JSLHR

Research Article

Does the Inclusion of a Genome-Wide Polygenic Score Improve Early Risk Prediction for Later Language and Literacy Delay?

Philip S. Dale,[a] Sophie von Stumm,[b] Saskia Selzam,[c] and Marianna E. Hayiou-Thomas[d]

Purpose: The ability to identify children early in development who are at substantial risk for language/literacy difficulties would have great benefit both for the children and for the educational and therapeutic institutions that serve them. Information that is relatively easily available prior to the age of 3 years, such as late talking, family history of language/literacy difficulties, and socioeconomic status, have some but very limited predictive power. Here, we examine whether the inclusion of a DNA-based genome-wide polygenic score that has been shown to capture children's genetic propensity for educational attainment (EA3) adds enough prediction to yield a clinically useful score.
Method: Data are longitudinal scores of 1,420 children from the Twins Early Development Study, who were assessed at ages 2 and 3 years on language and nonverbal ability and at 12 years of age on oral language, word decoding, and reading comprehension. Five risk factors were examined: expressive vocabulary, nonverbal ability (these two from parent report), family history, mothers' education, and EA3. Analyses were conducted both for continuous and categorically defined measures of risk and outcome.
Results: Language and literacy abilities at 12 years of age were significantly but modestly predicted by the risk factors, with a small but significant added prediction from EA3. Indices of diagnostic validity for poor outcomes, such as sensitivity and area under the curve statistics, were poor in all cases.
Conclusions: We conclude that, at present, clinically useful prediction from toddlerhood remains an unattained goal.
Supplemental Material: https://doi.org/10.23641/asha.12170331

Impairments in the mastery and use of language and literacy skills have substantial and far-reaching impact on the lives of children and adults, extending beyond education to social relationships, mental health, and employment (Law et al., 2009; Weismer, 2014). On the plausible assumption (though difficult to rigorously evaluate) that intervention is more likely to be effective early than later, a high priority in recent research has been to develop measures that can accurately identify young children who are at substantial risk for language and literacy difficulties. Such identification would have great benefit both for the children and for the educational and therapeutic institutions that serve them by enabling the provision of services to the children who need them most and as early as possible. On the whole, research has had only very limited success in this pursuit. One particularly important challenge is the great variability in individual growth trajectories in the first decade of life (Dale et al., 2014). This variability necessitates long-term longitudinal studies, which explore prediction at least out to early adolescence when abilities become relatively stable (Hayiou-Thomas et al., 2014).

An extensive body of research has documented prediction from the late preschool period to later language and literacy. A range of language abilities, including vocabulary, grammar, and speech, accounts for some of the variance, but the most strongly supported risk markers from this period are in the domain of emergent literacy, including phonological awareness and memory, letter knowledge,

[a]Department of Speech & Hearing Sciences, University of New Mexico, Albuquerque
[b]Department of Education, University of York, United Kingdom
[c]King's College London, United Kingdom
[d]Department of Psychology, University of York, United Kingdom
Correspondence to Philip S. Dale: dalep@unm.edu

Editor-in-Chief: Sean M. Redmond
Editor: Mabel L. Rice
Received August 22, 2019
Revision received October 25, 2019
Accepted January 27, 2020
https://doi.org/10.1044/2020_JSLHR-19-00161

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

and rapid automatized naming (Duff & Snowling, 2015; Marinus & Castles, 2015). A family history of early language or literacy learning difficulties also contributes to prediction (Snowling & Melby-Lervåg, 2016), along with some nonverbal abilities. Though not a child-based measure, socioeconomic status (most often assessed during early childhood as maternal education) also has predictive significance (Hoff, 2006). Although each of these risk factors contributes to prediction, their combined prediction is far from being sufficiently sensitive at an individual level (e.g., sensitivity, specificity) to be clinically useful. For example, although Hayiou-Thomas et al. (2020) found substantial odds ratios for low performance at 12 years of age based on the number of risk factors at 4.5 years of age, compared to none, in the range of 3.65–17.71 (depending on whether there were many risk factors compared to only a few), showing a strong group prediction; the positive predictive values (PPVs) for later delay for multiple risk factors for individual children were in the range of .33–.55, which is poor.

The most widely studied potential marker in toddlerhood is late talking: late emergence and/or late growth of early vocabulary and the emergence of grammar (Rescorla & Dale, 2013). Although late talking (e.g., fewer than 50 words at 24 months or lowest 10%) is associated with a higher rate of later language impairment, there is enormous variability. Half or more of all late talkers will move into the normal range by the end of the preschool period, though they may still have lower scores than matched controls (Rescorla, 2013). Dollaghan (2013) has provided a particularly comprehensive review and integration of 14 studies which followed up late talkers and demonstrated how poor the prediction is at an individual level. Notably, the pooled positive and negative likelihood ratios were just 2.04 and 0.66, respectively, not only far from the 10 and 0.1 values often recommended as being clinical trustworthy but not even reaching the criteria for a "suggestive indication" of 3 and 0.3. Some of the variables mentioned above cannot be measured below the age of 3 years as they have not yet emerged, but the risk factors of family history and nonverbal abilities are potentially available at an early age, as is maternal education. These factors have seldom been investigated together, but the available evidence (e.g., Dale et al., 2003) suggests that the prediction is poor.

Genome-Wide Polygenic Scores and EA3

Although emergent literacy skills are not generally testable in toddlers, they are known to have significant genetic influence (Samuelsson et al., 2007). This suggests that a DNA-based measure might be valuable. In this study, we examine whether inclusion of such a measure adds enough prediction to yield a clinically useful score in predicting language and literacy performance at 12 years of age, both across the full range of performance and specifically with respect to placement in a low-performance category.

A relatively recent new technique for evaluating genetic influence on a trait is centered on the use of genome-wide polygenic scores (GPSs). A fundamental insight underlying this approach is the realization that, although many academic, cognitive, and language skills are substantially heritable (30%–60% of the variance), this is the result of a very large number of genetic factors, each with very small influence. Even the largest associations account for less than 0.05% of the variance (Allegrini et al., 2019). An especially clear example of this is shown in a recent study, which focused on six genes known from pedigree and/or molecular studies to be associated with language/literacy development and disorders: ATP2C2, CMIP, CNTNAP2, DCDC2, FOXP2, and KIAA0319, which together included a total of 1,229 single-nucleotide polymorphisms (SNPs). Even when summed, these genetic variants accounted for only 0.18% of the variance in a measure of expressive language at the age of 8 years (Newbury et al., 2019). Thus, rather than searching for a few DNA units with a large effect, current research is directed toward identification of a much larger number of elements in genome-wide association (GWA) studies. This has become feasible due to the construction of "SNP chips," arrays that are able to genotype (identify) inexpensively and quickly the most common type of inherited DNA differences, a difference in a single-nucleotide base, called an SNP. Such a scan typically produces around a million units of information for an individual's DNA. Each of these units can be correlated with the trait of interest, but because of the extremely high number of such tests and the low actual impact of any unit, the results for individual units are not reliable or directly interpretable. Instead, this information is used to select a set of elements with the highest correlation; typically, hundreds or even thousands of SNPs are selected and then summed. It is acknowledged that SNPs may be included in the sum which are not in fact related, and that relevant SNPs may be missed; it is the sum that is important. To demonstrate the reliability and validity of the sum, the GPS, the formula is first computed using a "training set" of observations and then validated by determining how well it predicts the trait in the "replication sample." Once the GPS is confirmed in this way, the formula can be applied in new research studies. An important caveat for this approach is that it requires extremely large studies, both in number of subjects and number of DNA variants studied. This can be thought of as a problem of distinguishing signal from noise.

By far, the most successful GPS has been defined on the basis of the educational attainment of adults. This is due in part to the fact that educational attainment, usually indexed as total number of years of education, is obtained as a demographic marker in most GWA studies. Based on pooled data from many such studies, totaling 1.1 million individuals, a GPS for educational achievement accounted for 12% of the variance (Lee et al., 2018). This GPS, named EA3 because it is the third version of the formula, also predicts educational achievement during the period of schooling. Within the Twins Early Development Study (TEDS), from which the present sample is drawn, the proportion of variance in educational achievement predicted by EA3 increases from 5% at 7 years of age to 7% at 12 years of

age (both of these are with respect to national curriculum school grades) to 14% at 16 years of age for the end of compulsory education General Certificate of Secondary Education examination (Allegrini et al., 2019; von Stumm et al., in press). Within the same sample, Ayorech et al. (2019) utilized multinomial logistic regression to predict educational/occupational status at 18 years of age in terms of four broad categories: pursuing a university degree; taking a gap year; pursuing full-time employment or apprenticeship; or not being in education, employment, or training. The prediction of category placement was estimated with an effect size of Nagelkerke r^2 of 10%. Utilizing a previous version of the GPS known as EA2, Smith-Woolley et al. (2018) obtained significant prediction of performance for TEDS participants on the university entrance examination (A levels; 4% of the variance), university enrollment (dichotomous; 5% of the variance), university quality rating (U.K. league tables; 2%), and university achievement (possession and type of degree, e.g., first class; 0.7%). The last of these effect sizes is the lowest; this is likely because of restricted range for students attending university and also because both earning a degree and its type are not standardized across universities and thus not on the same scale.

There are two notable aspects of the success of EA3. The first is the evidence it provides of the need for very large samples. Earlier versions, EA1 and EA2, which were based on 125,000 and 294,000 subjects, respectively, accounted for only 2% and 3% of the variance in educational attainment (Allegrini et al., 2019). An additional 800,000 subjects were needed to improve the prediction from 3% to 12%. The second one is foreshadowed by the fact that EA3 actually predicts educational achievement at the age of 16 years even more successfully than it predicts total educational attainment in adults, the outcome that was used to develop the formula. Because there is high pleiotropy for genetic influence, that is, the effect of genetic units is widespread and not limited to a single trait, prediction of traits other than the target one may be very good. The influence of EA3 is considerably wider than this; in particular, it has been shown to predict (again, in the TEDS sample) 9.9% of the variance in IQ at 16 years of age (Allegrini et al., 2019). GPS formulas do exist for other traits, such as reading and IQ, but they are based on far smaller sample sizes in the training set, generally below 10,000. Due to the overwhelming importance of sample size in defining the GPS, along with pleiotropic effects, cross-trait prediction can be even higher than within-trait prediction. For example, EA3 actually predicts IQ at 16 years of age better (9.9%) than an IQ-based GPS does (6.7%; Allegrini et al., 2019). Given the broad success of EA3, its clear cognitive connection to language and literacy, and the small size of GPS training sets for language and literacy variables, we have chosen to examine the predictive contribution of EA3.

The Current Study

In this article, we evaluate the prediction from five risk factors discussed above, potentially available in the toddler period: family history, maternal education, expressive vocabulary, nonverbal abilities, and EA3 (the GPS for educational attainment). We utilize two analytic approaches, one focused on the full range of predictor and outcome measures and the other focused on a cumulative risk measure based on low scores on each measure. We address the following research questions:

1. Using continuous measures representing the full distribution of risk and outcome measures, how well can receptive language, word decoding, and reading comprehension at 12 years of age be predicted by early expressive vocabulary and nonverbal ability, family history of early language/literacy difficulties, mothers' education, and a GPS for educational attainment (EA3)?
   (a) Specifically, does EA3 make an additional, unique contribution to the prediction?
   (b) Conversely, do the other nongenetic measures make a contribution to the prediction beyond that of EA3?
2. Using categorically defined (lowest 10%; a frequent criterion in clinical work and research on disorders) measures for each risk and outcome measure, how well can impairments at 12 years of age for receptive language, word decoding, and reading comprehension be predicted by categorically defined risk factors of early expressive vocabulary and nonverbal ability, family history of early language/literacy difficulties, mothers' education, and a GPS for educational attainment (EA3)?
   (a) Specifically, does EA3 make an additional, unique contribution to the prediction?
   (b) Conversely, do the other nongenetic measures make a contribution to the prediction beyond that of EA3?
3. For both questions above, is prediction from the risk factors assessed at 3 years of age superior to that from 2 years of age? Note that only expressive vocabulary and nonverbal ability change from 2 to 3 years of age.

Method

Ethical approval for TEDS has been provided by the King's College London Ethics Committee (reference: PNM/09/10-104).

Participants

The present sample is drawn from TEDS, a longitudinal twin study that recruited more than 10,000 twin pairs born in England and Wales in 1994 through 1996. The recruitment process and the sample are described in detail elsewhere (Haworth et al., 2013; Rimfeld et al., 2019). The TEDS sample was at its inception representative of the U.K. population in comparison with census data and

despite some attrition remained considerably representative at 12 years of age (and later) with respect to ethnicity, maternal education and employment, and paternal employment (Haworth et al., 2013; Oliver & Plomin, 2007). For 1,812 individuals of European ancestry from TEDS, complete risk measures were available, including genotype data (both twins from DZ [i.e., paternal] pairs were included, as they differ genetically, but only one twin from each MZ [i.e., identical] pair was included, chosen randomly) and information on expressive vocabulary and nonverbal ability at ages 2 and 3 years, mother's educational qualifications, and family history concerning early language/literacy difficulties. For 1,420 of these individuals, outcome measures for at least one of oral language, word decoding, and reading comprehension at 12 years of age were available. The sample included 789 females and 631 males; etiologically, 568 twins were from MZ pairs, 429 from DZ same-sex pairs, and 423 from DZ opposite-sex pairs. Based on information provided at the time of recruitment, the present analysis sample is slightly more advantaged ($d = 0.22$) in terms of mother's educational qualifications ($M = 4.06$, $SD = 2.00$) than the full TEDS sample ($M = 3.63$, $SD = 1.96$).

Risk Factor Measures

Parent-Reported Expressive Vocabulary

The parent-reported expressive vocabulary measures at ages 2 and 3 years were shortform adaptations of the vocabulary portion of the MacArthur–Bates Communicative Development Inventories (MBCDI): Words and Sentences (Fenson et al., 2007). It is composed of a list of words from which parents are asked to check those that they have heard their child say. The score is the total number of words checked. The 100 words on the 2-year version were selected (Fenson et al., 2000) to give good prediction of the total score from the longer list on the MBCDI. The list of words was then "anglicized" for appropriate spelling in a U.K. setting. The 3-year vocabulary measure was developed in accordance with similar design principles, replacing about half the words in order to be more appropriate for older children based on literature review and pilot testing. Additional information about these measures is provided in the work of Dale et al. (2003).

Parent-Reported Nonverbal Ability

The Parent Report of Children's Abilities (PARCA; Oliver et al., 2002; Saudino et al., 1998) was developed within the TEDS project. It consists of both parent report questions and parent-administered items (e.g., copying shapes, imitating actions). Because the aim of this study was to include measures that were relatively easily available, we included only information provided in the parent report portion of the instrument. At 2 years of age, the PARCA consists of 26 questions assessing quantitative skills, spatial abilities, symbolic play, planning and organizing, adaptive behaviors, and memory. At 3 years of age, there are 22 questions assessing similar domains, extending also to reasoning and number. Items included modified items from existing parent questionnaires as well as original items written specifically for the PARCA. Questions were phrased in terms of specific activities, and parents were asked to indicate if they had seen their child perform the activity, as research indicates that parents are most accurate in this recognition format. A sample question at 2 years of age is "Can your child put a simple piece, such as a square or an animal, into the correct place on a puzzle board"; an example at 3 years of age is "Does your child draw simple pictures that other people can recognize, such as a person, house, or car?"

Mother's Educational Qualifications

Mothers' highest educational achievement was rated on an 8-point scale, as in the study of Dale et al. (2003) and other TEDS publications. On this scale, 1 = *no qualifications*, 2 = *below standard for a pass on the school leaving examination*, 3 = *O levels* (passing score on school leaving examination), 4 = *A levels* (exam at 18 years of age, generally required for university entrance), 5 and 6 = *tertiary vocational qualifications*, 7 = *an undergraduate degree*, and 8 = *a postgraduate degree*.

Family History

When children were 9 years of age, parents provided information about family history. If any first-degree relative was reported as having either early language or reading difficulty, we classified this as a positive family history, that is, family history = 1.

GPS EA3

Saliva and buccal cheek swab samples were collected by mail, DNA was extracted and genotyped, and molecular analysis was used for these same individuals to compute GPSs based on the latest GWA study for years of education ($N = 766{,}345$), called "educational attainment EA3," as it is the third and largest version of the study (Lee et al., 2018). (This sample is smaller than the full Lee et al. sample of 1.1 million, due to the fact that data contributed to their sample by 23andMe are not publically available.) Beta weights for the prediction of each SNP to educational attainment were drawn from EA3 and then applied to genotypes for each TEDS participant, utilizing a Bayesian approach (LDpred; Vilhjálmsson et al., 2015) that applies shrinkage to individual SNP weights to improve prediction. Additional information about this computation is provided in Supplemental Method S1 and in the study of Selzam et al. (2019).

Categorical and Cumulative Risk Measures

For categorical analyses, a criterion for dichotomous cutoff was defined as the value on the continuous measure, which would capture as close as possible to the lowest 10%. Scores below this level were scored as 1; scores above it were scored as 0. Family history was inherently categorical. In addition to the individual risk measures, cumulative risk measures at ages 2 and 3 years, ranging from 0 to 5, were computed as the total number of risk variables for which the score was 1. This measure was used to predict the continuous outcome measures at 12 years as well as low outcome at 12 years.

Outcome Measures at 12 Years

Language Composite

Receptive language was assessed at 12 years using a battery of audio-streamed, web-administered measures indexing vocabulary (Wechsler Intelligence Scale for Children–Third Edition as a Process Instrument; Kaplan et al., 1999), syntax (Listening Grammar, Test of Adolescent and Adult Language–Third Edition; Hammill et al., 1994), nonliteral-semantics, and understanding of inferences (Test of Language Competence–Level 2; Wiig et al., 1989). Previous analysis showed substantial phenotypic and genetic overlap among these four measures (Dale et al., 2010). A composite score was formed by averaging the standardized scores for each measure. (All four measures were required to be available, and a similar requirement held for the other two composites.)

Word Decoding Composite

The Test of Word Reading Efficiency, Form B (Torgesen et al., 1999) was included in a test booklet sent to families by mail (one test booklet for each twin) and administered to each twin separately by telephone. In addition, children completed an online adaptation of the Woodcock–Johnson III Reading Fluency Test (Woodcock et al., 2001). These three measures were combined into a composite by averaging the standardized scores for each measure.

Reading Comprehension Composite

Sentence-level reading comprehension was assessed using a web-based version of the Reading Comprehension subtest of the Peabody Individual Achievement Test (Markwardt, 1997), in which children read a sentence and chose the matching picture from a set of four. In addition, children completed a web version of the GOAL Formative Assessment in Literacy for Key Stage III (GOAL, 2002), which includes a wide range of literal and inferential comprehension questions based on a set of stimulus sentences and short paragraphs. These two measures were combined into a composite by averaging the standardized scores for each measure.

Analysis

Because patterns of prediction may vary between the full range of measures and performance at the low end, two sets of analysis were conducted. In the first, both predictors (other than family history, which is inherently categorical) and outcomes were treated as continuous, full-range measures in multiple regression analyses. In the second, predictors and outcomes were treated as dichotomous variables with cutoffs defining risk or low outcome. A cumulative risk score was computed across the five measures and related to outcome. Diagnostic validity was computed as sensitivity, specificity, PPV, and negative predictive value based on matching the predicted incidence of low outcome as closely as possible to the actual incidence. In addition, as a measure of diagnostic validity independent of cutoff, receiver operating characteristic (ROC) analyses were used to determine the area under the curve (AUC).

Results

Descriptive Information

Table 1 provides descriptive information for each of the risk and outcome measures for the full sample and by gender. It also includes the percentage of children who exhibited risk as defined above for each of the predictor and low performance on each of the outcome measures. With respect to predictor variables analyzed categorically, males were more likely than females to be at risk with respect to expressive vocabulary at 3 years of age and to EA3 and consequently higher cumulative risk at 2 years of age. With respect to outcome variables, the only gender difference was that males were more likely than females to be scoring low with respect to reading comprehension at 12 years. As there were few gender differences overall and they were small, gender was not treated as a distinct risk factor.

Predictive Analyses for Continuous, Full-Distribution Measures

Multiple regression predictions using all predictive and outcome measures in continuous mode (other than family history) produced modest predictions, as shown in Table 2. In the initial regressions, all five risk factors were observed to contribute significantly to the prediction of all three outcomes, except that family history did not contribute unique variance for receptive language at 12 years in either the 2- or 3-year analyses. The R^2 values for the total prediction (numeric Columns 2 and 5) are systematically higher for prediction from 3 years of age than from 2 years of age. The table also includes the unique contribution made by EA3 when entered last, and the unique combined contribution of the other four variables when entered after EA3. In each case, the variable or variables added last made a significant contribution.

Although these correlations are modest, the predictions are not trivial. This is illustrated in Figure 1. (Standard errors are not presented in this and the following figure due to lack of visual clarity that would result due to closeness of the lines. That information is provided in Supplemental Table S1.) Here, the results of the regression analysis (somewhat different for each measure) are used to compute predicted scores at outcome; those predictions are then categorized into quintiles, and the actual mean scores at outcome are computed. It can be seen that the top and bottom quintiles according to prediction are separated by approximately 0.8 of a standard deviation. On a standardized test, this would correspond to approximately 12 points.

Categorical Risk Factors and Cumulative Risk Analyses

No participant was observed to have all five risk factors at 2 years of age, and only two participants were

Table 1. Means (standard deviations) for continuous risk and outcome measures for total sample and by sex, and percentage meeting criteria for low score for each measure.

Measure	Total sample		Females		Males	
	M (*SD*)	% low	*M* (*SD*)	% low	*M* (*SD*)	% low
Vocabulary-2	49.7 (24.7)	10.5	52.7 (24.3)	7.4	**45.9 (24.7)**	**14.4**
					< .001	< .001
Vocabulary-3	65.2 (23.9)	10.0	66.1 (23.2)	8.7	64.0 (24.6)	11.6
Nonverbal cognition-2	18.4 (3.2)	11.8	18.6 (3.2)	10.1	**18.1 (3.1)**	**13.9**
					.001	.027
Nonverbal cognition-3	14.1 (3.0)	12.2	14.4 (3.0)	9.5	**13.7 (3.1)**	**15.5**
					< .001	< .001
Mother's education	4.1 (2.0)	16.8	4.1 (2.0)	17.9	4.07 (2.0)	15.4
Family history		20.5		19.8		21.4
EA3[a]	0.091 (1.025)	10.0	0.053 (0.997)	9.5	0.138 (1.058)	10.6
Cumulative risk-2	0.696 (0.85)	16.4	0.646 (0.83)	15.0	**0.758 (0.87)**	18.1
					.015	
Cumulative risk-3	0.694 (0.87)	16.4	0.654 (0.84)	16.0	0.745 (0.90)	17.3
Receptive language-12[a]	0.014 (0.73)	9.9	0.031 (0.75)	10.0	–0.008 (0.72)	9.9
Decoding-12[a]	0.002 (0.86)	10.0	0.021 (0.84)	9.5	–0.024 (0.88)	10.6
Reading comprehension-12[a]	0.009 (0.89)	9.9	–0.011 (0.87)	9.4	0.033 (0.91)	10.6

Note. Scores for gender scoring significantly higher risk score (lower vocabulary, nonverbal cognition, or mother's education; higher family history) or higher outcome are bolded, and *p* value is included underneath.

[a]*z* score; *M* may not equal 0 and *SD* may not equal 1, as standardization is on a larger sample. Similarly, % low may not equal 10%.

observed to have all five risk factors at 3 years of age, and therefore, that value for cumulative risk was omitted from the figures that follow. As shown in Supplemental Figure S1, approximately half of the children had no risk factors, one third had one, 13% had two, 2.7% had three, and 0.8% had four. The distributions of risk factors at ages 2 and 3 years are very similar, as expected, given that only vocabulary and cognition can change between the two ages. The two cumulative risk measures were correlated at $r = .806$ ($\kappa = .63$ if risk total is considered as a categorical measure). Correlations among the five individual risk factors are provided in Supplemental Table S2. Most pairwise phi coefficients were significant; low EA3 was the exception, in that it was not related to any child variable at either age, only to low maternal education.

The relation between each of the five risk factors and low performance on each outcome measure, as indexed by phi coefficients (Pearson coefficients for binary variables), was significant, though modest, in every case other than nonverbal delay predicting reading comprehension, as shown in Supplemental Figure S2. In general, receptive language at 12 years of age is more predictable (measured by phi) than word decoding or reading comprehension, though the differences are small.

The relation of cumulative risk at ages 2 and 3 years to low performance on each of the outcome measures is illustrated in Figures 2a and 2b. The relation is generally monotonic and is approximately equal for all three outcomes. Spearman correlations between number of risk conditions at ages 2 and 3 years and poor outcomes are shown in Table 3, along with comparable correlations when low EA3 is not included in the cumulative risk measure. Although the correlations are consistently higher for the measure with EA3, the differences are small and are significant only for receptive language, using Steiger's (1980) formulae. There is little evidence for the predictions from 3 years of age to be higher than from 2 years of age; the differences are inconsistent in direction, and none is significant (all *p*s > .3).

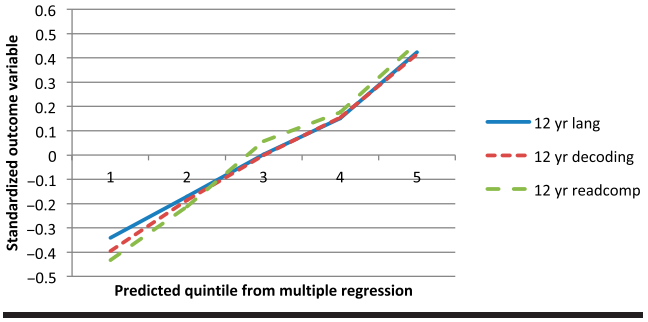
Table 2. Multiple regression analyses predicting 12-year outcomes from five continuous risk variables at ages 2 and 3 years.

Outcome measure	*n*	R^2 for prediction from 2 years of age	R^2 unique prediction from EA3	R^2 unique prediction from 4 variables[a] at 2 years of age	R^2 for prediction from 3 years of age	R^2 unique prediction from EA3	R^2 unique prediction from 4 variables[a] at 3 years of age
12-year language	1,077	.140	.023	.067	.168	.033	.069
12-year decoding	1,163	.118	.020	.057	.134	.029	.054
12-year reading comprehension	1,338	.132	.033	.049	.143	.033	.057

Note. All variance contributions are significant at $p < .01$.

[a]Family history, mother's education, expressive vocabulary, nonverbal skills; note that only expressive vocabulary and nonverbal skills can change between 2 and 3 years of age.

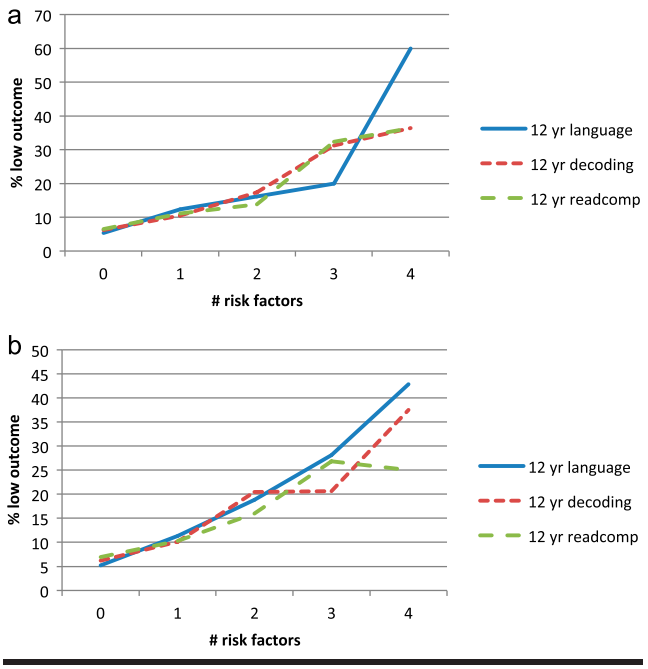
Figure 1. Mean outcomes by predicted quintile from multiple regressions.



The central analyses for Research Question 2 concern predictions for individual children. This requires a criterion for low score on the cumulative risk measure. For each outcome, approximately 10% are classified as low performers. At both ages, a cutoff of greater than or equal to 2 captures 16.5% of the sample. These are the closest it is possible to get to 10%. Table 4 summarizes the results of analyses based on this classification.

As can be seen, the odds ratios are all significantly greater than 1.0 for risk factors above the proposed cutoff of 2, but only modestly so, and the prediction is very poor with respect to sensitivity and PPV. As a more comprehensive evaluation of the validity of prediction, not tied to any specific cutoff of risk, six ROC analyses were conducted

Figure 2. (a) Risk factors at 2 years and percentage of low outcomes at 12 years. (b) Risk factors at 3 years and percentage of low outcomes at 12 years.



(2 cumulative risk predictors × 3 outcomes). These analyses assess the validity of a classification measure across all possible settings of the criterion by plotting sensitivity against specificity for each possible predictor cutoff. A sample ROC is shown in Figure 3, and all six are included as Supplemental Figures S3–S8. The AUC statistic is a measure of the extent to which that curve diverges from the diagonal, which indicates chance performance. Thus, it is a comprehensive assessment of the potential diagnostic validity and hence utility of a measure. For cumulative risk at 2 years and receptive language at 12 years, AUC = .655. AUC values for the entire set of six predictions range from .615 to .671, which would be considered poor (Caspi et al., 2016). It has been suggested that AUC values need to be approximately .8 or better to demonstrate adequate discrimination for consideration in decision making about intervention.

Discussion

For the continuous, full-distribution measures (Research Question 1), the GPS EA3 did significantly improve prediction, but only very modestly, and even with it included, prediction was poor. The other four measures contributed an even larger amount of unique prediction. For the categorical measures of risk factors and language/literacy outcomes (Research Question 2), the predictions were significantly better with EA3 included, but only for the receptive language measure. Waiting until 3 years of age (Research Question 3) slightly improved prediction of the continuous outcome measures, but not the prediction of poor outcomes. Overall, sensitivity, PPV, odds ratios, and AUC statistics all demonstrated predictions far from clinically useful.

Given the relative success of EA3 in predicting educational attainment, why are these results so poor? There are several possibilities, not mutually exclusive. First, EA3 was developed for educational attainment in the broad sense, in contrast to the focus on language and literacy outcomes in this study. Educational attainment (and success along the way) reflects a wide range of abilities and characteristics that extend beyond language and literacy. Notably, Krapohl et al. (2014) found that a substantial portion of the heritability of educational achievement was shared with genetic influence on personality (a composite including conscientiousness), behavior problems, and educational self-efficacy. The success of EA3 may reflect individual differences in these broader characteristics, and thus, language and literacy abilities and difficulties may constitute only a very small part of the story. We note that research is in progress to develop a language-specific GPS, but there is not yet a sufficient sample to yield a GPS score with good predictive validity.

Second, research documenting the predictive power of EA3 has typically examined that variable either in isolation as a predictor or without the extensive demographic and early childhood measures available here. In the present case, we hypothesize that some of the predictive power of

Table 3. Spearman correlations between number of risk factors and poor outcome.

Predictor (cumulative risk score)	Receptive language (n = 1,077)	Word decoding (n = 1,163)	Reading comprehension (n = 1,338)
2 years, including EA3	.176[a]	.163	.139
2 years, without EA3	.146[a]	.151	.134
3 years, including EA3	.194[b]	.164	.131
3 years, without EA3	.165[b]	.152	.124

Note. All correlations are significant at $p < .001$; correlations with the same superscript are significantly different (p = .005 and .006, respectively).

EA3 may be pre-empted by those other measures that are themselves predicted by EA3. As shown in Supplemental Table S2, EA3 is correlated in this sample with maternal education, consistent with the finding of Wertz et al. (2019) that children's EA3 is correlated with their mother's EA3 and actual educational attainment.

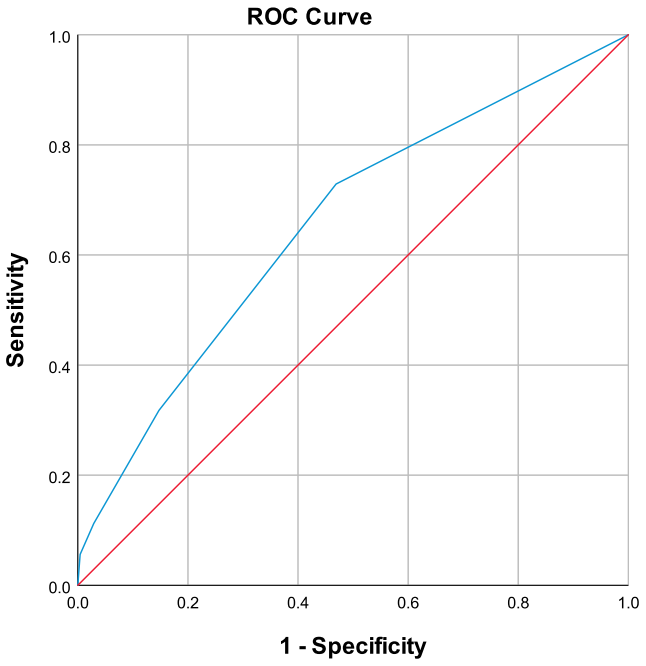
A third factor that may reduce prediction stems from the use of twins in the present research. Early language development in twins is typically slower than for singletons, though the "twin effect" is reduced or absent by the early school years (Thorpe, 2006). Likely causes of delay include prenatal and perinatal factors, which are known to impact language development and are at elevated risk in twins, including gestational diabetes, prolonged time to spontaneous respiration, and fetal growth restriction (Taylor et al., 2018), as well as twin effects on parental language input and interaction (Thorpe, 2006). Valid research on the twin effect and its developmental course requires the use of measures that have been normed in the general, that is, singleton population (Rice et al., 2018, 2020), and this is often not possible for twin studies and other large-sample designs, where, as in this study, a relative criterion of low performance within the available sample was used. This methodological difference, population based versus sample specific, has implications for categorizing children either as normally developing, late talkers or recovered to the normal range. Recently, Rice et al. (2014, 2018, 2020) have studied the twin effect in a large and representative Western Australian population with standardized measures, from the ages of 2 to 6 years. The twin effect was evident at 2 years of age and continued to be substantial at 4 years of age but had largely resolved by the age of 6 years, when considering continuous measures of language performance (Rice et al., 2018). Similarly, when using a categorical cutoff of −1 *SD* to identify language impairment, the proportion of children meeting this criterion was reduced from 36% at 4 years of age to 28% at 6 years of age, again suggesting that the twinning effect diminishes between these two ages, although the rate of language impairment was still substantially higher than the 8% expected in a general population (Rice et al., 2020). The findings of the Western Australia project taken together suggest a period of particularly dynamic and variable change in language development between the ages of 4 and 6 years that is specific to twins. Moreover, the twinning effect seems to be more marked for MZ than DZ twins, adding further

Table 4. Prediction of low performance on receptive language, word decoding, and reading comprehension from cumulative risk measure at 2 (CR2) and 3 (CR3) years of age.

Cumulative risk cutoff	Typical at 12	Low at 12	Sensitivity	Specificity	PPV	NPV	*OR* (95% CI)
Receptive language from 2 years of age							
CR2 < 2	827	74	.31	.85	.19	.92	2.58 [1.65, 4.03]
CR2 ≥ 2	143	33					
Word decoding from 2 years of age							
CR2 < 2	895	76	.34	.85	.21	.92	3.10 [2.04, 4.72]
CR2 ≥ 2	152	40					
Reading comp from 2 years of age							
CR2 < 2	1,023	93	.30	.85	.18	.92	2.42 [1.62, 3.62]
CR2 ≥ 2	182	40					
Receptive language from 3 years of age							
CR3 < 2	831	68	.36	.86	.22	.92	3.43 [2.22, 5.29]
CR3 ≥ 2	139	39					
Word decoding from 3 years of age							
CR3 < 2	894	74	.36	.85	.22	.92	3.32 [2.19, 5.03]
CR3 ≥ 2	153	42					
Reading comp from 3 years of age							
CR3 < 2	1,021	91	.32	.85	.19	.92	2.56 [1.72, 3.81]
CR3 ≥ 2	184	42					

Note. PPV = positive predictive value; NPV = negative predictive value; *OR*= odds ratio; CI = confidence interval.

Figure 3. Receiver operating characteristic (ROC) curve plotting sensitivity versus specificity for cumulative risk measure at 2 years and low receptive language at 3 years.



variability in twin samples such as ours, which include both MZ and DZ twins. Consequently, the long-range prediction of outcomes from early childhood is likely to be weaker in twin samples compared to the general population, whether focusing on the classification of language impairments or on continuous measures across the whole range of ability.

Fourth, the considerations above primarily focus on EA3, but the other risk factors have their own limitations. There is much evidence, for example, that early language and nonverbal behavior are so variable, even when they are reliable at any given point, that there is little prediction from them. The validity of family history information is limited by memory and by limited understanding of the nature of variability in the general population. Mothers' education is correlated, but far from perfectly, with actual interaction patterns. We also acknowledge that our measures did not include an assessment of receptive vocabulary. Psyridou et al. (2018) examined the reading outcomes of children through 16 years of age as related to vocabulary assessments at 2–2.5 years of age. They found the combination of family risk, expressive vocabulary, and receptive vocabulary had the largest negative impact on reading, whereas late talkers without receptive vocabulary difficulties tended to become typical readers.

Research Question 3 focused on the comparison of prediction from 2 versus 3 years of age (note that only vocabulary and nonverbal cognition can change from 2 to 3 years of age). Although the full-range predictions reported in Table 2, both for overall prediction and for the contribution of EA3, do show a significant improvement from 2 to 3 years of age, the changes are very small (< 3% and < 1% of the variance, respectively) with little practical significance. The analyses focused on risk factors and low outcomes (correlations in Table 3 and odds ratios in Table 4) do not identify any significant change, despite some trend for values at 3 years of age to be higher. The latter finding is interesting because it is inconsistent with what might be expected on the basis of genetic contribution, namely, that stability should increase as genetic endowment comes to play a larger role, perhaps due to lessened family environment effects as children spend more time outside the home. Between 2 and 3 years of age, the heritability of late talking (lowest 10% on vocabulary) in the larger TEDS sample has been estimated to increase from .28 at 2 years of age to .49 at 3 years of age (Dale & Hayiou-Thomas, 2013). In any case, a practical implication of these results is that waiting until 3 years of age does not substantially improve prediction.

Limitations and Future Directions

One limitation of the present findings is that they are based on a relatively small sample, due to the need for complete genetic and family information, along with complete vocabulary nonverbal cognition measures at ages 2 and 3 years, and at least one of the three outcomes measures. Although these criteria identified 1,420 children, generous for full-range correlational analysis, only 106–133 children were classified as showing low performance on each measure at 12 years of age. Replications of this general approach would be highly valuable.

The use of a GPS based on educational attainment, necessitated by the need for a large sample size in developing the measure, might also be seen as a limitation, though we have argued above that this may not be as important as it seems, given pleiotropy and the close relation of language/literacy and educational success. Certainly, it should be a high priority to develop a GPS based on literacy with a large data set, but that is likely to be far off. On the whole, behavior genetic research is limited less now by the cost and effort required to obtain molecular data than it is by the availability of quality phenotype measurement in large samples.

Another limitation is the use of vocabulary only as the early language measure. Leonard (2014) has suggested that reduced vocabulary, while often present in children who will manifest language/literacy difficulties later, is not a necessary or core feature of the condition. He suggests that grammatical difficulties are more central, and valid measures of morphosyntax are not possible until grammatical development is more advanced. Measures of tense and agreement appear to be especially relevant, at least for children learning English. Dale et al. (2003) did include a MacArthur–Bates Communicative Development Inventories–type measure of early grammatical development, which

focused primarily on combinatorial abilities, but they found it did not contribute significantly to prediction.

Taken together, these issues suggest that it may simply not be possible to predict from early childhood to later language/literacy development from the kinds of measures that were included in the current study. A promising alternative approach is to change focus from measures of knowledge, such as vocabulary, to process-oriented measures. Fernald and Marchman (2012), for example, have found that efficiency in identifying referents of familiar words, assessed by reaction time and accuracy in looking at a target picture from a presented pair, provides some differentiation between late-talking toddlers at 18 months who will "bloom" by 30 months from those who will not. Although sensitivity (55%–61%) and PPV (52%–66%) were not yet in the clinical range, they were better than in this study. In its present form, the looking-while-listening task is not feasible on a large scale, but technical improvements can be expected.

Our results overall are consistent with previous research, which has found that it is difficult to predict long-range language and literacy outcomes from toddlerhood. Although the EA3 GPS adds a little predictive power, it is not sufficient to be clinically useful.

Acknowledgments

The Twins Early Development Study (TEDS) is supported by a program grant to Robert Plomin from the U.K. Medical Research Council (MR/M021475/1), with additional support from the U.S. National Institutes of Health (AG046938). The research leading to these results has also received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/Grant Agreement 602768. S. V. is supported by a Jacobs Fellowship (2017–2019). We are grateful for the continuing participation of the TEDS twins and their families, which have made this research possible.

Data used for this submission may be made available on request to TEDS, through their data access mechanism (see http://www.teds.ac.uk/researchers/teds-data-access-policy). We will then consider requests for sharing data for appropriate research purposes.

References

Allegrini, A. G., Selzam, S., Rimfeld, K., von Stumm, S., Pingault, J. B., & Plomin, R. (2019). Genomic prediction of cognitive traits in childhood and adolescence. *Molecular Psychiatry, 24,* 819–827. https://doi.org/10.1038/s41380-019-0394-4

Ayorech, Z., Plomin, R., & von Stumm, S. (2019). Using DNA to predict educational trajectories in early adulthood. *Developmental Psychology, 55,* 1088–1095. https://psycnet.apa.org/doi/10.1037/dev0000682

Caspi, A., Houts, R. M., Belsky, D. W., Harrington, H., Hogan, S., Ramrakha, S., Poulton, R., & Moffitt, T. E. (2016). Childhood forecasting of a small segment of the population with large economic burden. *Nature Human Behaviour, 1,* Article No. 0005. https://doi.org/10.1038/s41562-016-0005

Dale, P. S., Harlaar, N., Hayiou-Thomas, M. E., & Plomin, R. (2010). The etiology of diverse receptive language skills at 12 years. *Journal of Speech, Language, and Hearing Research, 53*(4), 982–992. https://doi.org/10.1044/1092-4388(2009/09-0108)

Dale, P. S., & Hayiou-Thomas, M. E. (2013). Outcomes for late talkers: A twin study. In L. A. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 241–257). Brookes.

Dale, P. S., McMillan, A. J., Hayiou-Thomas, M. E., & Plomin, R. (2014). Illusory recovery: Are recovered children with early language delay at continuing elevated risk? *American Journal of Speech-Language Pathology, 23*(3), 437–447. https://doi.org/10.1044/2014_AJSLP-13-0116

Dale, P. S., Price, T. S., Bishop, D. V. M., & Plomin, R. (2003). Outcomes of early language delay: I. Predicting persistent and transient language difficulties at 3 and 4 years. *Journal of Speech, Language, and Hearing Research, 46*(3), 544–560. https://doi.org/10.1044/1092-4388(2003/044)

Dollaghan, C. (2013). Late talker as a clinical category: A critical evaluation. In L. A. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions and outcomes* (pp. 91–112). Brookes.

Duff, F. J., & Snowling, M. J. (2015). Reading disorders. In E. L. Bavin & L. R. Naigles (Eds.), *The Cambridge handbook of child language* (2nd ed., pp. 681–702). Cambridge University Press. https://doi.org/10.1017/CBO9781316095829.031

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur–Bates Communicative Development Inventories: User's guide and technical manual* (2nd ed.). Brookes. https://doi.org/10.1037/t11538-000

Fenson, L., Pethick, S., Renda, C., Cox, J., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics, 21,* 95–115. https://psycnet.apa.org/doi/10.1017/S0142716400001053

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development, 83*(1), 203–222. https://doi.org/10.1111/j.1467-8624.2011.01692.x

GOAL. (2002). *GOAL Formative Assessment: Key Stage 3*. Hodder & Stoughton.

Hammill, D. D., Brown, V. L., Larsen, S. C., & Wiederholt, J. L. (1994). *Test of Adolescent and Adult Language (TOAL-3)*. Pro-Ed.

Haworth, C. M. A., Davis, O. S. P., & Plomin, R. (2013). Twins Early Development Study (TEDS): A genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Research and Human Genetics, 16*(1), 117–125. https://doi.org/10.1017/thg.2012.91

Hayiou-Thomas, M. E., Dale, P. S., & Plomin, R. (2014). Language impairment from 4 to 12 years: Prediction and etiology. *Journal of Speech, Language, and Hearing Research, 57*(3), 850–864. https://doi.org/10.1044/2013_JSLHR-L-12-0240

Hayiou-Thomas, M. E., Smith-Woolley, E., & Dale, P. S. (2020). *Breadth versus depth: Cumulative risk model and continuous measure prediction of poor language and reading outcomes at 12*. Manuscript under review.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review, 26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

Kaplan, E., Fein, D., Kramer, J., Delis, D., & Morris, R. (1999). *The Wechsler Intelligence Scale for Children–Third Edition as a Process Instrument (WISC-III-PI)*. Psychological Corp.

Krapohl, E., Rimfeld, K., Shakeshaft, N. G., Trzaskowski, M., McMillan, A., Pingault, J.-B., Asbury, K., Harlaar, N., Kovas, Y., Dale, P. S., & Plomin, R. (2014). The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proceedings of the National Academy*

of Sciences, 111(42), 15273–15278. https://doi.org/10.1073/pnas.1408777111

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling developmental language difficulties from school entry into adulthood: Literacy, mental health, and employment outcomes. *Journal of Speech, Language, and Hearing Research, 52*(6), 1401–1416. https://doi.org/10.1044/1092-4388(2009/08-0142)

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghazian, O., Zacher, M., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Linnér, R. K., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics, 50,* 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

Leonard, L. (2013). Alternative routes to language impairment. In L. A. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 363–376). Brookes.

Leonard, L. (2014). *Children with specific language impairment* (2nd ed.). MIT Press. https://doi.org/10.7551/mitpress/9152.001.0001

Marinus, E., & Castles, A. (2015). Precursors to reading: Phonological awareness and letter knowledge. In E. L. Bavin & L. R. Naigles (Eds.), *The Cambridge handbook of child language* (2nd ed., pp. 661–680). Cambridge University Press. https://doi.org/10.1017/CBO9781316095829.030

Markwardt, F. C., Jr. (1997). *Peabody Individual Achievement Test–Revised–Normative Update*. Pearson Assessments.

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. *Journal of Speech, Language, and Hearing Research, 62*(9), 3381–3396. https://doi.org/10.1044/2019_JSLHR-L-19-0001

Oliver, B., Dale, P. S., Saudino, K. J., Petrill, S. A., Pike, A., & Plomin, R. (2002). The validity of a parent-based assessment of cognitive abilities in three-year-olds. *Early Child Development and Care, 172*(4), 337–348. https://doi.org/10.1080/03004430212713

Oliver, B., & Plomin, R. (2007). Twins' Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics, 10,* 96–105. https://doi.org/10.1375/twin.10.1.96

Psyridou, M., Eklund, K., Poikkeus, A.-M., & Torppa, M. (2018). Reading outcomes of children with delayed early vocabulary: A follow-up from age 2–16. *Research in Developmental Disabilities, 78,* 114–124. https://doi.org/10.1016/j.ridd.2018.05.004

Rescorla, L. A. (2013). Late-talking toddlers: A 15-year follow-up. In L. A. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 219–239). Brookes.

Rescorla, L. A., & Dale, P. S. (2013). *Late talkers: Language development, interventions, and outcomes*. Brookes.

Rice, M. L., Taylor, C. L., Zubrick, S. R., Hoffman, L., & Earnest, K. K. (2020). Heritability of specific language impairment and nonspecific language impairment at ages 4 and 6 years across phenotypes of speech, language, and nonverbal cognition. *Journal of Speech, Language, and Hearing Research, 63*(3), 793–813. https://doi.org/10.1044/2019_JSLHR-19-00012

Rice, M. L., Zubrick, S. R., Taylor, C. L., Gayán, J., & Bontempo, D. E. (2014). Late language emergence in 24-month-old twins: Heritable and increased risk for late language emergence in twins. *Journal of Speech, Language, and Hearing Research, 57*(3), 917–928. https://doi.org/10.1044/1092-4388(2013/12-0350)

Rice, M. L., Zubrick, S. R., Taylor, C. L., Hoffman, L., & Gayán, J. (2018). Longitudinal study of language and speech of twins at 4 and 6 years: Twinning effects decrease, zygosity effects disappear, and heritability increases. *Journal of Speech, Language, and Hearing Research, 61*(1), 79–93. https://doi.org/10.1044/2017_JSLHR-L-16-0366

Rimfeld, K., Malanchini, M., Spargo, T., Speckernell, G., Selzam, S., McMillan, A., Dale, P. S., Eley, T. C., & Plomin, R. (2019). Twins early development study: A genetically sensitive investigation into behavioural and cognitive development from infancy to emerging adulthood. *Twin Research and Human Genetics, 22*(Special Issue 6), 509–513. https://doi.org/10.1017/thg.2019.56

Samuelsson, S., Olson, R., Wadswork, S., Corley, R., DeFries, J. C., Willcutt, E., Hulslander, J., & Byrne, B. (2007). Genetic and environmental influences on prereading skills and early reading and spelling development in the United States, Australia, and Scandinavia. *Reading and Writing, 20,* 51–75. https://doi.org/10.1007/s11145-006-9018-x

Saudino, K. J., Dale, P. S., Oliver, B., Petrill, S. A., Richardson, V., Rutter, M., Simonoff, E., Stevenson, J., & Plomin, R. (1998). The validity of parent-based assessment of the cognitive abilities of 2-year-olds. *British Journal of Developmental Psychology, 16*(3), 349–362. https://doi.org/10.1111/j.2044-835X.1998.tb00757.x

Selzam, S., Ritchie, S. J., Pingault, J.-B., Reynolds, C. A., O'Reilly, P. F., & Plomin, R. (2019). Comparing within- and between-family polygenic score prediction. *American Journal of Human Genetics, 105*(2), 351–363. https://doi.org/10.1016/j.ajhg.2019.06.006

Smith-Woolley, E., Ayorech, Z., Dale, P. S., von Stumm, S., & Plomin, R. (2018). The genetics of university success. *Scientific Reports, 8,* 14579. https://doi.org/10.1038/s41598-018-32621-w

Snowling, M. J., & Melby-Lervåg, M. (2016). Oral language deficits in familial dyslexia: A meta-analysis and review. *Psychological Bulletin, 142*(5), 498–545. https://doi.org/10.1037/bul0000037

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Taylor, C. L., Rice, M. L., Christensen, D., Blair, E., & Zubrick, S. R. (2018). Prenatal and perinatal risks for late language emergence in a population-level sample of twins at age 2. *BMC Pediatrics, 18,* Article No. 41. https://doi.org/10.1186/s12887-018-1035-9

Thorpe, K. (2006). Twin children's language development. *Early Human Development, 82*(6), 387–395. https://doi.org/10.1016/j.earlhumdev.2006.03.012

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Pro-Ed.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H.-H., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discover, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Kathiresen, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., … Price, A. L. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics, 97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

von Stumm, S., Smith-Woolley, E., Ayorech, Z., McMillan, A., Rimfeld, K., Dale, P. S., & Plomin, R. (in press). Predicting educational achievement from genomic measures and socioeconomic status. *Developmental Science*.

Weismer, S. E. (2014). Specific language impairment. In L. Cummings (Ed.), *The Cambridge handbook of communication disorders* (pp. 73–87). Cambridge University Press. https://doi.org/10.1017/CBO9781139108683.007

Wertz, J., Moffitt, T. E., Agnew-Blais, J., Arseneault, L., Belsky, D.W., Corcoran, D. L., Houts, R., Matthews, T., Prinz, J. A., Richmond-Rakerd, L. S., Sugden, K., Williams, B., & Caspi, A. (2019). Using DNA from mothers and children to study parental investment in children's educational attainment. *Child Development*. Advance Online Publication. https://doi.org/10.1111/cdev.13329

Wiig, E. H., Secord, W., & Sabers, D. (1989). *Test of Language Competence–Expanded Edition*. Psychological Corporation.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Riverside.