin, E.L., Eadie, P., Reilly, S., Morgan, A.T., Wake, M., 2023. Using machine-learning methods to identify early-life predictors of 11-year language outcome. J. Child Psychol. Psychiatry 64 (8), 1242–1252. https://doi.org/10.1111/jcpp.13733.

Gasparini, L., Shepherd, D.A., Wake, M., Morgan, A.T., 2024a. Languagepredictions. GitHub. https://github.com/lottiegasp/languagepredictions.

Gasparini, L., Shepherd, D.A., Wang, J., Wake, M., Morgan, A.T., 2024b. Identifying early language predictors: a replication of Gasparini et al. (2023) confirming applicability in a general population cohort. Int. J. Lang. Commun. Disord. 59, 2352–2366. https://doi.org/10.1111/1460-6984.13086.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., Smoller, J.W., 2019. Polygenic prediction via bayesian regression and continuous shrinkage priors. Nat. Commun. 10 (1), 1776. https://doi.org/10.1038/s41467-019-09718-5.

Glascoe, F.P., 1999. Using parents' Concerns to detect and address developmental and behavioral problems. J. Spec. Pediatr. Nurs. 4 (1), 24–35. https://doi.org/10.1111/j.1744-6155.1999.tb00077.x.

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S.H., Gustafsson, J., Hulme, C., 2015. Sentence repetition is a measure of children's language skills rather than working memory limitations. Dev. Sci. 18 (1), 146–154. https://doi.org/10.1111/desc.12202.

Lange, K., Kerr, J.A., Mansell, T., O'Sullivan, J.M., Burgner, D.P., Clifford, S.A., Olds, T., Dwyer, T., Wake, M., Saffery, R., 2022. Can adult polygenic scores improve prediction of body mass index in childhood? Int. J. Obes. 46 (7), 1375–1383. https://doi.org/10.1038/s41366-022-01130-2.

Law, J., Charlton, J., Wilson, P., Rush, R., Gilroy, V., McKean, C., 2023. The development and productivity of a measure for identifying low language abilities in children aged 24–36 months. BMC. Pediatr. 23 (1), 495. https://doi.org/10.1186/s12887-023-04079-x.

Lever, J., Krzywinski, M., Altman, N., 2016. Model selection and overfitting. Nat. Methods 13 (9), 703–704. https://doi.org/10.1038/nmeth.3968.

McGregor, K.K., 2020. How we fail children with developmental language disorder. Lang. Speech Hear. V Sch. 51 (4), 981–992. https://doi.org/10.1044/2020_LSHSS-20-00003.

McKean, C., Law, J., Mensah, F., Cini, E., Eadie, P., Frazer, K., Reilly, S., 2016. Predicting meaningful differences in school-entry language skills from child and family factors measured at 12 months of age. Int. J. Early. Child 48 (3), 329–351. https://doi.org/10.1007/s13158-016-0174-0.

McKean, C., Wraith, D., Eadie, P., Cook, F., Mensah, F., Reilly, S., 2017. Subgroups in language trajectories from 4 to 11 years: the nature and predictors of stable, improving and decreasing language trajectory groups. J. Child Psychol. Psychiatry 58 (10), 1081–1091. https://doi.org/10.1111/jcpp.12790.

Mountford, H.S., Braden, R., Newbury, D.F., Morgan, A.T., 2022. The genetic and molecular basis of developmental language disorder: a review. Children 9 (5), 586. https://doi.org/10.3390/children9050586.

Murray, G.K., Lin, T., Austin, J., McGrath, J.J., Hickie, I.B., Wray, N.R., 2021. Could polygenic risk scores be useful in psychiatry?: a review. JAMa Psychiatry 78 (2), 210. https://doi.org/10.1001/jamapsychiatry.2020.3042.

Newbury, D.F., Gibson, J.L., Conti-Ramsden, G., Pickles, A., Durkin, K., Toseeb, U., 2019. Using polygenic profiles to predict variation in language and psychosocial outcomes in early and middle childhood. J. Speech. Lang. Hear. Res. 62 (9), 3381–3396. https://doi.org/10.1044/2019_JSLHR-L-19-0001.

Norbury, C.F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., Pickles, A., 2016. The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. J. Child Psychol. Psychiatry 57 (11), 1247–1257. https://doi.org/10.1111/jcpp.12573.

Pain, O., Glanville, K.P., Hagenaars, S.P., Selzam, S., Fürtjes, A.E., Gaspar, H.A., Coleman, J.R.I., Rimfeld, K., Breen, G., Plomin, R., Folkersen, L., Lewis, C.M., 2021. Evaluation of polygenic prediction methodology within a reference-standardized framework. PLoS. Genet. 17 (5), e1009021. https://doi.org/10.1371/journal.pgen.1009021.

Pearson Education, 2008. Clinical Evaluation of Language Fundamentals–Fourth Edition Technical Report. https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/celf-4/celf-4-technical-report.pdf.

Plomin, R., Von Stumm, S., 2022. Polygenic scores: prediction versus explanation. Mol. Psychiatry 27 (1), 49–52. https://doi.org/10.1038/s41380-021-01348-y.

Polley, E., LeDell, E., Kennedy, C., van der Laan, M., 2021. SuperLearner: Super Learner Prediction (Version R package version 2.0-28) [Computer software]. https://CRAN.R-project.org/package=SuperLearner.

Qi, Q., Li, Y., Chomistek, A.K., Kang, J.H., Curhan, G.C., Pasquale, L.R., Willett, W.C., Rimm, E.B., Hu, F.B., Qi, L., 2012. Television watching, leisure time physical activity, and the genetic predisposition in relation to body mass index in women and men. Circulation 126 (15), 1821–1827. https://doi.org/10.1161/CIRCULATIONAHA.112.098061.

R Core Team. (2020). *R: A language and Environment for Statistical Computing* (Version 3.6.3) [Computer Software]. R Foundation for Statistical Computing. https://www.Rproject.org/.

Redmond, S.M., Ash, A.C., Christopulos, T.T., Pfaff, T., 2019. Diagnostic accuracy of sentence recall and past tense measures for identifying children's language impairments. J. Speech. Lang. Hear. Res. 62 (7), 2438–2454. https://doi.org/10.1044/2019_JSLHR-L-18-0388.

Reilly, S., Cook, F., Bavin, E.L., Bretherton, L., Cahir, P., Eadie, P., Gold, L., Mensah, F., Papadopoullos, S., Wake, M., 2018. Cohort profile: the early language in Victoria study (ELVS). Int. J. Epidemiol. 47 (1), 11–20. https://doi.org/10.1093/ije/dyx079.

Reilly, S., McKean, C., Levickis, P., 2014. Late Talking: Can it Predict Later Language Difficulties? Centre for Research Excellence in Child Language, pp. 1–2. Research Snapshot No. 2. https://www.mcri.edu.au/sites/default/files/media/documents/crec_rs2_late-talkers-1_design_v0.1_0.pdf.

Rescorla, L., 1989. The language development survey: a screening tool for delayed language in toddlers. J. Speech. Hear. Disord. 54 (4), 587–599. https://doi.org/10.1044/jshd.5404.587.

Richards, B., Skoletsky, J., Shuber, A.P., Balfour, R., Stern, R.C., Dorkin, H.L., Parad, R. B., Witt, D., Klinger, K.W., 1993. Multiplex PCR amplification from the CFTR gene using DNA prepared from buccal brushes/swabs. Hum. Mol. Genet. 2 (2), 159–163. https://doi.org/10.1093/hmg/2.2.159.

RStudioTeam, 2020. *RStudio: Integrated development Environment for R* (Version 1.3.959) [Computer Software]. RStudio, PBC. http://www.rstudio.com/.

Sanson, A., Johnstone, R., 2004. Growing up in Australia" takes its first steps. Fam. Matters. 67, 46–53.

Schiavon, M., Burton, B.K., Hemager, N., Greve, A.N., Spang, K.S., Ellersgaard, D., Plessen, K.J., Jepsen, J.R.M., Thorup, A.A.E., Werge, T., Nordentoft, M., Nudel, R., 2024. Language, motor ability and related deficits in children at familial risk of schizophrenia or bipolar disorder. Schizophr. Bull. sbae181. https://doi.org/10.1093/schbul/sbae181.

Semel, E., Wiig, E.H., Secord, W.A., 2006. Clinical Evaluation of Language Fundamentals- Fourth Edition, Australian Standardised Edition, 4th edition. Harcourt Assessment. https://www.pearsonclinical.com.au/products/view/86.

Smith, J., Wang, J., Grobler, A.C., Lange, K., Clifford, S.A., Wake, M., 2019. Hearing, speech reception, vocabulary and language: population epidemiology and concordance in Australian children aged 11 to 12 years and their parents. BMJ Open. 9 (Suppl 3), 85–94. https://doi.org/10.1136/bmjopen-2018-023196.

Van Barreveld, M., Scheper, A., Vissers, C., Duinmeijer, I., Hakvoort, B., 2025. Social–Emotional functioning and quality of life in language disorders: a systematic review of development from childhood to adolescence. Int. J. Lang. Commun. Disord. 60 (3), e70039. https://doi.org/10.1111/1460-6984.70039.

van der Laan, M.J., Polley, E.C., Hubbard, A.E., 2007. Super learner. Stat. Appl. Genet. Mol. Biol. 6 (1). https://doi.org/10.2202/1544-6115.1309. Article 25.

Verhoef, E., Allegrini, A.G., Jansen, P.R., Lange, K., Wang, C.A., Morgan, A.T., Ahluwalia, T.S., Symeonides, C., Eising, E., Franken, M.-C., Hypponen, E., Mansell, T., Olislagers, M., Omerovic, E., Rimfeld, K., Schlag, F., Selzam, S., Shapland, C.Y., Tiemeier, H., Barwon Infant Study Investigator Group, 2023. Genome-wide analyses of vocabulary size in infancy and toddlerhood: associations with attention-deficit/hyperactivity disorder, literacy, and cognition-related traits. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2023.11.025. S0006322323017572.

Verhoef, E., Shapland, C.Y., Fisher, S.E., Dale, P.S., St Pourcain, B., 2020. The developmental origins of genetic factors influencing language and literacy: associations with early-childhood vocabulary: developmental origins of genetic factors influencing language and literacy. J. Child Psychol. Psychiatry. https://doi.org/10.1111/jcpp.13327.

von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C., Vandenbroucke, J.P., for the STROBE initiative, 2007. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370 (9596), 1453–1457. https://doi.org/10.1016/S0140-6736(07)61602-X.

Wallace, I.F., Berkman, N.D., Watson, L.R., Coyne-Beasley, T., Wood, C.T., Cullen, K., Lohr, K.N., 2015. Screening for speech and language delay in children 5 years old and younger: a systematic review. Pediatrics 136 (2), e448–e462. https://doi.org/10.1542/peds.2014-3889.

Wang, J., le Clercq, C.M.P., Sung, V., Carew, P., Liu, R.S., Mensah, F.K., Burt, R.A., Gold, L., Wake, M., 2018. Cross-sectional epidemiology of hearing loss in Australian children aged 11–12 years old and 25-year secular trends. Arch. Dis. Child 103 (6), 579–585. https://doi.org/10.1136/archdischild-2017-313505.

Ziegenfusz, S., Paynter, J., Flückiger, B., Westerveld, M.F., 2022. A systematic review of the academic achievement of primary and secondary school-aged students with developmental language disorder. Autism. Dev. Lang. Impair. 7, 239694152210993. https://doi.org/10.1177/23969415221099397.