Received: 9 November 2023 | Accepted: 11 June 2024

DOI: 10.1111/1460-6984.13086

RESEARCH REPORT

International Journal of Language & Communication Disorders

Identifying early language predictors: A replication of Gasparini et al. (2023) confirming applicability in a general population cohort

Loretta Gasparini[1,2] | Daisy A. Shepherd[1,2] | Jing Wang[1,2] | Melissa Wake[1,2,3] | Angela T. Morgan[2,4,5]

[1]Department of Paediatrics, The University of Melbourne, Parkville, VIC, Australia

[2]Murdoch Children's Research Institute, Parkville, VIC, Australia

[3]Liggins Institute, The University of Auckland, Grafton, Auckland, New Zealand

[4]Department of Audiology and Speech Pathology, The University of Melbourne, Parkville, VIC, Australia

[5]Royal Children's Hospital Melbourne, Parkville, VIC, Australia

Correspondence

Loretta Gasparini, Royal Children's Hospital, 50 Flemington Rd, Parkville, VIC 3052, Australia.
Email:
lgasparini@student.unimelb.edu.au

Melissa Wake and Angela T. Morgan are joint senior authors.

Funding information

Victorian Government's Operational Infrastructure Support Program; NHMRC Investigator grant, Grant/Award Number: 1195955; MCRI PhD Top Up Scholarship; Australian Government Research Training Program (RTP) Scholarship

Abstract

Background: Identifying language disorders earlier can help children receive the support needed to improve developmental outcomes and quality of life. Despite the prevalence and impacts of persistent language disorder, there are surprisingly no robust predictor tools available. This makes it difficult for researchers to recruit young children into early intervention trials, which in turn impedes advances in providing effective early interventions to children who need it.

Aims: To validate externally a predictor set of six variables previously identified to be predictive of language at 11 years of age, using data from the Longitudinal Study of Australian Children (LSAC) birth cohort. Also, to examine whether additional LSAC variables arose as predictive of language outcome.

Methods & Procedures: A total of 5107 children were recruited to LSAC with developmental measures collected from 0 to 3 years. At 11–12 years, children completed the Clinical Evaluation of Language Fundamentals, 4th Edition, Recalling Sentences subtest. We used SuperLearner to estimate the accuracy of six previously identified parent-reported variables from ages 2–3 years in predicting low language (sentence recall score $\geq$ 1.5 SD below the mean) at 11–12 years. Random forests were used to identify any additional variables predictive of language outcome.

Outcomes & Results: Complete data were available for 523 participants (52.20% girls), 27 (5.16%) of whom had a low language score. The six predictors yielded fair accuracy: 78% sensitivity (95% confidence interval (CI) = [58, 91]) and 71% specificity (95% CI = [67, 75]). These predictors relate to sentence complexity, vocabulary and behaviour. The random forests analysis identified similar predictors.

Conclusions & Implications: We identified an ultra-short set of variables that predicts 11–12-year language outcome with 'fair' accuracy. In one of few replication studies of this scale in the field, these methods have now been conducted across two population-based cohorts, with consistent results. An imminent practical implication of these findings is using these predictors to aid recruitment into early language intervention studies. Future research can continue to refine the accuracy of early predictors to work towards earlier identification in a clinical context.

KEYWORDS
language disorders, sensitivity and specificity, longitudinal studies, machine learning, random forests, SuperLearner

WHAT THIS PAPER ADDS

What is already known on the subject

- There are no robust predictor sets of child language disorder despite its prevalence and far-reaching impacts. A previous study identified six variables collected at age 2–3 years that predicted 11–12-year language with 75% sensitivity and 81% specificity, which warranted replication in a separate cohort.

What this study adds to the existing knowledge

- We used machine learning methods to identify a set of six questions asked at age 2–3 years with ≥ 71% sensitivity and specificity for predicting low language outcome at 11–12 years, now showing consistent results across two large-scale population-based cohort studies.

What are the potential or clinical implications of this work?

- This predictor set is more accurate than existing feasible methods and can be translated into a low-resource and time-efficient recruitment tool for early language intervention studies, leading to improved clinical service provision for young children likely to have persisting language difficulties.

INTRODUCTION

Language disorders are defined as language difficulties that impact everyday functioning (Bishop et al., 2017). They lead to poorer socio-behavioural, academic, employment and quality-of-life outcomes (Conti-Ramsden & Durkin, 2012; Eadie et al., 2018; Yew & O'Kearney, 2013; Ziegenfusz et al., 2022). Prevalence estimates range from 7% to 10% of children, and an estimated 6–7% of children have a language disorder that cannot be attributed to any specific condition or environmental factor (denoted developmental language disorder—DLD; Calder et al., 2022; Norbury et al., 2016). Language disorders are unlikely to resolve without specialist intervention (Bishop et al., 2017). Language disorders, including DLD, are largely unknown to the community, under-researched and critically underserviced (McGregor, 2020).

Identifying lasting language disorder at an early age across the population is desirable, so clinicians and educators can provide targeted support while avoiding overservicing. This is complicated, however, because early developmental delays often resolve. Conversely, lasting delays may not manifest in very young children on currently available tests suitable for at-scale use. Thus, 6% of all children shift between classifications of typical and low language between 4 and 11 years (McKean et al., 2017), which yields a high error rate considering the estimated prevalence of 7–10% (Calder et al., 2022; Norbury et al., 2016). To best support the children who will have persisting difficulties using limited available resources, we need early measures that accurately predict language outcome in late childhood.

The difficulty in accurately identifying young children with persisting language difficulties hinders interventional research. McGregor (2020) describes how a lack of awareness of, services for, and research about DLD feed each other, resulting in systematically failing children with DLD. Providing effective, early intervention to children is desirable, because research suggests that modifiable factors influencing language development throughout childhood are already established in the preschool years (McKean et al., 2015). Currently, a standard approach to interventional research is to recruit 'late-talking' children (those with small productive vocabularies at 2 years) into trials, which has been found to have 84% specificity, but only 45% sensitivity, for predicting lasting language outcomes (Reilly et al., 2014). One early language intervention trial using this recruitment approach reported yielding null results as they found that the majority of recruited late-talking children caught up spontaneously, regardless of the intervention (Wake et al., 2011). This spontaneous improvement of children in both the intervention and control groups drowns out the impact of the intervention for the children who would otherwise have persisting difficulties and for whom therapy is most beneficial. There is a critical need to design high-quality language intervention studies with adequate statistical power to reveal the effects of an intervention (Donolato et al., 2023). Accurately identifying later persisting language difficulties in early life can enable intervention programs to more precisely recruit those children who require support, increasing their power to yield meaningful results. A more robust evidence base for early intervention as well as accurate early detection methods could lead to more efficient use of services. This would mean the children who are in most need of language support receives it.

A 2024 systematic review by the U.S. Preventive Services Task Force found insufficient evidence to support screening for speech and language disorders in children ≤ 5 years in the general population (Feltner et al., 2024). Although they identified screening tools with adequate concurrent validity for detecting speech and language delays, they noted that most instruments were unable to discern between children whose delay would resolve without needing intervention, and those who would develop a persisting language disorder. Replications were also sparse. Furthermore, many instruments with adequate accuracy were either lengthy surveys with over 50 items or tools requiring assessment by a speech–language therapist.

While progress has been made identifying short sets of language predictors (e.g., Armstrong et al., 2018; Borovsky et al., 2021; Rudolph & Leonard, 2016; Stott et al., 2002; Wilson et al., 2022), they all still have limitations when it comes to predictive validity or feasibility for widespread application. For an instrument to be suitable for early identification of children likely to have persisting language difficulties across the population, it must (1) have satisfactory accuracy in the early years both in identifying true cases of low and typical language outcome, (2) over enduring timeframes into late childhood, (3) with replicable results, and it must be (4) time- and (5) resource-efficient to administer.

Together, this replication study, along with the previous study by Gasparini, Shepherd, Bavin, et al. (2023), uniquely, to our knowledge, tick these five boxes (e.g., see cited studies in Feltner et al., 2024). Using data from the Early Language in Victoria Study (ELVS), Gasparini, Shepherd, Bavin, et al. (2023) identified short sets of parent-reported survey items (maximum eight, approximating 1 min administration time) that predicted 11-year language outcome with 70–85% sensitivity and specificity (considered 'fair' to 'good' accuracy according to their preregistered thresholds, based on previous studies citing > 70% sensitivity and specificity as acceptable to screen for developmental delays; Council on Children With Disabilities et al., 2006; Wallace et al., 2015). This study was one of only very few to date in the language development field to use machine learning methods to identify childhood language predictors (see also Armstrong et al., 2018; Borovsky et al., 2021). These predictors accorded with previous literature identifying early predictors of language outcome, namely sentence complexity (Armstrong et al., 2018; Borovsky et al., 2021; Rudolph & Leonard, 2016; Sansavini et al., 2021), vocabulary (Sansavini et al., 2021; Zambrana et al., 2014) and behaviour (Law et al., 2012).

The Gasparini, Shepherd, Bavin, et al. (2023) predictor sets are not suitable to imminently translate into a clinical screening instrument. While the point estimates may be of acceptable levels, the estimated sensitivity had wide 95% confidence intervals (CIs) reaching as low as 58% due to the low population prevalence of language disorder. The identified predictor sets require an exploration of novel predictors and use of very large samples to further increase accuracy and precision. In addition to greater accuracy, more robust evidence on effective early language intervention, and the weighting of benefits versus adverse effects of screening are needed before implementing population-wide detection of children likely to have persisting language difficulties (Feltner et al., 2024).

In contrast, we consider the Gasparini, Shepherd, Bavin, et al. (2023) predictor sets to have good potential for recruiting young children likely to have persisting language difficulties into early intervention trials. Early language intervention remains desirable yet elusive (McKean et al., 2015) and children today need better evidence-based services (McGregor, 2020). While the field may still be far from implementing population-wide early language screening, there is an urgent need to leverage incremental advances in knowledge to improve services, giving children with language disorder the best opportunities to thrive. Recruitment methods for early interventional trials must strike a balance between accurately identifying true cases and being feasible to implement. For instance, one intervention study recruited the five children from each participating classroom with the lowest language scores at 4 years, rather than relying on clinical cut-offs (West et al., 2021). Previous work has found that recruiting late-talking children with 45% sensitivity for persisting language difficulties is not suitable to enrich a trial sample with enough true cases to yield precise results (Wake et al., 2011). Thus, the Gasparini, Shepherd, Bavin, et al. (2023) predictor sets have the potential to increase the number of true cases of children who would have persisting language difficulties recruited into early intervention trials.

Prediction methods can be vulnerable to overfitting; when high accuracy occurs only as a result of artefacts in the dataset that do not generalize to the wider population (Lever et al., 2016). The methods of Gasparini, Shepherd, Bavin, et al. (2023) methods warrant replication leveraging existing data from a separate cohort study to assess how well the results generalize to the wider Australian population. Growing Up in Australia: The Longitudinal Study of Australian Children (LSAC), like ELVS, is a population-based cohort study that recruited infants in the early 2000s. LSAC regularly collected measures on the family and home environment, health, development, communication and language throughout infancy, childhood, and adolescence. Thus, we deemed LSAC an appropriate cohort for replicating the methods of Gasparini, Shepherd, Bavin, et al. (2023).

AIMS

The current study replicated a previous study of early predictors of later language (Gasparini, Shepherd, Bavin, et al., 2023) using the LSAC birth cohort dataset.

Aim 1: External validation of the ELVS predictor set

Here we aimed to estimate the accuracy of sets of previously identified parent-reported predictors in the LSAC cohort and compared these results to the original analysis using ELVS data. Note that because of differences in the measures collected by ELVS and LSAC, we here replicated the supplementary results of Gasparini, Shepherd, Bavin, et al. (2023) (see their Supporting information 10 at https://osf.io/fpdzk/), rather than their primary results. This analysis identified six predictors collected at 24 or 36 months (n = 757) that predicted low language ability at 11 years with 78% Area Under the receiving operating characteristic Curve (AUC, 95% CI = [68, 88]), 75% sensitivity (95% CI = [58, 88]) and 81% specificity (95% CI = [78, 83]). We expected that sensitivity and specificity would attenuate to some degree in this study, possibly due to the previous study's results being overfitted to noise in the data, or methodological design differences between the LSAC and ELVS cohorts (see Supporting information 1). Although we expected some attenuation of accuracy, we expected that sensitivity and specificity may remain 'fair' (≥ 70%).

Aim 2: Replication of the ELVS study methodology

Here we aimed to identify and estimate the accuracy of any additional variables that may contribute to predicting language outcome in late childhood in the LSAC cohort. We reproduced all the methods from Gasparini, Shepherd, Bavin, et al. (2023) using data from LSAC. We expected predictors of language outcome arising in the Aim 2 analysis to accord largely with Gasparini, Shepherd, Bavin, et al. (2023) and previous research: factors relating to syntax, vocabulary, gestures, communication, parental stress, socioeconomic position and parent–child interactions. We expected that the predictor sets using the variables identified in the Aim 2 analysis may have higher accuracy than those in the Aim 1 analysis, although they will not have been externally validated.

Identifying a set of predictors with satisfactory and replicable predictive accuracy will facilitate earlier identification of children likely to have persisting language difficulties through to late childhood.

METHODS

We preregistered the hypotheses and statistical analysis methods on Open Science Framework (OSF) on 27 July 2022 at https://osf.io/jk32c/, prior to data access. We published and timestamped protocol amendments and all Supporting information on the OSF repository. LSAC is approved by the Australian Institute of Family Studies Ethics Committee, and caregivers provided written, informed consent. Permission was granted to use the LSAC data for the current study. The data necessary to reproduce the analyses presented here can be made accessible via the Australian Data Archive. As this is a confirmatory, replication study of Gasparini, Shepherd, Bavin, et al. (2023), Supporting information 1 summarizes all differences between Gasparini, Shepherd, Bavin, et al. (2023) and the current study. In short, LSAC is a general population cohort with fewer exclusion criteria than ELVS. Thus, we expect LSAC to be a cohort more representative of the Australian population, which would increase the external validity of Gasparini, Shepherd, Bavin, et al. (2023) if the results replicate. The outcome measure between the two studies is different (see section: Outcome variable) which may introduce measurement error. We expect all other minor differences to have negligible effects on the results. We follow the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines (von Elm et al., 2007) and provide a checklist in Supporting information 2.

Study design and participants

LSAC recruited a representative birth cohort of 5107 infants across Australia born between March 2003 and February 2004 using a clustered design (Edwards, 2014). 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. The baseline recruitment rate of families approached in 2004 was 57.2%. The recruited families completed wave 1 at recruitment and were followed up in biennial waves thereafter. Of the families recruited in wave 1, 73.7% ($n = 3764$) remained in the study until wave 6, of whom 93.3% ($n = 3513$) then consented to being invited to participate in Child Health CheckPoint. The Child Health CheckPoint (henceforth LSAC CheckPoint) recruited 53.3% ($n = 1874$) of those families into a one-off detailed cross-sectional assessment of the study child and one parent, nested between LSAC waves 6 and 7 when the children were aged 11–12 years (Clifford et al., 2019). The LSAC CheckPoint cohort was found to represent the broader Australian population in terms of geographical distribution, but on average is more socioeconomically advantaged, has parents with higher levels of education and fewer Aboriginal and Torres Strait Islander families and families from non-English-speaking backgrounds (Clifford et al., 2019). We include all LSAC CheckPoint children with an available sentence repetition measure ($n = 1441$, 16.1% of families approached to join LSAC and 28.2% who completed wave 1) in our analyses unless otherwise specified (section: Additional analyses). Figure 1 provides a flow chart of participant recruitment and attrition.

Data collection procedures

LSAC collected data when the participants were aged 0–1 years (waves 1 and 1.5) and 2–3 years (wave 2). The 11–12-year language outcome measure was collected during the 2.75–3.5-h LSAC CheckPoint assessment centre visit (nested between LSAC waves 6 and 7), where participants completed numerous assessments of physical and social–cognitive health and development (Smith et al., 2019) (see section: Outcome variable).

Variables

Predictor variables

Variables from LSAC wave 2 (collected at age 2–3 years) were selected for the Aim 1 analysis based on the Gasparini, Shepherd, Bavin, et al. (2023) results. Table 1 shows the predictors that constituted the predictor sets with the highest accuracy in the Supporting information 10 analysis of Gasparini, Shepherd, Bavin, et al. (2023; their supporting information is available at https://osf.io/fpdzk/), which involved only variables collected by both ELVS and LSAC. Supporting information 3 includes the source of each variable.

The Aim 2 analysis comprises selected variables from LSAC waves 1, 1.5 and 2. While Gasparini, Shepherd, Bavin, et al. (2023) included all variables collected between 8 and 36 months in their corresponding analysis, LSAC waves 1–2 contained substantially more variables (2102 at wave 1, 259 at wave 1.5 and 2674 at wave 2). Thus, it was not computationally feasible to perform the analysis over the complete set of variables in the current study. Hence, we selected a constrained list of variables to encompass areas we consider to be related to language development based on previous literature. As such, the current Aim 2 analysis includes 799 variables from LSAC wave 1, 121 from wave 1.5 and 965 from wave 2. Supporting information 3 summarizes the variables included in the Aim 2 analysis, grouped by construct, with justifications. The full list of variables included in this analysis can be found in Supporting information 4. To avoid convergence issues when running the random forests, we reduced factors with > 8 levels to ≤ 8 by merging conceptually similar levels.

Outcome variable

LSAC CheckPoint administered the Recalling Sentences subtest from the CELF-4 (Semel et al., 2006), where participants heard audio recordings of sentences via audio file

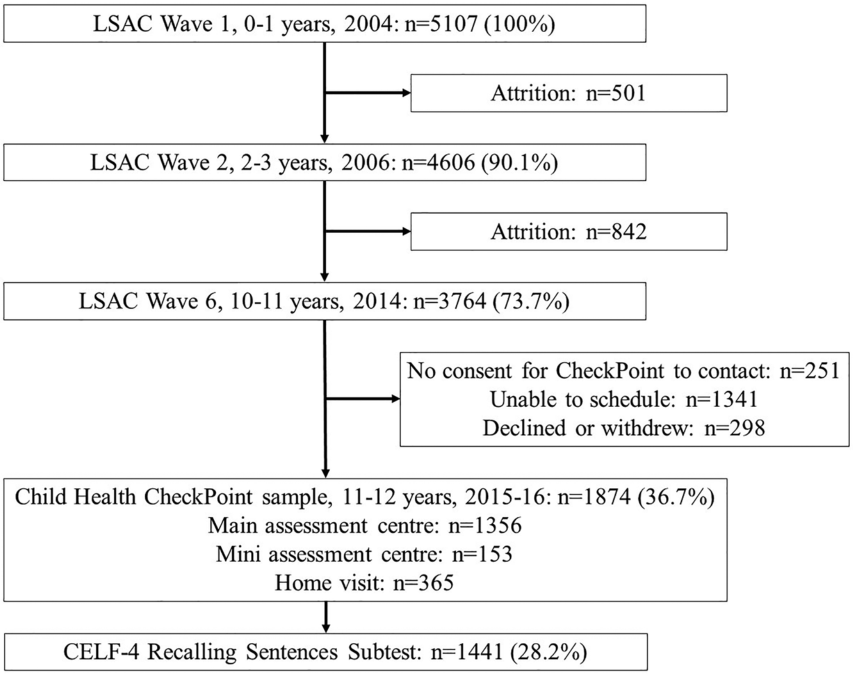
LSAC Wave 1, 0-1 years, 2004: n=5107 (100%)
Attrition: n=501
LSAC Wave 2, 2-3 years, 2006: n=4606 (90.1%)
Attrition: n=842
LSAC Wave 6, 10-11 years, 2014: n=3764 (73.7%)
No consent for CheckPoint to contact: n=251
Unable to schedule: n=1341
Declined or withdrew: n=298
Child Health CheckPoint sample, 11-12 years, 2015-16: n=1874 (36.7%)
Main assessment centre: n=1356
Mini assessment centre: n=153
Home visit: n=365
CELF-4 Recalling Sentences Subtest: n=1441 (28.2%)

FIGURE 1 Participant flowchart.

TABLE 1 Predictors included in the Aim 1 analysis externally validating the ELVS study predictor set. This includes the construct each predictor maps to, and examples of, previous literature that identified the same constructs as language predictors.

Question	Options	Construct
Tick the sentence that sounds most like the way the study child currently talks	This dolly big; This dolly big and this dolly little	Sentence complexity[a,b,d,e]
Child says ‘circle’	No; Yes	Productive vocabulary[e,f]
Child says ‘accident’	No; Yes	Productive vocabulary[e,f]
What are your concerns? Other behaviour difficulties	No; A little; Yes	Behaviour[c]
Child says ‘kangaroo’	No; Yes	Productive vocabulary[e,f]
Child says ‘forget/forgot’	No; Yes	Productive vocabulary[e,f]

Sources: [a]Armstrong et al. (2018); [b]Borovsky et al. (2021); [c]Law et al. (2012); [d]Rudolph and Leonard (2016); [e]Sansavini et al. (2021); [f]Zambrana et al. (2014).

on an iPad and were asked to repeat them verbatim. We used the CELF-4 Recalling Sentences scaled score as the outcome measure of this study. This differs from the Gasparini, Shepherd, Bavin, et al. (2023) outcome measure, the CELF-4 core language score, because LSAC CheckPoint only administered the CELF-4 Recalling Sentences subtest due to time constraints. The CELF-4 is used as a diagnostic tool for language disorder, and a core score >1.5 standard deviations (SD) below the mean was found to have 100% sensitivity and 89% specificity for diagnosing language disorder (Pearson Education, 2008). Using ELVS data, the LSAC CheckPoint Investigators found that the CELF-4 Recalling Sentences scaled score had the highest accuracy (AUC = 0.96) in predicting the total CELF-4 core, expressive and receptive language scores (Smith et al., 2019), which motivated them to administer the CELF-4 Recalling Sentences subtest in LSAC CheckPoint. Other studies have found sentence repetition to strongly agree with broad language skills (Botting et al., 2001; Klem et al., 2015), thus we consider this subtest a suitable indicator of low language outcome. The CELF-4 Recalling Sentences subtest has total scores ranging from 0 to 96 and age-related scaled scores ranging from 1 to 18 with a mean of 10 and SD of 3. In this study, we operationalize low language ability (as a potential indicator, but not a diagnosis, of language disorder) as a scaled score ≤ 5.5 (1.5SD below the population mean). We also conducted sensitivity analyses using different dichotomization thresholds and outcome measures (see section: Additional analyses).

Statistical methods

We conducted statistical analyses in RStudio (R Core Team, 2020; RStudio Team, 2020) and code is available on the OSF repository. We provide more details of the statistical methods in Supporting information 5 (SuperLearner, Aims 1 and 2) and 6 (random forests, Aim 2).

Aim 1: External validation of the ELVS predictor set

To address Aim 1, we estimated the accuracy of the predictor sets identified by Gasparini, Shepherd, Bavin, et al. (2023) and listed in Table 1. To achieve this, we reproduced the Supporting information 10 analysis of Gasparini, Shepherd, Bavin, et al. (2023; their supporting information can be found at https://osf.io/fpdzk/) that estimated the accuracy of a predictor set with common variables between ELVS and LSAC. Unless otherwise specified, the analysis described below is identical to Gasparini, Shepherd, Bavin, et al. (2023) Supporting information 10. This analysis used SuperLearner, an ensemble algorithm that utilizes various prediction algorithms, evaluates the accuracy of each algorithm and weights them accordingly in a new single prediction algorithm that is expected to perform at least as well as any of the individual methods (van der Laan et al., 2007). We conducted the Aim 1 analysis using data from individuals with complete data on the predictors and outcome.

We ran models that included the six predictors of the best-fitting predictor set in Gasparini, Shepherd, Bavin, et al. (2023). We applied 10-fold cross-validation and included a variety of individual algorithms in our SuperLearner algorithm (see Supporting information 5 for details). We calculated the sensitivity, specificity and AUC of the predictor sets with their estimated 95% CIs. We calculated and reported sensitivity and specificity at a cut-point where the two are balanced, to optimize both (see section: Practical implications for how a more flexible approach is possible in future uses).

Gasparini, Shepherd, Bavin, et al. (2023) also included variables collected at 8–12 months, but they yielded unsatisfactory accuracy (< 70%). They also combined predictors collected from 8 to 36 months, but accuracy did not improve over the 24–36-month predictors. For completeness we ran these analyses but only report them in Supporting information 10.

Aim 2: Replication of the ELVS study methodology

To address Aim 2, we reproduced the entire methodology of Gasparini, Shepherd, Bavin, et al. (2023). This involved ranking many LSAC variables by how well they predict the language outcome, selecting sets of the most predictive variables and estimating the accuracy of these sets for predicting language outcome. This analysis used random forests to estimate the 'importance' (see below) of 1885 variables collected at 0–1 or 2–3 years in predicting the outcome measure (see section: Predictor variables for how these 1885 variables were selected). Random forests is a tree-based machine-learning algorithm used for classification and regression. It is suitable for investigating the role of a large number of variables and ranking them by order of 'importance' (Breiman, 2001). 'Importance' is a technical statistical term, where a larger importance value indicates a closer relationship between a predictor and the outcome (Strobl et al., 2009). We conducted this analysis to establish whether any additional variables to those in Aim 1 are worth considering for their predictive value, for instance, any variables not collected by ELVS but available in LSAC. As all the predictor sets containing variables collected before 2 years yielded unsatisfactory accuracy in Gasparini, Shepherd, Bavin, et al. (2023), we included wave 1 and 1.5 (collected at 0–1 years) variables in this study to see if any predictor set would yield satisfactory (> 70%) accuracy.

We removed from analysis potential predictors with a large amount of missingness (> 50%) and applied multiple imputation for variables with < 50% missing data using 100 iterations (Liaw & Wiener, 2002; and see Supporting information 6). We created separate random forests for variables from data collection waves 1 and 1.5 combined (both collected at participant age 0–1 years) and wave 2 (age 2–3 years). For each of 100 different databases with different imputed values, we created a random forest of 300 unbiased conditional inference trees (Strobl et al., 2009). We estimated the 'importance' of each variable using conditional permutation importance without replacement (Strobl et al., 2009), where a larger importance value indicates a closer relationship between the given predictor and outcome. We averaged variable importance across the 100 imputed datasets which resulted in all the predictors ranked by their estimated importance in predicting the language outcome.

We used these rankings to select sets of variables from each timepoint (ages 0–1 and 2–3 years) that have high 'importance' for predicting the language outcome, accord with previous research and are feasible for adapting into an approximately 1-min parent-reported survey. Specifically, we wanted maximum eight variables that we deemed easy to understand and answer by any caregiver and appropriate for a clinician or researcher to ask in various contexts. We ran SuperLearner with the same prediction algorithms and parameters as the Aim 1 analysis (consistent with Gasparini, Shepherd, Bavin, et al., 2023), and report the sensitivity, specificity (at cut-points where the two are optimized) and AUC alongside their 95% CIs for each predictor set.

Additional analyses

We conducted several additional analyses to support our findings from the above analyses. We ran univariate logistic regressions on all variables included in the final predictor sets, to check whether the effects were in the directions we expected. We also ran univariate logistic regressions on a constrained number of predictors which we expected from previous literature might have a relationship with the outcome, to allow comparison with other literature (see Supporting information 9).

To determine whether predictor set accuracy appears stable regardless of how we operationalize low language, we reproduced the SuperLearner analyses with the cut-off of low and typical language at 1.25 and 2 SD below the mean CELF-4 Recalling Sentences scaled score. We also reproduced the analyses with alternative language outcome measures at separate ages: the Academic Rating Scale (ARS)—Language and Literacy score (Rock & Pollack, 2002) at age 12–13 and the Rice Test of Grammaticality Judgement (GJT) prime score (Rice et al., 2009) at age 14–15, all dichotomized at 1.25, 1.5 and 2 SD below the sample mean.

We ran random forests using complete cases only (individuals with no missing predictors) as a sensitivity check to assess the consistency of our results and evaluate the robustness of our imputation approach.

We wanted to determine whether the sets of predictors we identified would have adequate accuracy in the subgroup of multilingual children by assessing the classification accuracy of the predictors in the subgroup of participants whose parents report speaking an additional language to English at home at 0–1 or 2–3 years.

RESULTS

Participant characteristics

An 11–12-year CELF-4 Recalling Sentences score was available for 1441 (28.2%) children, and not available for 3666 (71.8%) of participants recruited at baseline. Comparing these two groups shows that the included participants were over-represented with English being the main language spoken to the child and Parent 1 having higher education levels (see Supporting information 7). This accords with a previous examination of representativeness of the LSAC CheckPoint cohort (Clifford et al., 2019). The included sample had a mean 11–12-year CELF-4 Recalling Sentences score of 10.12 (SD = 2.82), aligning with the standardized population scores (mean = 10, *SD* = 3).

Aim 1: External validation of the ELVS predictor set

Table 1 shows the variables included in the Aim 1 analysis. Table 2(a) shows the AUC, sensitivity, and specificity of the Aim 1: 2–3-year (wave 2) predictor set. The SuperLearner model had higher accuracy than any single algorithm. The predictor set had 'fair' (> 70%) sensitivity and specificity.

Figure 2 illustrates how the estimated sensitivity was similar between Gasparini, Shepherd, Bavin, et al. (2023) using ELVS data and the current study using LSAC data, increasing slightly in the current study but with overlapping 95% CIs that spanned from unsatisfactory (58%) to good/excellent (88–91%) values. Specificity attenuated in the current study, but in both cases surpassed the preregistered threshold of 'fair' accuracy (≥ 70%), although the 95% CI lower limit dropped to 67% in the LSAC cohort.

Aim 2: Replication of the ELVS study methodology

In Supporting information 8 we list the 30 variables with the highest estimated variable importance values from each data collection wave. We indicate whether we included them in the next part of the analysis and justify our decision to include or exclude.

The 0–1-year predictor set with the highest AUC included three variables (one each about child support, parental smoking, and parental concern about the child's language comprehension). The 2–3-year predictor set with the highest AUC included five variables (two vocabulary items and one question each about the child showing pleasure when they succeed, parental education and the child's sentence complexity). The 0–3-year predictor set with the highest AUC included nine variables: five variables from 0–1 years (child support, parental smoking, child's language comprehension, parental education, medications taken during pregnancy) and four from 2–3 years (two vocabulary items, child showing pleasure when they succeed, parental education).

Table 2(b) shows the AUC, sensitivity, and specificity of each of these final predictor sets, by data collection wave. In all cases, an individual prediction algorithm had higher accuracy than the combined SuperLearner model (conditional tree forest for 0–1 years, elastic net regression for 2–3 years, and Bayesian additive regression trees for 0–3 years). The 2–3-year predictor set had 'fair' (≥ 70%) sensitivity and specificity and the 0–1-year set was unsatisfactory (≤ 68%).

TABLE 2 Estimated AUC, sensitivity and specificity (and respective 95% CIs) by data collection wave obtained using SuperLearner. (a) Includes the same variables as in the ELVS cohort (Gasparini, Shepherd, Bavin, et al., 2023) and (b) includes variables of high importance according to random forests. 'Typical' and 'Low' indicate the number of participants included in each group after missing data were removed.

	n		**AUC**	**Sensitivity**	**Specificity**
	Typical	**Low**			
(a) Aim 1: External validation of the ELVS predictor set					
Wave 2 (2–3 years)	496	27	0.72 (0.61, 0.83)	0.78 (0.58, 0.91)	0.71 (0.67, 0.75)
(b) Aim 2: Replication of the ELVS study methodology					
Wave 1 (0–1 years)	1091	52	0.68 (0.60, 0.77)	0.63 (0.49, 0.76)	0.68 (0.65, 0.71)
Wave 2 (2–3 years)	1059	43	0.78 (0.69, 0.86)	0.74 (0.59, 0.86)	0.70 (0.67, 0.73)
Waves 1–2 (0–3 years)	958	42	0.81 (0.74, 0.88)	0.76 (0.61, 0.88)	0.71 (0.68, 0.74)

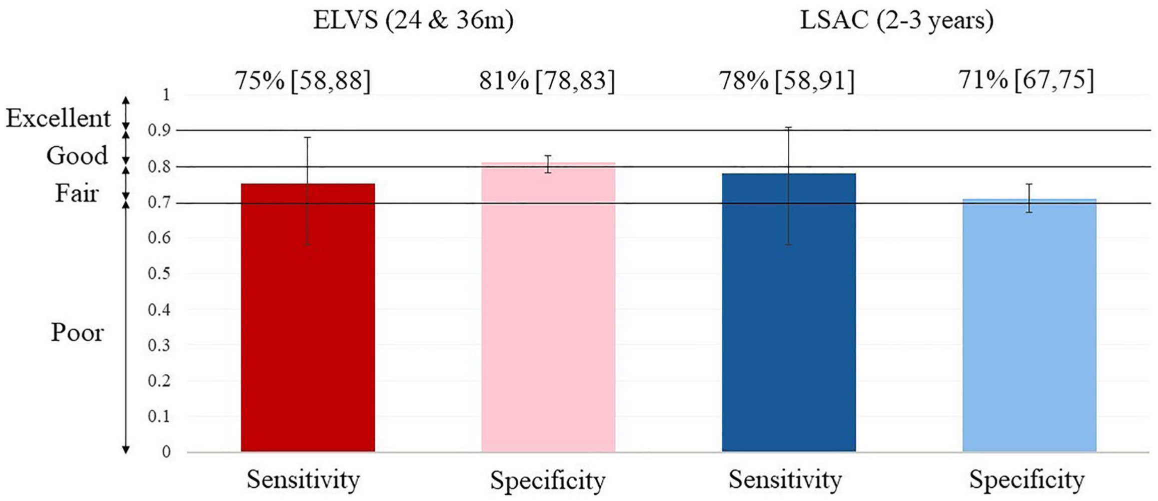

FIGURE 2 Estimated sensitivity (dark) and specificity (light) of the set of six predictors when using ELVS data (Gasparini, Shepherd, Bavin, et al., 2023) ($n = 757$, in red on the left) and in the current study using LSAC data ($n = 523$, in blue on the right). Error bars represent 95% CIs.

The 0–3-year set also yielded fair sensitivity and specificity but was not substantially better than the 2–3-year set. Table 3 lists the variables included in each of the three Aim 2 predictor sets. Supporting information 3 includes the source of each variable.

Additional analyses

Univariate analysis

Univariate logistic regression results are reported in Supporting information 9. For most of the Aim 1: 2–3-year variables, the lower limit of the odds ratio 95% CI was greater than 1 (mostly at around 2 or 3), meaning for these predictors there is reasonable confidence that the odds of a low language outcome is at least 2 or 3 times greater depending on the predictor level. The only exception was parental concern about their child's behaviour, where the odds ratio 95% CI spanned values below and above 1. Effects were in the expected directions according to previous literature: if a child was reported to produce more complex sentences (Armstrong et al., 2018; Borovsky et al., 2021; Rudolph & Leonard, 2016; Sansavini et al., 2021) or the given vocabulary items (Sansavini et al., 2021; Zambrana et al., 2014) at 2–3 years, they were more likely to be in the typical language group at 11–12 years.

Regarding variables included in the final Aim 2 predictor sets (Table 3), there were various univariate odds ratio 95% CIs that spanned values below and above 1 and others that were comfortably above the null value of 1. Again, in cases where the odds ratio 95% CI did not cross the null value, effects were generally in the expected directions according to previous literature.

TABLE 3 The Aim 2 predictor sets, selected by replicating the ELVS study methodology.

Question	Options
Wave 1 (0–1 years)	
Currently, does Parent 1 personally receive income from child support or maintenance (from ex-partner)?	No; Yes
On average, about how many cigarettes did you (Parent 1) smoke per day during this pregnancy, per day in the first 3 months?	None; ≤ 10; 11–20; 1–30; 31–40; 41–50; 51 or more
Do you have any concerns about how your child understands what you say to him/her?	No; A little; Yes
What was the highest year of primary or secondary school Parent 1 completed?[a]	Year 12 or equivalent; Year 11 or equivalent; Year 10 or equivalent; Year 9 or equivalent; Year 8 or below; Never attended school; Still at school
During the pregnancy with child, did you/child's mother take any medicines or tablets on a doctor's prescription?[a]	Yes; No
Wave 2 (2–3 years)	
Child says 'kitchen'[b]	No; Yes
Child says 'today'[b]	No; Yes
Child shows pleasure when he/she succeeds (e.g., claps for self)[b]	Not true/Rarely; Somewhat true/Sometimes; Very true/Often;
Has Parent 2 completed a trade certificate, diploma, degree or any other educational qualification?[b]	Yes; No
Tick the sentence that sounds most like the way the study child currently talks:	Don't read book; Don't want you read that book

Notes: [a] Included in the combined wave 1–2 predictor set only, not included in the wave 1 predictor set.
[b] Included in the combined wave 1–2 predictor set, as well as wave 2 predictor set.

Aim 1: External validation of the ELVS predictor set (SuperLearner)

Sensitivity analyses using alternative outcome measures generally failed to yield similar results to the main analysis (generally AUC < 70% but the CELF-4 Recalling Sentences subtest with 2 SD below the mean cut-off yielded AUC > 70%). We report results in Supporting information 10 and discuss this limitation below (section: Strengths and limitations).

Aim 2: Replication of the ELVS study methodology (random forests)

Random forests using complete cases analysis failed to converge due to high data missingness (0–1 years: $n = 1$, 2–3 years: $n = 38$). We discuss this limitation below (section: Strengths and limitations).

Aim 2: Replication of the ELVS study methodology (SuperLearner)

Sensitivity analyses of the Aim 2 predictor sets using alternative cut-offs of the 11–12-year Recalling Sentences score (1.25 and 2 SD below the mean) tended to yield similar scores to the main analysis (AUC < 70% for 0–1 years and AUC ≥ 70% for 2–3 years and 0–3 years). Sensitivity analyses using the alternative outcome measures (ARS at 12–13 years and GJT at 14–15 years) yielded unsatisfactory results (AUC < 70%). We report the results in Supporting information 10.

Subgroup analysis

The subgroup of children whose family spoke another language at home had very few participants in the low language groups (0–1 years: $n = 6$, 2–3 years: $n = 8$, 0–3 years: $n = 4$). As the number of predictors approximated the number of participants, we opted not to run these models, as precision would have been very low even if the models had managed to converge. We discuss this as a direction for future research (section: Future directions).

DISCUSSION

We have identified a set of six variables that can be asked at 2–3 years with 'fair' accuracy (≥ 71% sensitivity and specificity) for predicting which children will have low language skills at 11–12 years, replicated across two Australian population-based cohort studies: LSAC (in the current study) and ELVS (Gasparini, Shepherd, Bavin, et al., 2023). These predictors relate to constructs that have also been identified as early language predictors in other population-based studies, namely there is one predictor about the child's sentence complexity (Armstrong et al., 2018; Borovsky et al., 2021; Rudolph & Leonard, 2016; Sansavini et al., 2021), four vocabulary items (Sansavini et al., 2021; Zambrana et al., 2014) and one predictor about the child's behaviour (Law et al., 2012) (see Table 1 for the list of predictors).

Our Aim 2 analysis involved replicating Gasparini, Shepherd, Bavin, et al. (2023) entire methodology. We identified a set of five separate variables collected at 2–3 years (Table 3) with fair accuracy (≥ 70% sensitivity and specificity) for predicting language outcome. This Aim 2 predictor set includes similar constructs as the Aim 1 set, namely, two vocabulary items, one question on sentence complexity and one about temperament (related to general behaviour). This supports the inclusion of such variables in future prediction models. There was also a question about parental education, which accords with previous research (e.g., Tomblin et al., 1997). The Aim 2 predictor set did not improve in accuracy over the Aim 1 set and has not been externally validated across two cohorts. Like in Gasparini, Shepherd, Bavin, et al. (2023), variables collected before 2 years did not achieve satisfactory accuracy (< 70%) and combining predictors from 0 to 3 years did not improve accuracy compared with just including 2–3-year predictors. Thus, we will henceforth discuss the Aim 1: 2–3-year predictor set unless otherwise specified.

As expected, the Aim 1: 2–3-year predictor set reached slightly lower accuracy (lower specificity, but similar sensitivity at our selected cut-off) using the LSAC data than in the original ELVS dataset (Gasparini, Shepherd, Bavin, et al., 2023). This may be because the original models using ELVS data could have overfit to noise in the data, exaggerating the predictive accuracy estimates. The variation in results could also be due to methodological differences between LSAC and ELVS (see Supporting information 1). By replicating across two cohorts, we yield more conservative but more robust and generalizable estimates of the predictor set's accuracy for predicting the language outcome.

Strengths and limitations

Our study offers several strengths. It is a replication study, which yielded consistent results to the original study (Gasparini, Shepherd, Bavin, et al., 2023). This is important in prediction studies to increase confidence that results reflect true effects (Lever et al., 2016). Our use of machine learning and ensemble techniques offers many advantages. SuperLearner (Aims 1 and 2) does not restrict the user to selecting a single prediction algorithm, and thus minimizes the effects of arbitrary decisions, by running and evaluating multiple algorithms and parameters. Random forests (Aim 2) is non-parametric, robust to correlated predictors and can manage many variables with relatively few observations. Both methods use cross-validation which, along with replication, reduces overfitting (Lever et al., 2016).

LSAC recruited participants from across all of Australia, spanning socioeconomic positions, cultural backgrounds, and metropolitan and rural areas (excluding only very remote areas). Despite its broad inclusion criteria, the LSAC CheckPoint sample is more socioeconomically advantaged, has parents with higher levels of education and fewer Aboriginal and Torres Strait Islander families and families from non-English-speaking backgrounds than the wider Australian population (Clifford et al., 2019). LSAC did not collect data on participants' ethnicity other than whether they are Aboriginal and Torres Strait Islander. Thus, we cannot assume that our results generalize to the wider Australian population in terms of socioeconomic position and ethnicity.

We operationalized low language outcome based on the CELF-4 Recalling Sentences subtest at 11–12 years of age. Collecting a language outcome in late childhood is a strength because it allows us to identify children with persisting language difficulties (a characteristic of language disorder) as opposed to children whose language skills recover spontaneously. However, given that the language outcome did not include clinician-reported diagnosis of language disorder, it is possible we misclassified some children with language disorder whose language skills had improved by 11–12 years because of language therapy. Another limitation of our language outcome here was that LSAC CheckPoint only administered the CELF-4 Recalling Sentences subtest, compared with Gasparini, Shepherd, Bavin, et al. (2023) where the entire CELF-4 was administered. As noted earlier, sentence repetition has been found to have high, but not perfect, agreement with broader language skills (Botting et al., 2001; Klem et al., 2015; Smith et al., 2019). Thus, is it possible we misclassified some children if their sentence repetition skills are strong, but broader language skills are a relative weakness, or vice versa. Yet our results do converge with Gasparini, Shepherd, Bavin, et al. (2023), whose outcome measure was not the Recalling Sentences subtest, but the overall CELF-4 core language score, a known, robust assessment for language disorder.

The sensitivity analyses using alternative outcome measures did not achieve satisfactory results. This is likely due to the language assessment instruments capturing different underlying constructs: sentence repetition, requiring receptive and expressive skills (CELF-4 Recalling Sentences subtest), teacher-reported academic language and literacy skills (ARS) and grammaticality judgement, requiring grammatical competence and meta-linguistic awareness (GJT). We consider the CELF-4 Recalling Sentences subtest to be the most appropriate as the main language outcome, as it has high agreement (AUC = 0.96) with the overall CELF-4 core language score (Smith et al., 2019), a validated and widely used diagnostic instrument for language disorder (Pearson Education, 2008), along with converging evidence that sentence repetition strongly agrees with broad language skills (Botting et al., 2001; Klem et al., 2015). The alternative outcome measures ARS and GJT have not been validated for the purposes of identifying language disorder. The differences in results between outcome measures may also be because the predictors lose accuracy throughout adolescence. However, because LSAC did not collect the CELF-4 Recalling Sentences subtest repeatedly between ages, we cannot demonstrate whether this is the case and whether these results are robust to different outcome ages or larger sample sizes.

Running random forests with only complete cases failed to converge as very few participants had complete information for all the selected variables across LSAC waves 1 and 2 and the language outcome. It is possible that the variable importance measures yielded by the random forests were influenced by imputed values rather than real patterns in the data. However, we expect that the effect of this on the results is minor for two core reasons. Firstly, the imputations were repeated 100 times to minimize the effects of individual imputed values on the results. Secondly, the next part of the analysis involved using SuperLearner with complete cases only, so imputed values could not influence the final sensitivity and specificity estimates.

As in Gasparini, Shepherd, Bavin, et al. (2023), the sensitivity estimates have rather wide 95% CIs due to the low population prevalence of language disorder. A larger sample size, for instance by combining data across cohort studies, would yield more precise sensitivity estimates.

Future directions

A future study should evaluate how well this predictor set performs when collected as a standalone predictor set, rather than amongst hundreds of other questions, as in LSAC and ELVS. This study should also record how long it takes parents to complete the six questions and could collect researchers' and parents' experiences administering and completing the predictors respectively.

Future studies could assess how well the predictor set generalizes to populations underrepresented by the current cohort, namely those who experience more disadvantage or who are ethnically minoritized. While we opted for a population-wide approach in identifying predictors, future studies could identify precise predictors in subgroups where language outcomes remain difficult to anticipate, for example, multilingual or Autistic children. Future studies could also examine whether the predictor set would yield similar sensitivity and specificity in populations outside of Australia or translated into other languages.

Parent-reported measures are useful for predicting language outcomes because they currently yield comparable accuracy to assessment of language and communication skills by a trained examiner in the early years (Feltner et al., 2024), are inexpensive and low burden. However, 70–80% accuracy still misclassifies a substantial number of children. Future research could investigate whether greater accuracy can be achieved when combining parent-reported with other measures. For example, we will next add polygenic scores to the predictor set (Gasparini, Shepherd, Lange, et al., 2023).

Future studies examining predictors of persisting language difficulties should collect a language outcome that has been validated to show high accuracy for identifying children with language disorder. The CELF-4 core language score has been validated for this purpose (Pearson Education, 2008), and has also been updated with the CELF-5 (Wiig et al., 2013). When very quick assessments are needed, the CELF-4 Recalling Sentences subtest has shown promise as a proxy for the CELF-4 core language score in an agreement analysis using ELVS data (Smith et al., 2019, with converging evidence from Botting et al., 2001; Klem et al., 2015 using different instruments), but replication of this result would further justify its use for this purpose. When possible, studies should collect repeated measures of the language outcome through middle and late childhood and triangulate direct assessments with clinician-reported diagnosis of language disorder and history of language intervention, to account for children whose language skills have improved because of intervention. Future studies could also use the methodology used in Gasparini, Shepherd, Bavin, et al. (2023) and the current study to identify the best predictors of alternative outcomes, such as literacy, or subcategories of language like grammatical competence.

Practical implications

We do not consider our predictor set to be imminently suitable for screening purposes. In addition to requiring greater accuracy and precision, more robust evidence on effective early language intervention is needed before considering whether population-wide detection of children likely to have language disorder is desirable (Feltner et al., 2024). To that end, our predictor set could enable researchers to recruit children who are likelier to have language disorder into early language intervention trials. Currently, a standard approach of recruiting late-talking children has 45% sensitivity for predicting 4-year outcomes (Reilly et al., 2014). If used to recruit children into early intervention trials, our predictor set would roughly halve the number of recruited false positives who have no need for language intervention, compared with recruiting late-talking children (even when considering the lower limit of our 95% CIs, our predictor set has 13% higher sensitivity for predicting 11–12-year language than late-talking for predicting 4-year language). This would increase the power of trials to detect the extent to which an intervention uniquely contributes to improving language outcomes, rather than the majority of recruited children's language catching up spontaneously regardless of the intervention (Wake et al., 2011).

Our next steps are to add polygenic scores to the predictor set (Gasparini, Shepherd, Lange, et al., 2023), and to develop a digital interface that could estimate probability of persisting language difficulties from the parent-reported responses (similar to West et al.'s, 2021, 'LanguageScreen' for older children). Users could opt to prioritize sensitivity over specificity or vice-versa according to their needs. Such a tool could help recruit children into early intervention trials, eventually leading to improved clinical service provision where language difficulties are likely to persist.

CONCLUSIONS

We identified a set of six parent-reported variables that can be asked when children are aged 2–3 years with 'fair' accuracy (≥ 71% sensitivity and specificity) for predicting which children will have low language skills at 11–12 years. We estimate parents could answer these six questions in under 1 min. In one of few replication studies of this scale in the language development field, these results have now been replicated across two Australian population-based cohort studies using modern machine learning and ensemble analysis methods. This predictor set is not currently suitable for clinical screening purposes, given it will misclassify about one quarter to one third of children, and there are no current intervention pathways to warrant population-wide screening for language difficulties at 2–3 years. This predictor set can, however, be translated into a time- and resource-efficient tool for researchers to identify young children who are likely to have persisting language difficulties throughout childhood, for example, to recruit them into early intervention trials. This will help to strengthen the evidence base on the nature of language disorders in the early years, and how to best support these children to thrive.

ACKNOWLEDGEMENTS

L.G. was supported by an Australian Government Research Training Program (RTP) Scholarship and MCRI PhD Top Up Scholarship. A.M. was supported by an NHMRC Investigator grant (number 1195955). Research at the Murdoch Children's Research Institute (MCRI) was supported by the Victorian Government's Operational Infrastructure Support Program. The funding organizations are independent of all researchers and were not involved in any of the study design, the collection, analysis, and interpretation of data, the writing of the report or the decision to submit the manuscript for publication.

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 Sentence were developed with support from the Centre of Research Excellence in Child Language (Australian National Health & Medical Research Council, grant number 1023493). We thank the LSAC and CheckPoint study participants, staff and students for their contributions. For insightful contributions thanks to Varnika Aggarwal, Emma Baker, Charlotte Boulton, Sarah Horton, Jenny Hua, Katherine Lange, Fatma Lelik, Katie McBain, Jatender Mohal, Hannah Morgan, Lottie Morison, Hayden Naar, Olivia Van Reyk, the University of Potsdam First Language Acquisition group, the University of Groningen Neurolinguistics and Language Development research group, attendees of the Society for Research in Child Development (SRCD) Biennial Meeting 2023 and three anonymous reviewers.

Open access publishing facilitated by Murdoch University, as part of the Wiley - Murdoch University agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data necessary to reproduce the analyses presented here can be made accessible via the Australian Data Archive, see https://growingupinaustralia.gov.au/data-and-documentation/accessing-lsac-data.

REGISTRY

The hypotheses and analyses presented here were preregistered with Open Science Framework (OSF) on 27/07/2022 at https://osf.io/jk32c/registrations and all amendments were published and timestamped at https://osf.io/jk32c/. Supporting information including the analytic code necessary to reproduce the analyses presented in this paper are publicly accessible at https://osf.io/jk32c/.

ORCID

Loretta Gasparini https://orcid.org/0000-0002-1561-5572

REFERENCES

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. *Journal of Speech, Language, and Hearing Research*, 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. *Journal of Child Psychology and Psychiatry*, 58(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

Borovsky, A., Thal, D. & Leonard, L.B. (2021) Moving towards accurate and early prediction of language delay with network science and machine learning approaches. *Scientific Reports*, 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). *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(6), 741–748. Cambridge Core. https://doi.org/10.1017/S0021963001007600

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

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. *Journal of Paediatrics and Child Health*, 16149. https://doi.org/10.1111/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

Conti-Ramsden, G. & Durkin, K. (2012) Postschool Educational and Employment Experiences of Young People With Specific Language Impairment. *Language, Speech, and Hearing Services in Schools*, 43(4), 507–520. https://doi.org/10.1044/0161-1461(2012/11-0067)

Council on Children With Disabilities, Section on Developmental Behavioral Pediatrics, Bright Futures Steering Committee, & Medical Home Initiatives for Children with Special Needs Project Advisory Committee. (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

Donolato, E., Toffalini, E., Rogde, K., Nordahl-Hansen, A., Lervåg, A., Norbury, C. & Melby-Lervåg, M. (2023) Oral language interventions can improve language outcomes in children with neurodevelopmental disorders: a systematic review and meta-analysis. *Campbell Systematic Reviews*, 19(4), e1368. https://doi.org/10.1002/cl2.1368

Eadie, P., Conway, L., Hallenstein, B., Mensah, F., McKean, C. & Reilly, S. (2018) Quality of life in children with developmental language disorder: quality of life in children with DLD. *International Journal of Language & Communication Disorders*, 53(4), 799–810. https://doi.org/10.1111/1460-6984.12385

Edwards, B. (2014) Growing Up in Australia: the Longitudinal Study of Australian Children: entering adolescence and becoming a young adult. *Family 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

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. *Journal of Child Psychology & Psychiatry*, 64(8), 1242–1252. https://doi.org/10.1111/jcpp.13733

Gasparini, L., Shepherd, D.A., Lange, K., Wang, J., Verhoef, E., Bavin, E.L., Reilly, S., St Pourcain, B., Wake, M. & Morgan, A.T. (2023) *Combining genetic and behavioral predictors of 11-year language outcome: A multi-cohort study* [Preregistration]. Open Science Framework. https://osf.io/mrxdg/

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. *Developmental Science*, 18(1), 146–154. https://doi.org/10.1111/desc.12202

Law, J., Rush, R., Anandan, C., Cox, M. & Wood, R. (2012) Predicting language change between 3 and 5 years and its implications for early identification. *Pediatrics*, 130(1), e132–e137. https://doi.org/10.1542/peds.2011-1673

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

Liaw, A. & Wiener, M. (2002) Classification and regression by RandomForest. *R News*, 2(3), 18–22.

McGregor, K.K. (2020) How we fail children with developmental language disorder. *Language, Speech, and Hearing Services in Schools*, 51(4), 981–992. https://doi.org/10.1044/2020_LSHSS-20-00003

McKean, C., Mensah, F.K., Eadie, P., Bavin, E.L., Bretherton, L., Cini, E. & Reilly, S. (2015) Levers for language growth: characteristics and predictors of language trajectories between 4 and 7 years. *PLoS ONE*, 10(8), e0134251. https://doi.org/10.1371/journal.pone.0134251

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. *Journal of Child Psychology and Psychiatry*, 58(10), 1081–1091. https://doi.org/10.1111/jcpp.12790

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. *Journal of Child Psychology and Psychiatry*, 57(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

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

R Core Team. (2020) *R: A language and environment for statistical computing* (3.6.3) [Computer software]. R Foundation for Statistical Computing. https://www.Rproject.org/

Reilly, S., McKean, C. & Levickis, P. (2014) *Late talking: Can it predict later language difficulties?* (Research Snapshot 2; pp. 1–2). Centre for Research Excellence in Child Language. https://www.mcri.edu.au/sites/default/files/media/documents/crec_rs2_late-talkers-1_design_v0.1_0.pdf

Rice, M.L., Hoffman, L. & Wexler, K. (2009) Judgments of omitted BE and DO in questions as extended finiteness clinical markers of specific language impairment (SLI) to 15 years: a study of growth and asymptote. *Journal of Speech, Language, and Hearing Research*, 52(6), 1417–1433. https://doi.org/10.1044/1092-4388(2009/08-0171)

Rock, D.A. & Pollack, J.M. (2002) *Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS-K): Psychometric Report for Kindergarten through First Grade. Working Paper Series*. (NCES-WP-2002-05; p. 199). National Center for Education Statistics (ED). https://files.eric.ed.gov/fulltext/ED470320.pdf

RStudio Team. (2020) *RStudio: Integrated development environment for R (1.3.959)* [Computer software]. RStudio, PBC. http://www.rstudio.com/

Rudolph, J.M. & Leonard, L.B. (2016) Early Language Milestones and Specific Language Impairment. *Journal of Early Intervention*, 38(1), 41–58. https://doi.org/10.1177/1053815116633861

Sansavini, A., Favilla, M.E., Guasti, M.T., Marini, A., Millepiedi, S., Di Martino, M.V., Vecchi, S., Battajon, N., Bertolo, L., Capirci, O., Carretti, B., Colatei, M.P., Frioni, C., Marotta, L., Massa, S., Michelazzo, L., Pecini, C., Piazzalunga, S., Pieretti, M., … & Lorusso, M.L. (2021) Developmental language disorder: early predictors, age for the diagnosis, and diagnostic tools. A scoping review. *Brain Sciences*, 11(5), 654. https://doi.org/10.3390/brainsci11050654

Semel, E., Wiig, E.H. & Secord, W.A. (2006) Clinical evaluation of language fundamentals—fourth edition*,* Australian Standardised Edition, 4th edition. Marrickville, Australia: 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

Stott, C.M., Merricks, M.J., Bolton, P.F. & Goodyer, I.M. (2002) Screening for Speech and Language Disorders: the reliability, validity and accuracy of the General Language Screen. *International Journal of Language & Communication Disorders*, 37(2), 133–151. https://doi.org/10.1080/13682820110116785

Strobl, C., Hothorn, T. & Zeileis, A. (2009) Party on! A new, conditional variable-importance measure for random forests available in the party package. *The R Journal*, 1(2), 14–17.

Tomblin, J.B., Smith, E. & Zhang, X. (1997) Epidemiology of specific language impairment: prenatal and perinatal risk factors. *Journal of Communication Disorders*, 30(4), 325–344. https://doi.org/10.1016/S0021-9924(97)00015-4

van der Laan, M.J., Polley, E.C. & Hubbard, A.E. (2007) Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), Article25. https://doi.org/10.2202/1544-6115.1309

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. *The Lancet*, 370(9596), 1453–1457. https://doi.org/10.1016/S0140-6736(07)61602-X

Wake, M., Tobin, S., Girolametto, L., Ukoumunne, O.C., Gold, L., Levickis, P., Sheehan, J., Goldfeld, S. & Reilly, S. (2011) Outcomes of population based language promotion for slow to talk toddlers at ages 2 and 3 years: let's Learn Language cluster randomised controlled trial. *BMJ*, 343(Aug 18 2), d4741–d4741. https://doi.org/10.1136/bmj.d4741

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

West, G., Snowling, M.J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H. & Hulme, C. (2021) Early language screening and intervention can be delivered successfully at scale: evidence from a cluster randomized controlled trial. *Journal of Child Psychology and Psychiatry*, 62(12), 1425–1434. https://doi.org/10.1111/jcpp.13415

Wiig, E.H., Semel, E. & Secord, W.A. (2013) *Clinical evaluation of language fundamentals–fifth edition (CELF-5)*. NCS Pearson. https://www.pearsonclinical.com.au/store/auassessments/en/Store/Professional-Assessments/Speech-%26-Language/Clinical-Evaluation-of-Language-Fundamentals-Australian-and-New-Zealand-Fifth-Edition/p/P100010122.html?tab=product-details

Wilson, P., Rush, R., Charlton, J., Gilroy, V., McKean, C. & Law, J. (2022) Universal language development screening: comparative performance of two questionnaires. *BMJ Paediatrics Open*, 6(1), e001324. https://doi.org/10.1136/bmjpo-2021-001324

Yew, S.G.K. & O'Kearney, R. (2013) Emotional and behavioural outcomes later in childhood and adolescence for children with specific language impairments: meta-analyses of controlled prospective studies: SLI and emotional and behavioural disorders. *Journal of Child Psychology and Psychiatry*, 54(5), 516–524. https://doi.org/10.1111/jcpp.12009

Zambrana, I.M., Pons, F., Eadie, P. & Ystrom, E. (2014) Trajectories of language delay from age 3 to 5: persistence, recovery and late onset. *International Journal of Language & Communication Disorders*, 49(3), 304–316. https://doi.org/10.1111/1460-6984.12073

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 & Developmental Language Impairments*, 7, 239694152210993. https://doi.org/10.1177/23969415221099397

SUPPORTING INFORMATION

All supporting information are available at https://osf.io/jk32c/. This paper's written plain language summary, visual abstract, and video plain language summary are all available in the "Plain language summaries" subdirectory therein.

How to cite this article: Gasparini, L., Shepherd, D.A., Wang, J., Wake, M. & Morgan, A.T. (2024) Identifying early language predictors: A replication of Gasparini et al. (2023) confirming applicability in a general population cohort. *International Journal of Language & Communication Disorders*, 1–15. https://doi.org/10.1111/1460-6984.13086