



City Research Online

City St George's, University of London

Citation: Ward, L., Bannard, C., Norbury, C. & Polisenska, K. (2026). The Utility and Robustness of Sentence Repetition as a Marker of Developmental Language Disorder: An Analysis of SCALES Project Data. *Journal of Speech, Language and Hearing Research*, 69(4), pp. 1669-1685. doi: 10.1044/2025_JSLHR-25-00058

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/36667/>

Link to published version: https://doi.org/10.1044/2025_JSLHR-25-00058

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

The Utility and Robustness of Sentence Repetition as a Marker of Developmental Language Disorder: An Analysis of SCALES Project Data

Leah Ward¹, Colin Bannard², Courtenay Norbury³ and Kamila Polišíenská^{1,4}

¹ Division of Psychology, Communication and Human Neuroscience,
The University of Manchester, United Kingdom

² Department of Linguistics and English Language,
The University of Manchester, United Kingdom

³ Division of Psychology and Language Sciences,
University College London, United Kingdom

⁴ Department of Language and Communication Science,
City St George's, University of London, United Kingdom

Author Note

Leah Ward  <https://orcid.org/0000-0003-3728-5591>

Colin Bannard  <https://orcid.org/0000-0001-5579-5830>

Courtenay Norbury  <https://orcid.org/0000-0002-5101-6120>

Kamila Polišíenská  <https://orcid.org/0000-0001-7405-6689>

Kamila Polišíenská is now at the Department of Language and Communication Science, City St George's, University of London.

This study was preregistered at <https://osf.io/ue9qb/>. The authors have no conflicts of interest to disclose.

Correspondence concerning the article should be addressed to Leah Ward, Division of Psychology, Communication and Human Neuroscience, The University of Manchester, Manchester, United Kingdom, M13 9PL. Email: leah.ward-2@postgrad.manchester.ac.uk

Abstract

Purpose: This study examined the diagnostic power and robustness of sentence repetition (SR) in distinguishing monolingual English-speaking children with developmental language disorder (DLD) from typically developing (TD) children, using a large-scale secondary analysis of the Surrey Communication and Language in Education Study (SCALES) dataset (Norbury, 2022). We evaluated SR's effectiveness across different scoring methods in age-matched groups across three different time-points.

Method: A total of 407 children were included. Bayesian beta regression models compared SR performance between groups across four scoring methods. SR's diagnostic accuracy was assessed using area under the curve (AUC), sensitivity, specificity, and likelihood ratios.

Results: Children with DLD consistently performed less accurately than TD peers on SR tasks, with differences persisting across ages five to 11 and across all scoring methods. AUC values indicated good to excellent discriminative ability, especially in younger groups, though sensitivity and specificity fell below clinical thresholds.

Conclusions: These findings support SR's diagnostic potential as a reliable language measure. SR may serve as a valuable tool due to its ease of use and scoring flexibility, though additional assessments are recommended for comprehensive language evaluation.

Keywords: *Developmental Language Disorder; Sentence Repetition; Screening Tools*

Introduction

Developmental language disorder (DLD) is characterised by impairments in acquisition, use, and understanding of spoken language (Bishop et al., 2017). DLD can have significant long-term impacts on academic achievement, social interaction and overall communication skills (Conti-Ramsden et al., 2001; Durkin & Conti-Ramsden, 2007). There is a critical need for accurate and reliable tools that can identify children with DLD or identify those who may require further diagnostic evaluation (McGregor, 2020). One promising marker is performance on sentence repetition (SR) tasks. The evidence base for the diagnostic usefulness of SR as a marker for DLD is somewhat problematic, due to the small size of the samples and diverse procedures used in studies. However, a recent meta-analysis (Ward et al., 2024) found the collective evidence to be robust. The current study seeks to build on this by investigating the diagnostic power and robustness of SR tasks in a large-scale, secondary data analysis of the Surrey Communication and Language in Education Study (SCALES; Norbury et al., 2022). Through a series of simulated case-control studies we assess the effectiveness of SR tasks in detecting DLD across different scoring methods and across children aged five to 11.

The use of SR within the Literature

The use of performance on SR tasks as a marker for DLD has gained traction since Conti-Ramsden et al. (2001) demonstrated that SR achieved better sensitivity (true positives; 90%) and specificity (true negatives; 85%) for identifying DLD than a third-person singular task, a past tense marking task, or a nonword repetition task. Archibald and Joanisse (2009) found SR tasks to be a better predictor of DLD than nonword repetition for English-speaking children, with sensitivity and specificity values both exceeding 80%. A meta-analysis by Pawlowska (2014) looked at the use of three marker tasks (verb tense, nonword repetition, and SR) of language impairment in English speaking children across 13 studies (four of which used SR). SR was found to be the best marker of the three, with Pawlowska noting that this may be due to the advantages of SR in being sensitive to a wide range of language and processing deficits. Their conclusion is that SR is promising as a

screening tool for DLD, but that it needs further research validation. This hesitancy is justified by the high levels of heterogeneity and small samples sizes seen across the studies reviewed. There is also evidence of SR's utility beyond English, with studies showing sensitivity and specificity above 80% for monolingual speakers of Hebrew (Armon-Lotem & Meir, 2016), Russian (ibid.), French (Fleckstein et al., 2018; Leclercq et al., 2014), German (Hamann & Abed Ibrahim, 2017), Mandarin (Wang et al., 2022), and Turkish (Kaçar-Kütükçü & Topbaş, 2024).

The methodological variation in the application of SR tasks across the literature has been considered in a couple of recent reviews. A scoping review by Rujas et al. (2021) looked at the broad use of SR across 203 studies published in English or Spanish. There were at least 50 original sentence repetition tasks seen across the studies with considerable variation seen in the uses of these tasks. From their review, SR is concluded to be useful for assessing language ability and specifically presents a "suitable" task for detecting the presence of DLD in children. A recent systematic review and meta-analysis by Ward et al. (2024) explored the use of SR tasks in the literature for distinguishing monolingual DLD and TD groups of children. There were 66 studies identified, which involved 19 different languages, 39 different SR tasks, and 4447 children. Across the studies, children with DLD were found to perform an average of 2.08 standard deviations (SD) worse than TD children on SR tasks. A key finding was that this power to differentiate the groups was robust to differences in (1) whether the tasks were standardised or not, (2) whether sentences were prerecorded or not, (3) whether scoring was conducted during the task or after, (4) the type of scoring method used, (5) whether clinical or population DLD samples were used, and (6) the language the task was conducted in. This supports the clinical application and usefulness of SR, as a task that can be adapted to contextual demands.

Ward et al. (2024) note key limitations of their review, similar to those mentioned by Pawlowska (2014). These include small sample sizes within individual studies, as well as comparisons of individual aspects of methodological variation being limited by high heterogeneity and potential confounding between studies. The current study seeks to

address these limitations by conducting a secondary data analysis within a single large-scale, population-based study. This approach allows for consistent methodology, larger sample sizes, and controlled comparisons within the same dataset, thereby reducing between-study confounds and increasing interpretability. Further, while reviews such as these provide valuable insight into the methodological variability of SR tasks, few studies have directly examined age-related effects on SR performance or diagnostic accuracy. The meta-analysis by Ward et al., (2024) included studies with participants ages between three and 16 years, and although encompassing a broad developmental range, age effects were not specifically examined. Similarly Rujas (2021) reported that most SR studies have focused on children aged between five and eight years, with fewer including younger or older participants. In the current study, age was controlled for within each time point to reduce potential confound. Although we did not formally assess age-related changes across time, examining SR performance at three developmental stages provides an informal indication of how performance on a specific SR task may change across the age range for which it was designed.

Methods of Scoring SR Performance

One key methodological variation in the application of SR tasks is in how the productions are scored. These measures of SR performance can for example involve binary scoring where the whole sentence production is marked as either correct or incorrect. For instance, for the target sentence “The dog is running”, a binary approach would assign a score of one only if the entire sentence was produced correctly. They can also be more detailed and involve a score per production in relation to the number of errors made, or whether specific targets within the sentence were produced correctly or not. Again, for a target such as “The dog is running,” scores may be assigned based on whether specific elements (e.g., the, dog, is, running) are produced correctly. Each method can bring with it different potential benefits. A binary method of scoring is the simplest method and has been suggested to be the most efficient method for use in clinical contexts because of this (Hamman & Abed Ibrahim, 2017). In addition, Ebert et al. (2019) found that binary scoring

methods can be reliably implemented, even by those with no background in language assessment. On the other hand, scoring which considers specific target structures within sentences provide the most detailed window onto a child's language ability (Komeili & Marshall, 2013).

In terms of which scoring method can better identify children with DLD, there have been mixed research findings. Some research (e.g., Hamaan & Abed Ibrahim, 2017; Taha et al., 2021; Theodorou et al., 2017) indicates that scoring method does not affect a SR task's ability to distinguish children with DLD from TD children (when looking to achieved sensitivity and specificity). However, other studies (e.g., Pham & Ebert, 2020; Wang et al., 2022) have found that binary scoring methods less reliably distinguish children with DLD from TD children (lower sensitivity and specificity values) when compared to more detailed methods of scoring.

One important consideration in interpreting these findings is the language in which the SR tasks were administered. The studies cited above span multiple languages, including German, Arabic, Greek, Vietnamese, and Mandarin. These languages vary in features such as syntactic structure and morphological richness, which may influence how reliably different measures capture language impairment (Leonard, 2017). For example, languages with rich morphology may benefit more from specific target or error-based scoring, whereas in languages like English, which is less morphologically complex, binary scoring might be just as effective.

In the meta-analysis by Ward et al. (2024), a subgroup analysis comparing four types of scoring method (binary, subsentence binary, error, and target structure scoring) revealed no effect of scoring decision on the ability of SR to distinguish DLD and TD groups across the studies. The current study directly compares how choice of scoring method influences the diagnostic power of SR for identifying DLD in English, within a single large-scale sample.

Current Study

In the current study, we compare SR performance in DLD and TD groups via a secondary data analysis using data from the Surrey Communication and Language in Education Study:

Intensive Data T2-T5, 2012-2020 (SCALES; Norbury et al., 2022). SCALES is a longitudinal UK population study of language development involving state-maintained schools across Surrey, England. SR performance within SCALES was measured using the School-aged Sentence Imitation Test (SASIT-E32; Marinis et al., 2011), a SR task designed to assess a range of morphosyntactic and syntactic structures in school-aged children. The SASIT-E32 has been shown to be valuable in identifying DLD in both monolingual and bilingual children (e.g., Komeili & Marshall, 2013; Chiat & Roy, 2013). We compare performance on this measure across four scoring methods: whole sentence accuracy (binary scoring), correct content words, correct function words, and correctly inflected verbs.

Using this data, children identified as meeting criteria for DLD were matched to TD children through age-matching across three separate time-points (ages 5-6, 7-8, and 10-11). In doing so, we aim to assess differences in SR performance across scoring methods. This study addresses this in two ways. First, Bayesian beta regression models, and specifically Bayes Factors, are used to compare SR performance between groups. The primary advantage of this Bayesian approach is that it allows us to quantify the strength of support for both the experimental and the null hypothesis (meaning we would be able to distinguish evidence of the absence of an effect from mere absence of evidence due to sample size). Second, we assess the ability of SR to detect DLD with measures of diagnostic accuracy (sensitivity, specificity, and likelihood ratios).

We aim to address the following questions:

1. Do monolingual children with DLD differ from age-matched groups of TD children on SR tasks?
2. What diagnostic accuracy does the SR task achieve?
3. Does the type of scoring used to measure performance on the SR task affect the difference in performance between DLD and TD groups and the diagnostic power of the SR task?

Based on the results of the Ward et al. (2024) meta-analysis, we hypothesise that children with DLD will perform with less accuracy on each method of scoring on the SASIT-

E32 in comparison to TD children. We predict that calculated Bayes factors will indicate support for this hypothesis. We also predict that each method of scoring on the SASIT-E32 will show good diagnostic accuracy, consistent with conventional benchmarks (including sensitivity and specificity exceeding 0.90, a positive likelihood ratio greater than 3, a negative ratio below 0.30, and an area under the curve of at least 0.70), in differentiating between children with DLD and TD children.

Method

Our study, including research questions and methods of analyses was preregistered at <https://osf.io/ue9qb/>. We address our research questions through a secondary data analysis. We use data collected as part of the Surrey Communication and Language in Education Study (SCALES; Norbury et al., 2022), which is publicly accessible through the UK Data Service with identifier <https://doi.org/10.5255/UKDA-SN-8968-1>. SCALES is a longitudinal UK population study of language development involving state-maintained schools across Surrey, England. SCALES protocol and consent procedures were developed in consultation with Surrey County Council and approved by the Royal Holloway Ethics Committee in Year 1 of the study and the UCL Research Ethics Committee (9733/002) in Year 6 of the study.

Data Collection Procedures

Data collection for SCALES occurred in two phases. In the first phase, all state-maintained schools in Surrey were invited to participate in 2011. This resulted in the screening of 7,267 children in reception classes (four or five years of age) across 161 schools. This is referred to as time-point 1 (T1) of the study. As part of this screening, teachers completed the Children's Communication Checklist-Short (CCC-S; Bishop, 2003a) resulting in a score per child out of 39. A subsample was then selected for in-depth assessment in the second phase of the study using a stratified random sampling process. All children reported to have "no phrase speech" (the teacher responded "no" to the question "Is the child combining words into phrases or sentences?") were selected for inclusion. The remaining children were stratified by sex, season of birth, and risk for DLD (higher scores on

the CCC-S) for selection. Specifically, there was an oversampling of girls and those with CCC-S scores that indicated a low language-level. Those from larger schools were also more likely to be selected. Those attending special schools for learning disabilities and those with English as an additional language were excluded.

This resulted in 529 children selected for intensive testing as part of phase two of SCALES. This testing occurred at four different timepoints – when the children were in year 1 (T2; five to six years olds), year 3 (T3; seven to eight years old), year 6 (T4; ten to 11 years old), and year 8 (T5; 12-13 years old). T5 data collection however coincided with Covid-19 school closings, resulting in greater than expected attrition. Testing involved an assessment battery which included measures of speech and language, non-verbal cognitive skills, motor skills, executive function, speed of processing, social cognition, emotion processing, and literacy. Teachers and caregivers also completed questionnaires at each time point. In the current research, a subset of this dataset is used.

Participants

A total of 407 monolingual English-speaking children from Surrey are included in the current research. Specifically, this included 407 children when they were in year 1 (T2), 391 children when they were in year 3 (T3), and 294 children when they were in year 6 (T3). This sample represents a subset of the initial dataset of 529 children who underwent intensive testing in phase two of SCALES, based on the following exclusions. Children were excluded if they were reported to have a biomedical condition (based on caregiver and teacher reports) at any of T2, T3, or T4. Additionally, children who failed the audiometric screening test (hearing assessment) at T2 were excluded. All included children scored above -2 SDs on a non-verbal composite ability score at T2, T3, and T4. This composite score was derived from performance on the block design and matrix reasoning subtests of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 2003). Finally, for inclusion at each specific time point, children were required to have SR scores for that time point.

Whether or not children met criteria for DLD differed between SCALES and the current study. For both, the criteria for DLD was based upon core language battery

measures assessing vocabulary, grammar, and narrative skills across expressive and receptive language. These included the Expressive and Receptive One Word Picture Vocabulary Test (EOWPVT and ROWPVT respectively; Martin & Brownell, 2011), the Test of Reception of Grammar – short form (TROG-S; Bishop, 2003b), the Assessment of Comprehension and Expression 6-11 Narrative Recall (ACE-Recall; Adams et al., 2001), and a narrative comprehension task derived from the ACE-Recall made by the SCALES team (ACE-Comp). These measures were selected to replicate existing epidemiological work (Tomblin et al., 1997) that has informed current diagnostic algorithms for language disorder in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). A full outline of the core language battery measures, including their purpose and known reliability, is provided in the supplementary materials of Norbury et al. (2016). DLD criteria within SCALES also included the SASIT-E32. However as SR performance is the variable of interest in the current research, it is not used for DLD classification here.

Inverse probability weighting was utilised in SCALES to produce weighted z-scores of performances (for each of the above core language measures). This weighting process was derived during SCALES (as outlined by Vamvakas et al., 2019) and was conducted to correct for sampling bias (as described earlier), ensuring that the tested sample reflected the full cohort in terms of factors such as gender, school size, and risk for language disorder. These weighted z-scores of language performance were used solely for DLD identification.

Five composite scores were created from these measures by averaging the z-scores of the individual tasks within each domain: expressive language (EOWPVT and ACE-Recall), receptive language (ROWPVT, TROG-S and ACE-Comp), vocabulary (EOWPVT, ROWPVT), grammar (TROG-S), and narrative skills (ACE-Recall and ACE-Comp). At T4, only four composite scores were created (all but the grammar composite), due to the TROG-S not being administered. At each time point, children were identified as meeting criteria for DLD if they scored 1.5SD below the mean on at least two of these composite scores. This threshold aligns with the -1.5 SD cut-off used in SCALES and approximately mirrors the widely reported 7.6% prevalence estimate for DLD in the UK population (Norbury et al.,

2016). Children identified as meeting criteria for DLD at any of the time points were classified as DLD across all the time points. This resulted in 67 children meeting criteria for DLD at T2, 66 at T3, and 50 at T4. Children who did not meet this criterion for DLD based upon the core language measures were classified as being TD. Based on this, there were 340 TD children at T2, 325 at T3, and 244 at T4.

The consequences of this decision to classify children meeting criteria for DLD at any one time point as DLD across all the time points were examined through a sensitivity analysis. This involved carrying out the same analysis outlined below with DLD groups where children are only identified as DLD at a timepoint if they met the criteria at that specific timepoint. The results of this, including a table summarising how many children met criteria for DLD at each time point specifically, are outlined in supplementary materials S1 and it was concluded that this methodological decision did not influence results.

Samples

There were three age-matched groups corresponding to the SCALES data collection time points. The time-points are based on UK school years, with T2 being year 1 (ages five to six years), T3 being year 3 (ages seven to eight), and T4 being year 6 (ages ten to 11). Table 1 provides a summary of the sample characteristics for the samples.

Table 1

Characteristics of the typically developing (TD) children, and children identified as meeting criteria for developmental language disorder (DLD)

Time-Point 2						
	<i>n</i>	Age in months	Male	Female	White Ethnic Origin (%)	IDACI
TD	340	71.83 (4.93)	162	178	90.88	3.89 (1.12)
DLD	67	71.10 (3.97)	37	30	83.58	3.33 (1.19)
Time-Point 3						
	<i>n</i>	Age in months	Male	Female	White Ethnic Origin (%)	IDACI
TD	325	95.38 (4.21)	152	173	90.46	3.87 (1.13)
DLD	66	95.36 (5.26)	37	29	83.33	3.35 (1.18)
Time-Point 4						
	<i>n</i>	Age in months	Male	Female	White Ethnic Origin (%)	IDACI
TD	244	134.08 (4.16)	111	133	91.80	3.94 (1.11)
DLD	50	133.18 (3.73)	26	24	86.00	3.32 (1.28)

Note. Age and IDACI are reported as means and standard deviations. TD = typically developing, DLD = developmental language disorder, IDACI = Income Deprivation Affecting Children Index. IDACI is post-code indicator of neighbourhood deprivation, reported here as a quintile rank with higher ranks indicating lower deprivation. Those not defined as “white ethnic origin” were defined as “other ethnic origin”.

Sentence Repetition Task

The SR task used within SCALES was the School-Aged Sentence Imitation Test (SASIT-E32; Marinis et al., 2011). The SASIT-E32 consists of 32 sentences and is designed to assess a range of morphosyntactic and syntactic structures, including auxiliaries, passives, object wh- questions, sentence complements, temporal and conditional sentence adjuncts, and object relative clauses. It was developed based on principles from the LITMUS (Language Impairment Testing in Multilingual Settings) framework (Marinis & Armon-Lotem, 2015). Full details of the task can be found here:

https://www.citystgeorges.ac.uk/__data/assets/word_doc/0010/118864/SASIT-E32-Protocol,-Structures-and-Scoresheet.doc

This was administered as per manual instructions. It is a short, structured SR task, taking about 5 to 10 minutes to complete. Using a PowerPoint Presentation, children listened to 32 pre-recorded sentences through headphones, verbally repeating each sentence upon hearing it. Responses were audio recorded and later scored.

The SASIT-E32 manual outlines multiple scoring options, allowing for flexibility in scoring and level of linguistic analysis. The current study looks at the following four methods of scoring, as used within SCALES:

- Binary Scoring: A simple measure of whole sentence accuracy. Each sentence was scored as correct (1) or incorrect (0), with a maximum total score of 32.
- Function Word Scoring: Each function word was scored as correct or incorrect/omitted. The total number of correctly produced function words was calculated (maximum score of 176). For example, in the sentence, “The sandwich was eaten by the postman”, there would be a total of four function words: “the”, “was”, “by”, and “the”.
- Content Word Scoring: Each content word was scored as correct or incorrect/omitted. The total number of correctly produced content words was calculated (maximum score of 121). For example, in the sentence, “The sandwich was eaten by the postman”, there would be a total of three content words: “sandwich”, “eaten”, and “postman”.
- Verb Inflection Scoring: Each verb inflection is scored as correct or incorrect. The total number of correctly inflected verbs was calculated (maximum score of 44). For example, in the sentence, “The sandwich was eaten by the postman”, one point would be awarded for the correct inflection of “was”.

Statistical Analysis

All analyses were run using R software (Version 4.3.1; R Core Team, 2023). This study looked at SR performance in those identified as meeting criteria for DLD and TD children.

RQ1: Do monolingual children with DLD differ from age-matched groups of TD children on SR tasks?

We first calculated effect sizes using Hedges' g , a measure of standardised mean difference (SMD) used in the meta-analysis by Ward et al. (2024). This allowed us to compare our results with existing literature. These were calculated for each method of scoring on the SR task for each of the age-matched groups.

To formally examine performance differences, Bayesian beta regression analyses were conducted. A beta regression was chosen due to its flexibility in modelling distributions of values between zero and one (such as proportions). They are flexible to distribution, being able to handle multiple instances of very high or very low scores (such as patterns of floor/ceiling levels of performance) and thus reducing the risk of type one or two error. A Bayesian approach was chosen to measure the strength of differences between group performance.

Proportion correct on the SR task was the dependent variable, and group (DLD or TD) was the independent variable. Separate models were run for each method of scoring SR performance and time point. Since the TD and DLD groups were matched for age by school year rather than individually, age was included in the analyses at each time point individually as a potential covariate on the effect. If cases of floor performance (proportion = 0) or ceiling performance (proportion = 1) were present, models were fit to accordingly account for the extreme scores (zero-inflated, one-inflated, or zero-one inflated beta regression).

The Bayesian models were fit using the [brms] package in R (Bürkner, 2017). Bayesian analysis requires us to set priors on our parameters, reflecting our knowledge state before observing the data. Our knowledge is then updated using the data in order to give the *posterior* parameter estimates. As we do not have any existing identical data, we cannot assume that there is an effect of group and so we set the mean of our prior on the group

term to be zero. We assume only that we might see values as extreme as those observed in previous work by chance and so set our prior on the group term to be a normal distribution with a mean of 0 and a standard deviation corresponding to the largest plausible group difference in log odds seen in previous work (see Dienes, 2014 for a recommendation of such an approach).. This was derived (in log odds) from the meta-analysis conducted by Ward et al. (2024), which reported an average effect size of $g = -2.08$ (95% CI [-2.32, -1.84]). We took the upper limit of this confidence interval (2.32) as the standard deviation of our prior. We have no basis on which to set informative priors on the other parameters (intercept and age) and so these were set to be wide and uninformative as $\text{Normal}(0,10)$.

To assess the strength of evidence Bayes factors were computed for each model using the method of bridge sampling. A Bayes factor of 1 indicates ambiguous evidence; a Bayes factor of > 1 indicates evidence in support of the hypothesis that there is a difference in performance between the DLD and TD groups; and a Bayes factor < 1 indicates evidence in support of the null hypothesis. We interpreted the Bayes factors using guidance by Lee and Wagenmakers (2013; adjusted from Jeffreys, 1961; see Appendix A).

RQ2: What diagnostic accuracy does the SR task achieve?

We calculated diagnostic accuracy measures to evaluate the ability of the SR task to identify the presence or absence of DLD in the groups. These measures were derived from receiver operating characteristic (ROC) curves, using the [cutpointr] package in R (Thiele, 2022). The ROC curve plots the true positive rate (sensitivity) against the false positive rate ($1 - \text{specificity}$) across various performance thresholds. The optimal cut-off was identified using Youden's Index as the threshold providing the best balance between sensitivity and specificity. This specifically being the point where the ROC curve was closest to the upper left corner (which would indicate perfect classification).

Sensitivity and specificity values were interpreted using guidelines by Plante and Vance (1994) whereby values of 90% and above demonstrate good classification accuracy, values of 80% to 89% demonstrate acceptable classification accuracy, and values below this indicate poor classification accuracy. Likelihood ratios (LRs) were then calculated to assess

how a change in SR score changes the probability of a child having DLD. Likelihood ratios were interpreted using recommendations by Dollaghan (2007). LR+ over 10 and LR- under 0.1 would indicate that the SR test could rule in and out DLD with high confidence. LR+ over 3 and LR- below 0.3 would imply a moderate confidence, and LR+ under 3 and LR- over 0.3 would indicate an uninformative test.

Area under the ROC curve (AUC) is also reported, summarising the overall accuracy of the SR task. Following guidelines by Hosmer and Lemeshow (2000), AUC values above 0.9 suggest outstanding discriminative ability, values between 0.8 to 0.9 suggest excellent discriminative ability, values between 0.7 to 0.8 suggest an acceptable ability, and values below this indicate an unacceptable ability.

Again, as groups were matched by school year, the influence of any residual age differences were assessed. This was done by plotting a separate ROC curve using age as the predictor. The resulting curve was observed to be very close to the diagonal, indicating that age alone did not discriminate between DLD and TD groups at that time point.

RQ3: Does the type of scoring used to measure performance on the SR task affect the difference in performance between age-matched DLD and TD groups and the diagnostic power of the SR task?

To answer this, Bayes Factors were compared for each regression model for each scoring method and time point, assessing the relative strength of evidence for group differences. Interpretations of Bayes factors (as outlined in RQ1) were directly compared to determine whether the level of support for the presence or absence of an effect differed across scoring methods.

To compare the diagnostic performance we compare AUC values for each scoring method at each time point using the [pROC] package in R (Robin et al., 2011). We used bootstrap methods with 10,000 bootstrap replicates to test whether one scoring method produced a significantly different AUC compared to another. This helped us to determine whether any of the scoring methods were better at distinguishing DLD from TD children.

435 **Results**

436 Performance on the SR task for each group, expressed as proportion correct across
437 all methods of scoring, is shown in Table 2. As shown, mean scores for some scoring
438 methods, and particularly for the TD group at later time points, began to approach ceiling
439 levels, with proportion correct nearing 1.00.

440 In their meta-analysis, Ward et al. (2024) found an average effect size of $g = -2.08$
441 (95% CI [-2.32, -1.84]) for the difference in performance between DLD and TD groups
442 across the studies reviewed. Effect sizes resulting from the present study for each group
443 comparison are shown in Table 3.

444

445 **Table 2**

446 *Summary of SR performance for every matched TD-DLD grouping*

Time-Point	Binary Scoring		Content Word Scoring		Function Word Scoring		Verb Inflection Scoring	
	TD	DLD	TD	DLD	TD	DLD	TD	DLD
T2	0.54	0.24	0.92	0.80	0.88	0.71	0.78	0.53
	(0.24)	(0.23)	(0.08)	(0.12)	(0.10)	(0.17)	(0.15)	(0.22)
T3	0.71	0.47	0.97	0.91	0.94	0.86	0.89	0.75
	(0.17)	(0.24)	(0.04)	(0.08)	(0.04)	(0.11)	(0.08)	(0.17)
T4	0.85	0.70	0.99	0.94	0.97	0.92	0.94	0.86
	(0.10)	(0.19)	(0.02)	(0.07)	(0.02)	(0.08)	(0.05)	(0.12)

447 *Note.* Performance is reported as group means and standard deviations of raw scores. TD =
448 typically developing, DLD = developmental language disorder.

449

450

451

452

453

454

455

Table 3

Summary of effect sizes for SR performance differences between each matched TD and DLD group

Time-Point	Binary Scoring	Content Word Scoring	Function Word Scoring	Verb Inflection Scoring
T2	-1.28 [-1.56, -1.00]	-1.33 [-1.60, -1.05]	-1.57 [-1.85, -1.29]	-1.47 [-1.75, -1.19]
T3	-1.34 [-1.63, -1.06]	-1.26 [-1.54, -0.98]	-1.46 [-1.74, -1.17]	-1.44 [-1.72, -1.15]
T4	-1.24 [-1.56, -0.92]	-1.11 [-1.43, -0.79]	-1.13 [-1.45, -0.81]	-1.19 [-1.51, -0.87]

Note. Effect sizes represent a calculation of standardised mean differences using Hedges g (Hedges, 1981). Effect sizes are reported with 95% confidence intervals.

Group Comparisons

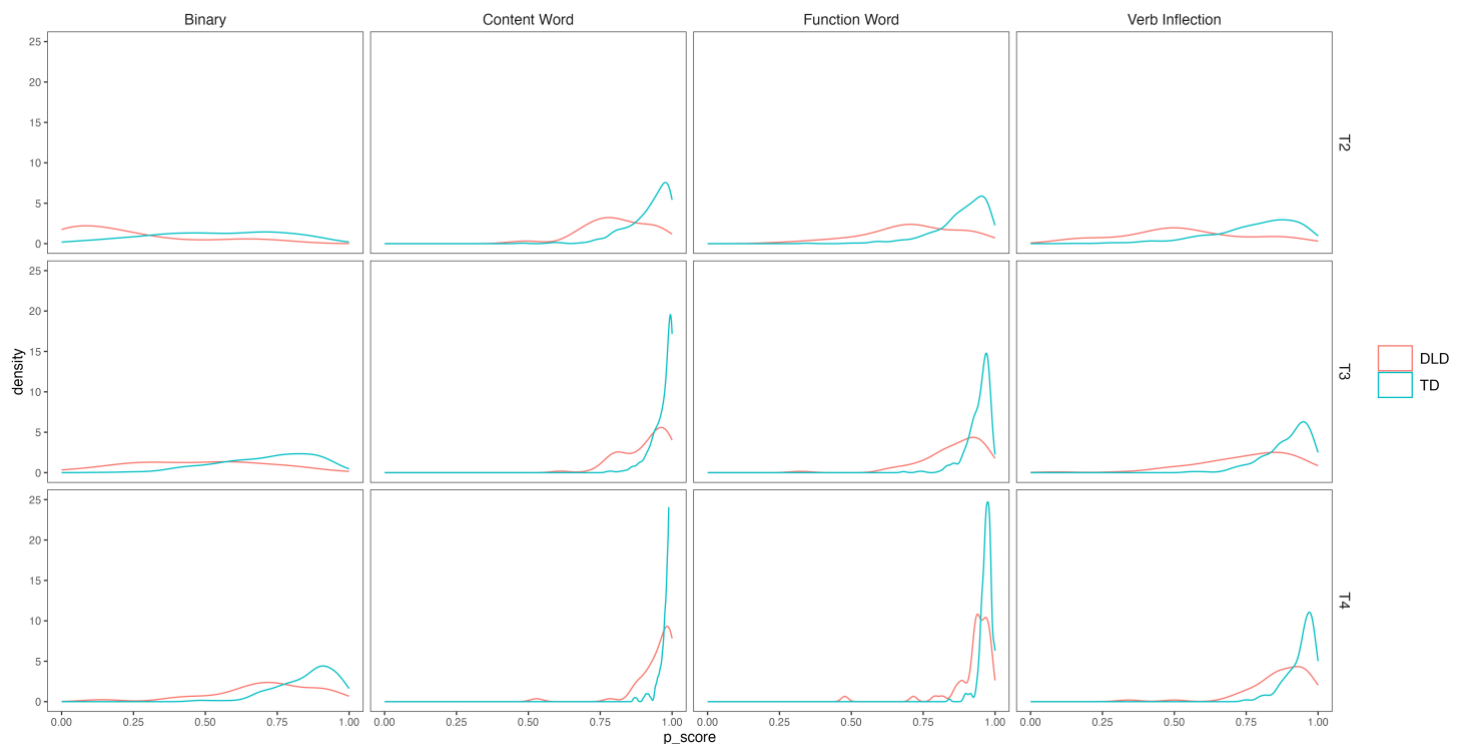
Comparisons of performance between DLD and TD groups on the SR task is visualised in Figure 1, for proportion correct by each of the four methods of scoring. Bayesian beta regression models were fit to compare the difference in performance between the groups. Overall, 12 models were fit for each of the four scoring methods for each of the three age-matched groups. All models were fit with a one-inflated beta regression model due to the presence of ceiling levels of performance in some children. This is except for the comparison of binary scores with children at T2, where the model was fit with a zero-one inflated model due to the presence of both ceiling and floor levels of performance. Parameters on all the final models converged successfully (all R^2 values < 1.01) and posterior predictive checks indicated good model fit. Supplementary materials S2 includes time-series plots of Markov chains showing parameter convergence for all models, and plots for posterior predictive checks comparing the observed and simulated distributions of task performance.

Table 4 presents the average marginal effects summarising the difference in SR performance between TD and DLD groups in interpretable terms of proportion correct. A full

outline of the posterior parameters for each model is provided in Appendix B. There was found to be no confounding influence of age across any of the time points, with 95% credible intervals for the age parameter overlapping zero across all models.

Figure 1

Density plot of the difference in SR performance between DLD and TD groups for each method of scoring.



Note. DLD = Developmental language disorder, TD = Typically developing, p_score = proportion correct on the SR task. Each row represents each group, being those matched at time point 2, 3, and 4 respectively.

Table 4

Average marginal effects of being TD (as opposed to meeting criteria for DLD) on SR performance, along with resulting Bayes Factors for the strength of evidence in support of an effect of group on SR performance.

Time-Point	Scoring Method	Average marginal effect	95% CIs	BF 10
T2	Binary	0.25	[0.19, 0.31]	1.01×10^{16}
	Content word	0.10	[0.07, 0.14]	1.43×10^{12}
	Function word	0.17	[0.13, 0.21]	2.48×10^{19}
	Verb Inflection	0.23	[0.18, 0.29]	4.17×10^{17}
T3	Binary	0.25	[0.19, 0.31]	1.87×10^{18}
	Content word	0.06	[0.04, 0.08]	4.82×10^{10}
	Function word	0.09	[0.07, 0.12]	2.34×10^{23}
	Verb Inflection	0.15	[0.11, 0.19]	2.73×10^{19}
T4	Binary	0.15	[0.10, 0.21]	1.87×10^{13}
	Content word	0.04	[0.03, 0.07]	4.07×10^{12}
	Function word	0.05	[0.03, 0.07]	5.31×10^{21}
	Verb Inflection	0.08	[0.05, 0.11]	5.85×10^{10}

Note. Average marginal effects refer to the proportion difference in SR performance from being TD as opposed to meeting criteria for DLD. A positive effect meaning better SR performance. BF 10 = Bayes factor value giving the evidence for H_1 over H_0 . Bayes factors were calculated using a bridge sampling procedure.

At T2 (UK school year 1, ages five to six years), the found marginal effects suggest that on average children with DLD scored 25% lower when compared to those who were TD using the binary method of scoring, 10% less using content word scoring, 17% less using function word scoring, and 23% less when using verb inflection scoring. At T3 (UK school year 3, ages seven to eight years), the found marginal effects suggest that on average children with DLD scored 25% less when compared to those who were TD using the binary method of scoring, 6% less using content word scoring, 9% less using function word scoring, and 15% less when using verb inflection scoring. And at T4 (UK school year 6, ages ten to

11 years), the found marginal effects suggest that on average children with DLD scored 15% less when compared to those who were TD using the binary method of scoring, 4% less using content word scoring, 5% correct using function word scoring, and 8% less when using verb inflection scoring.

Table 4 also outlines the Bayes factors which resulted from each model for the strength of evidence in support of an effect of group on SR performance. Following interpretation guidance from Lee and Wagenmakers (2013; adjusted from Jeffreys, 1961), there was extreme evidence for there being a difference in SR performance between the age-matched DLD and TD groups. This was true for all time points and there was no difference in this classification by method of scoring the SR productions.

Diagnostic Accuracy

Table 5 shows the optimal cut-off points for each method of scoring the SR task, along with the resulting AUCs, sensitivity, specificity, and likelihood ratios for each of the time points. Figure 2 plots the ROC curves for the scoring methods at each time point.

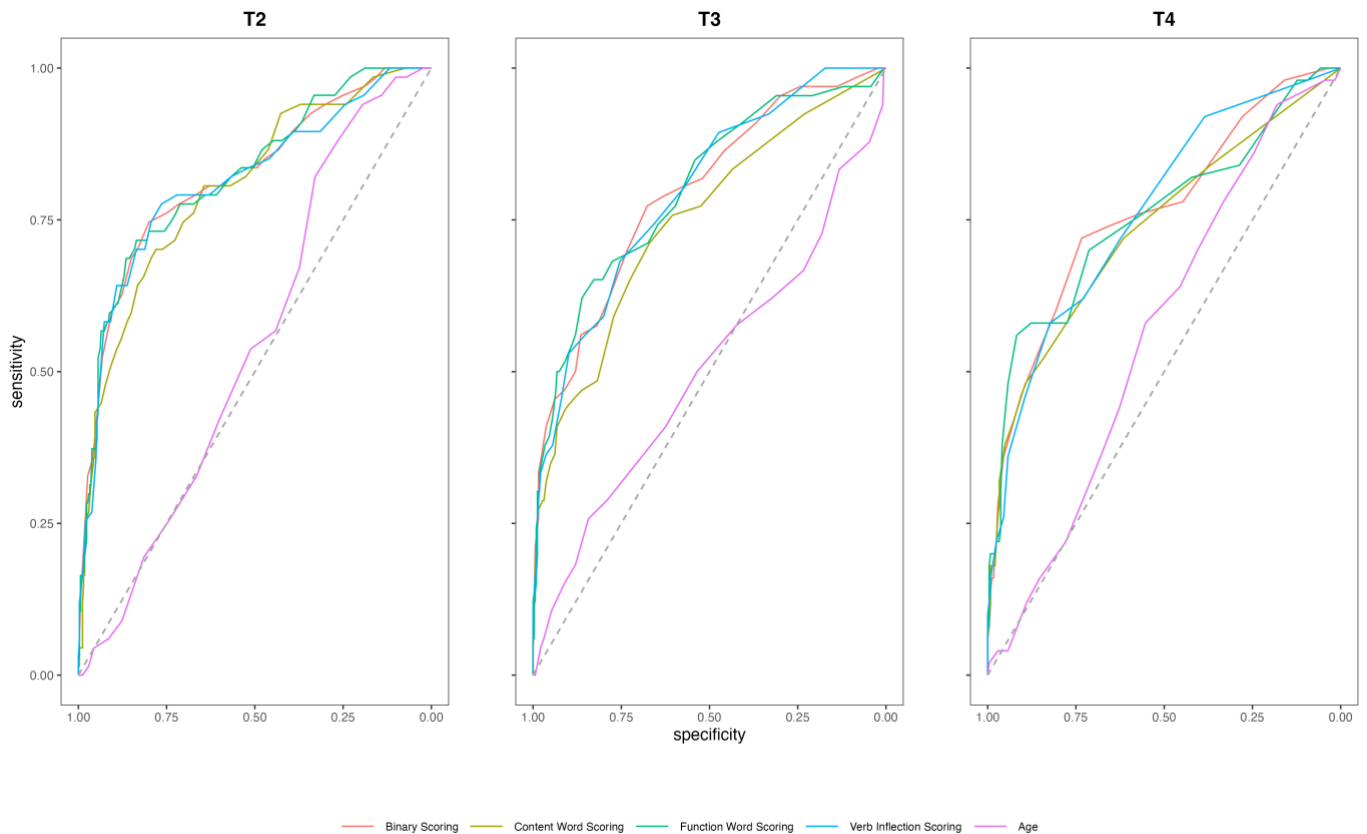
Table 5*Diagnostic Accuracy Metrics for Groups*

Time-Point	Method of Scoring	Optimal Cut-Off	AUC	Sensitivity	Specificity	LR+	LR-
T2	Binary	0.31	0.82	0.75	0.80	3.73	0.32
	Content Word	0.87	0.80	0.70	0.78	3.18	0.38
	Function Word	0.80	0.82	0.72	0.84	4.35	0.34
	Verb Inflection	0.68	0.81	0.78	0.76	3.30	0.29
T3	Binary	0.63	0.79	0.77	0.68	2.38	0.34
	Content Word	0.96	0.75	0.71	0.67	2.16	0.43
	Function Word	0.90	0.80	0.62	0.86	4.47	0.44
	Verb Inflection	0.84	0.80	0.68	0.75	2.76	0.42
T4	Binary	0.78	0.76	0.72	0.73	2.70	0.38
	Content Word	0.96	0.73	0.48	0.89	4.50	0.58
	Function Word	0.94	0.75	0.56	0.92	6.83	0.48
	Verb Inflection	0.89	0.76	0.58	0.82	3.29	0.51

Note. Optimal cut-off points refer to the proportion correct on the task. AUC = Area under the curve (calculated from receiver operator curves), LR+ = positive likelihood ratio, LR- = negative likelihood ratio.

Figure 2

ROC curves plotting sensitivity by specificity values for the diagnostic accuracy of the SR task for each group for each method of scoring



Note. Each column represents each group, being those matched at time point 2, 3, and 4 respectively.

Time-Point 2

At T2 (UK school year 1, ages five to six years), all scoring methods resulted in AUC values above 0.80, indicating that the SR task had an excellent overall ability to distinguish between children with and without DLD (using interpretation guidelines by Hosmer & Lemeshow, 2000). However, when examining individual diagnostic metrics, the picture was more mixed.

Sensitivity (correctly identifying DLD) and specificity (correctly identifying TD) values were above 0.70 for all scoring methods. No scoring method resulted in both values being

above the 0.80 threshold, typically considered “acceptable” for diagnostic use (Plante & Vance, 1994). Only specificity levels for binary scoring and function word scoring were above 0.80, showing “acceptable” levels of being able to correctly identify TD children.

Looking at likelihood ratios, all scoring methods resulted in LR+ values over 3.00, implying the SR task (at T2) was able to rule in DLD with moderate confidence (Dollaghan, 2007). Only verb inflection scoring resulted in a LR- score under 0.30 which indicates this scoring method was able to rule out DLD with moderate confidence.

To compare scoring method at T2, bootstrap tests were used to compare the AUCs of the ROC curves. No statistically significant difference was found between any of the scoring methods, indicating that the overall diagnostic accuracy of the SR task was not influenced by the choice of scoring.

Time-Point 3

At T3 (UK school year 3, ages seven to eight years), the SR task showed excellent overall discriminative ability for function word scoring and verb inflection scoring (AUCs above 0.80), and an acceptable overall discriminative ability for binary scoring and content word scoring (AUCs above 0.70).

Sensitivity and specificity values for all methods of scoring ranged from 0.62 to 0.86. Only the specificity for function word scoring was above 0.80 threshold, showing an “acceptable” level of identifying TD children. No pair of sensitivity and specificity were above this 0.80 level.

For likelihood ratios, only the LR+ value for function word scoring was above the 3.00, indicating a moderate confidence for ruling in DLD. No methods resulted in a LR- below 0.30 for ruling out DLD with moderate confidence.

Bootstrap comparisons of ROC curves for each scoring method revealed the AUC for content word scoring (AUC = 0.75) was significantly lower than that for binary scoring (AUC = 0.79; $D = -2.25$, $p = .025$), function word scoring (AUC = 0.80; $D = -2.68$, $p = .007$), and verb inflection scoring (AUC = 0.80, $D = -2.22$, $p = .027$). This suggests that content word scoring was less effective at distinguishing DLD and TD groups at T3.

Time-Point 4

At T4 (UK school year 6, ages ten to 11 years), all scoring methods resulted in AUCs above 0.70, indicating an acceptable discriminative ability of the SR task.

Sensitivity and specificity value for all methods of scoring ranged from 0.48 to 0.92. While content word scoring, function word scoring, and verb inflection scoring resulted in specificity values above 0.80 (an “acceptable” ability to identify TD), this came at a cost to sensitivity, which were all below 0.60 (a poor ability in identifying DLD). Binary scoring provided a better balance of sensitivity and specificity values, with both being above 0.70.

For likelihood ratios, content word scoring, function word scoring, and verb inflection scoring resulted in LR+ values above 3.00, but again this came as a cost to the corresponding LR- values.

No statistically significant difference was found between the AUCs for the different scoring methods at T4, suggesting that the overall level of diagnostic accuracy achieved by the SR task did not differ as a result of the scoring method.

Discussion

This study evaluated the effectiveness of SR in differentiating between English-speaking monolingual children with DLD and TD children in data from a large-scale cohort study. The diagnostic utility of SR tasks was assessed through the calculation of sensitivity, specificity, LRs, and AUCs. Comparisons between methods of scoring the SR task were also made to consider differences in effectiveness.

Group Differences in Performance

Our first research question asked whether children with DLD performed differently on the SASIT-E32 SR task compared to their TD peers. Consistent with our hypotheses, our findings revealed that children with DLD consistently performed with lower accuracy on the task across each time point. At T2 (ages five to six), those meeting criteria for DLD performed on average 10% to 25% worse on the task than their TD peers (depending on the scoring method). At T3 (ages seven to eight), this difference was 6% to 25% depending on the scoring method used. And at T4 (ages ten to 11), this difference had narrowed slightly to

4% to 15% depending on the scoring method used. This reduced difference between the groups at the later time points was likely a reflection of the groups approaching ceiling levels of performance, limiting between-group differentiation.

Crucially, Bayes factors demonstrated extreme evidence for the hypothesis of a difference in performance on the SASIT-E32 between the DLD and TD groups across all scoring methods and time points. This finding is in line with previous literature which suggests that children with DLD have difficulties with SR and will perform with lower accuracy than their peers within a targeted age group (e.g., Conti-Ramsden et al., 2001; Archibald & Joanisse, 2009; Pham & Ebert, 2020; Vang Christensen, 2019). Although the magnitude of the performance gap decreased over time, there remained strong evidence that the SASIT-E32 could consistently result in different performance between DLD and TD groups at each time point. This sustained difference over a six-year span supports the claim that SR performance varies as a result of DLD and this effect is generally robust.

We calculated effect sizes to enable direct comparisons to those reported in the review by Ward et al. (2024). Effect sizes in the present study ranged from -1.11 to -1.57, which, although smaller than the literature's average (-2.08), still reflect substantial differences in SR performance between DLD and TD groups. These differences in effect sizes could reflect differences in the nature of the DLD samples, as the present study utilised a population-based sample, whereas many studies included in the meta-analysis were drawn from clinical samples. However note that in the meta-analysis of Ward et al. (2024), there was no evidence that population samples had smaller effect sizes. It should though be noted that the Hedges' g method of effect size calculation assumes a normal distribution and equal variances (Turner & Bernard, 2006). Neither of these properties universally held for our data, with multiple instances of performance approaching ceiling. Therefore, while the calculation of effect sizes allows for easy comparison, there may be bias in its estimation. This supports our use of beta regression modelling, which is better suited for bounded, proportional data.

Diagnostic Accuracy

Our second research question examined whether SR tasks could effectively distinguish between children with DLD and their TD peers with acceptable diagnostic accuracy. At T2, the SASIT-E32 demonstrated excellent discriminative ability (AUC values between 0.80-0.82), and with sensitivity and specificity ranging between 0.72 and 0.84. At T3, there was a reduction in discriminative ability with AUC values ranging from 0.75 to 0.80 (acceptable to excellent level), and this reduced again by T4 with AUC values ranging from 0.73 to 0.76 (acceptable level). No scoring method achieved both sensitivity and specificity levels above 0.80 at any time point, a threshold often used as an “acceptable” clinical benchmark for diagnostic accuracy (Plante & Vance, 1994). Across all time points, specificity values tended to be higher than sensitivity, indicating the SASIT-E32 was more effective at correctly identifying TD children, than identifying children with DLD.

A similar mixed pattern emerged when examining likelihood ratios. At T2, LR+ values exceeded 3.00 across all scoring methods, indicating moderate confidence in the task’s ability to “rule in” DLD. However, only the verb inflection scoring method produced an LR- below 0.30, showing moderate confidence in the task’s ability to “rule out” DLD. This pattern only became stronger across time points, with LR+ values remaining relatively strong, but LR- values worsened.

While there is research demonstrating that SR can result in high diagnostic accuracy, such as the findings from Christensen and Hansson (2012), Hamann and Abed Ibrahim (2017), Tuller et al. (2018), and Wang et al. (2022), who all reported instances where both sensitivity and specificity values reached 1, the literature is generally mixed. Indeed, nearly 25% of studies reviewed by Ward et al. (2024) that reported such outcomes failed to achieve both sensitivity and specificity values above 0.80. Similarly, in their meta-analysis comparing marker tasks, Pawlowska (2014) noted that while LR for SR appeared “promising”, there was no conclusive evidence that SR can reliably diagnose DLD.

There is, then, precedent for the finding that SR falls short of the clinical diagnostic thresholds. Ward et al. (2024) found little evidence that task or experimental design systematically affects group-level performance differences (i.e., a difference in the size of the

performance between TD and DLD groups). However, the absence of evidence of an effect in that study does not allow us to rule out the possibility that design features might contribute to variability in diagnostic accuracy (which was not directly studied). For example, of the studies referenced to above, Christensen and Hansson (2012) and Wang et al. (2022) each presented original SR tasks developed for their studies. The first involving 18 sentences developed to assess the use of past tense with specific verbs, and the second 16 sentences, half with passives and half with aspect markers as core elements. In contrast, the SASIT-E32 used in the current study comprises 32 sentences spanning a broader range of morphosyntactic and syntactic structures. The SASIT-E32 was developed following principles from the LITMUS framework (Marinis & Armon-Lotem, 2015), similar to the SR tasks used by Hamann and Abed Ibrahim (2017) and Tuller et al. (2018). Such differences in item composition and linguistic complexity might be worth further exploration.

It is also the case that, of the studies that reported very high sensitivity and specificity, all but Wang et al. (2022) drew DLD participants from clinical samples, involving children with pre-existing diagnoses and often receiving language intervention. This is in contrast to the current study which used a population-based sample. These sampling and diagnostic differences could of course also contribute to the variation reported in diagnostic accuracy. However, as Ward et al. (2024) reported, there was no evidence that group-level performance differences (TD vs. DLD) were smaller in population samples than in clinical ones. Indeed, in some cases, population samples showed clearer group distinctions.

From a practical perspective, while we found clear and strong evidence of performance differences between the DLD and TD groups, it is important to understand that this does not necessarily translate into clinical utility. To reliably diagnose DLD, SR tasks like SASIT-E32 should be used in conjunction with other assessment tools and information sources (e.g., case history and observations from different contexts). These considerations align with the diagnostic criteria proposed by Bishop et al. (2016) in the CATALISE-2 consortium, which emphasise the importance of a comprehensive assessment approach, including the use of multiple data sources to diagnose DLD accurately. This does not mean

that SR tasks lack clinical value. We argue that SR tasks like the SASIT-E32 may serve as a useful complementary screening tool. Although sensitivity was lower than typically expected for screening purposes, the high specificity observed indicates that children performing below threshold were likely to represent true cases of DLD. This pattern suggests that the measure may work as an initial screening tool for identifying children who warrant further assessment, particularly in large-scale or resource-limited contexts where false positives could place undue strain on services. Nonetheless, the moderate sensitivity indicates that some children with DLD may not be detected, highlighting the need for future work to examine whether combining this measure with complementary indicators such as parent reports could enhance overall screening efficiency.

We observed a slight decline in diagnostic accuracy across the time points, with accuracy generally being strongest at T2 compared to T3 and then T4. This might suggest that clinically, the SASIT-E32 is more effective when used on younger children. The structures targeted by the SASIT-E32 may still be in the process of being acquired in early childhood but may be mastered by older children and even by those with DLD. As a result the task may become less sensitive to impairment. It is also possible that because scores on the task began to approach a ceiling level of performance as children got older, there was simply less room to differentiate performance between the groups.

This trend may be specific to the difficulty level of the SASIT-E32. A more challenging SR task containing more complex structures might maintain diagnostic accuracy in older children, and there is evidence that SR can be useful for identifying DLD in adults (e.g., Poll et al., 2016). This highlights the importance of aligning the linguistic complexity of a SR task with the target age group to achieve optimal diagnostic accuracy, particularly in contexts where standardised tests of age norms are no available. This pattern of decreasing diagnostic accuracy also reflects broader challenges in identifying reliable screening tools for school-age populations. For example, Bao et al. (2024) note the scarcity of tools with robust diagnostic accuracy for this age range. While our findings show moderate utility of SR tasks as a marker of impairment, the reduction in accuracy over time is consistent with the

challenges identified in the broader literature for maintaining diagnostic precision in older children. This emphasises the importance of further developing and refining SR tasks for this age group.

Methods of Scoring

Our third research question asked whether the different scoring methods used with the SASIT-E32 – binary scoring, function word scoring, content word scoring, and verb inflection scoring – would produce different outcomes. Across all timepoints, content word scoring yielded the highest levels of child SR performance (in terms of proportion correct) compared to the other methods. Performance on this measure approached ceiling across the time points, with mean proportion correct ranging from 92%-99% for TD groups and 80%-94% for DLD groups. In contrast, the lowest child SR performance levels were observed with the binary scoring method, with mean proportion correct ranging from 53%-85% for TD groups and 24%-70% for DLD groups. Despite these differences, all methods of scoring resulted in lower performance in the DLD groups compared to TD groups, and high Bayes factors indicated strong evidence for a group difference. Likewise, diagnostic accuracy did not vary meaningfully in most cases.

At T3, content word scoring resulted in a significantly lower AUC (0.75) than the other three scoring methods. Although not statistically significant, content word scoring also produced the lowest AUC at both T2 and T4. These findings align with Seef-Gabriel et al. (2010), who reported that both TD children and children with language impairments performed best on content word scoring when compared to function word and inflection scoring. They similarly observed the smallest group differences observed for content word scoring. This pattern may be due to children with DLD having more pronounced difficulties with function words and inflections (Chiat, 2001). However again, the ceiling effects observed with content word scoring especially, may have limited the task's ability to differentiate between groups for this method of scoring.

In practice, scoring on a detailed level (such as scoring function words/inflections) may not be practical, given the scoring time and training required. Binary scoring provides a

simpler alternative, being the quickest and easiest method to administer, and requiring little training to apply reliably (Ebert et al., 2019), and, as our results show, retains comparable diagnostic accuracy to more detailed methods. Indeed, in the literature, binary scoring has often been recommended as being most useful for application within a clinical context (e.g., Pham & Ebert, 2020; Taha et al., 2021), due to its efficiency and scalability. In contrast, detailed scoring has been used and recommended in studies aiming to detect specific linguistic deficits or to gain finer-grained profiles of children's language abilities (e.g., Hamann & Abed Ibrahim, 2017; Christensen & Hansson, 2012). Although a full cost–benefit analysis comparing the feasibility of different scoring approaches would be valuable, such data are not currently available. Nonetheless, these distinctions do reinforce that different scoring methods may serve different goals, and the best approach depends on whether the priority is broad screening, or research precision.

It is important to note that the task of picking a scoring method by weighing up the additional time and training against linguistic insight gained, may soon be reframed by technological advances. Progress in automated speech recognition and natural language processing means it may soon be feasible to automate the scoring of SR tasks. Automated scoring systems are already being developed and trialled (e.g., Liu et al., 2024, Albudoor & Peña, 2022), and show promise for producing reliable, fine-grained scoring outputs with minimal input from clinicians. As these tools become more widely available and accurate, the time and training barriers currently limiting the use of detailed scoring may be significantly reduced. In this context, more detailed scoring may become a viable and efficient option in the future, and thus any performance advantages they offer may become more easily attainable.

Strengths and Limitations

The current study involved an analysis of the SCALES dataset (Norbury, 2022), which allowed for not only a large sample of TD children but, because there was an oversampling at those “at risk” for language disorder, a large DLD sample. It also allowed for a one-gate design to be used, whereby children identified as meeting criteria for DLD came

from the same sample as those who were TD, which Dollaghan and Horner (2011) note is important for studies assessing diagnostic accuracy. Additionally, all children were recruited from schools within the same authority, which may have offered some consistency in language exposure.

However, the SCALES sample, in sampling school pupils from Surrey, was likely more affluent than the national average (Norbury et al., 2016). Within the work by Norbury et al. (2016), a significant association was found between socioeconomic background and language disorder (classified using the methods described earlier in this paper), with children from lower socioeconomic backgrounds being more likely to meet criteria for DLD. The underrepresentation of lower socioeconomic neighbourhoods in the sample limits the generalisability of the current findings.

Finally, while using primarily the same core composite measures and classification criteria as the SCALES project, classification of DLD in the current study differed slightly in that the SASIT-E32 was excluded. In contrast, the original SCALES classification included the SASIT-E32 as one of the core measures. This reclassification ensured that SR performance was not part of the criterion, but it may have led to small discrepancies in group membership relative to the original SCALES groupings (e.g., some children misclassified as TD or DLD).

A strength of the current study was the use of a Bayesian analytical framework, which allowed us to assess the strength of evidence for both the presence and absence of effects. This being important in language research, where small sample sizes and reliance on null-hypothesis significance testing can make it difficult to distinguish true absence of an effect from insufficient statistical power. The Bayesian approach also provides an informative measure of uncertainty by estimating the probability of different effect sizes given the data.

Implications

These findings have important implications for both research and clinical practice. SR tasks such as the SASIT-E32 can provide a valid and flexible tool for detecting group-level differences in language ability. Our results showed consistent performance differences

between DLD and TD groups across multiple scoring methods, supporting SR as a promising marker of language difficulty. It reinforced binary scoring as a practical and valid option clinically and in research.

Clinically, while the SASIT-E32 did not meet thresholds for standalone diagnosis, its strong specificity and LR+ values do suggest value as a screening tool. The task was most accurate in younger children, highlighting its potential role in early identification. Given its efficiency and potential for automation in the future, SR may serve as a useful component of a broader assessment battery in both research and applied settings.

Conclusions

This study provides promising evidence that SR tasks, specifically the SASIT-E32, consistently differentiate performance between English-speaking children with DLD and TD children. While the task's diagnostic accuracy did not consistently meet clinical thresholds for standalone use, its efficiency, ease of administration, and adaptability make it a valuable tool in both research and screening contexts. The flexibility to select scoring methods based on specific needs, along with the potential to automate scoring, further enhances the adaptability and utility of SR as a practical and scalable approach to assessing language ability.

Acknowledgements

This paper was completed as parts of L Ward's PhD research at The University of Manchester. Sponsorship and funding for the PhD and all resulting publications is provided by the UKRI Economic and Social Research Council (ESRC). Funding for SCALES was provided by Wellcome Trust [WT094836AIA] and ESRC [ES/R003041/1] grants to C Norbury.

Data Availability Statement

The dataset is publicly available (<http://doi.org/10.5255/UKDA-SN-8968-1>). A UK Data Service account is required to download the dataset. Some variables are not archived, and the weighted z-scores for each core language battery measure for T2, T3, and T4 were provided by C Norbury.

References

- Adams, C., Cooke, R., Crutchley, A., Hesketh, A., & Reeves, D. (2001). *Assessment of comprehension and expression 6-11*. GL assessment.
- Albudoor, N., & Peña, E. D. (2022). Identifying language disorder in bilingual children using automatic speech recognition. *Journal of Speech, Language, and Hearing Research*, 65(7), 2648-2661.
- Archibald, L. M., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research*, 52(4), 899-914. [https://doi.org/10.1044/1092-4388\(2009/08-0099\)](https://doi.org/10.1044/1092-4388(2009/08-0099))
- Armon-Lotem, S., & Meir, N. (2016). Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: Evidence from Russian and Hebrew. *International journal of language & communication disorders*, 51(6), 715-731. <https://doi.org/10.1111/1460-6984.12242>
- Bao, X., Komesidou, R., & Hogan, T. P. (2024). A Review of Screeners to Identify Risk of Developmental Language Disorder. *American Journal of Speech-Language Pathology*, 33(3), 1548-1571. <https://doi.org/10.1016/j.xpro.2024.103418>
- Bishop, D.V.M. (2003a). *Children's communication checklist-2*. Pearson.
- Bishop, D.V.M. (2003b). *Test for reception of grammar (TROG-2) (2nd edn)*. Pearson Assessment.
- Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Catalise Consortium. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS one*, 11(7), e0158753. <https://doi.org/10.1371/journal.pone.0158753>
- Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Klee, T. M. (2017). CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development. Phase 2. *Journal of Child Psychology and Psychiatry*, 58(10), 1068-1080. <https://doi.org/10.1111/jcpp.12721>

- 879 Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal*
880 *of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>.
- 881 Chiat, S. (2001). Mapping theories of developmental language impairment: Premises,
882 predictions and evidence. *Language and cognitive processes*, 16(2-3), 113-142.
883 <https://doi.org/10.1080/01690960042000012>
- 884 Chiat, S., & Roy, P. (2013). Early predictors of language and social communication
885 impairments at ages 9–11 years: A follow-up study of early-referred children. *Journal*
886 *of Speech, Language, and Hearing Research*, 56(6), 1824-1836.
- 887 Christensen, R. V., & Hansson, K. (2012). The use and productivity of past tense
888 morphology in specific language impairment: An examination of Danish. *Journal of*
889 *Speech, Language, and Hearing Research*, 55(6), 1671-1689.
890 [https://doi.org/10.1044/1092-4388\(2012/10-0350\)](https://doi.org/10.1044/1092-4388(2012/10-0350))
- 891 Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific
892 language impairment (SLI). *Journal of child psychology and psychiatry*, 42(6), 741-
893 748. <https://doi.org/10.1111/1469-7610.00770>
- 894 Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in*
895 *psychology*, 5, 781. <https://doi.org/10.3389/fpsyg.2014.00781>
- 896 Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication*
897 *disorders*. Paul H. Brookes Publishing Co.
- 898 Dollaghan, C. A., & Horner, E. A. (2011). Bilingual language assessment: A meta-analysis of
899 diagnostic accuracy. *Journal of Speech, Language, and Hearing Research*, 54(4),
900 1077-1088.
- 901 Durkin, K., & Conti-Ramsden, G. (2007). Language, social behaviour, and the quality of
902 friendships in adolescents with and without a history of specific language impairment.
903 *Child development*, 78(5), 1441-1457. [https://doi.org/10.1111/j.1467-](https://doi.org/10.1111/j.1467-8624.2007.01076.x)
904 [8624.2007.01076.x](https://doi.org/10.1111/j.1467-8624.2007.01076.x)

- 905 Ebert, K. D., Rak, D., Slawny, C. M., & Fogg, L. (2019). Attention in bilingual children with
906 developmental language disorder. *Journal of Speech, Language, and Hearing*
907 *Research*, 62(4), 979-992. https://doi.org/10.1044/2018_JSLHR-L-18-0221
- 908 Fleckstein, A., Prévost, P., Tuller, L., Sizaret, E., & Zebib, R. (2018). How to identify SLI in
909 bilingual children: A study on sentence repetition in French. *Language Acquisition*,
910 25(1), 85-101. <https://doi.org/10.1080/10489223.2016.1192635>
- 911 Hamann, C., & Abed Ibrahim, L. (2017). Methods for identifying specific language
912 impairment in bilingual populations in Germany. *Frontiers in communication*, 2, 16.
913 <https://doi.org/10.3389/fcomm.2017.00016>
- 914 Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related
915 estimators. *Journal of Educational Statistics*, 6(2), 107-128.
916 <https://doi.org/10.3102/10769986006002107>
- 917 Hosmer, D. W., Jr., & S. Lemeshow (2000). *Applied Logistic Regression* (2nd ed). Wiley.
- 918 Jeffreys, H. (1961) *Theory of Probability* (3rd Ed.). Clarendon Press.
- 919 Kaçar-Kütükçü, D., & Topbaş, S. (2024). LITMUS Turkish sentence repetition test: The best
920 items, effect of scoring and diagnostic accuracy. *Applied Neuropsychology: Child*, 1-
921 14. <https://doi.org/10.1080/21622965.2024.2400483>
- 922 Komeili, M., & Marshall, C. R. (2013). Sentence repetition as a measure of morphosyntax in
923 monolingual and bilingual children. *Clinical linguistics & phonetics*, 27(2), 152-162.
924 <https://doi.org/10.3109/02699206.2012.751625>
- 925 Leclercq, A. L., Quémart, P., Magis, D., & Maillart, C. (2014). The sentence repetition task: A
926 powerful diagnostic tool for French children with specific language impairment.
927 *Research in developmental disabilities*, 35(12), 3423-3430.
928 <https://doi.org/10.1016/j.ridd.2014.08.026>
- 929 Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*.
930 Cambridge University Press. <https://doi.org/10.1017/CBO9781139087759>
- 931 Leonard, L. B. (2017). *Children with specific language impairment*. MIT press.

- 932 Liu, D., Yang, J., Albrecht-Buehler, I., Qin, H., Li, S., Hu, Y., Nassereldine, A., & Xiong, J.
 933 (2024). Automatic Screening for Children with Speech Disorder Using Automatic
 934 Speech Recognition: Opportunities and Challenges. *Proceedings of the AAAI*
 935 *Symposium Series*, 4(1), 308-313.
- 936 Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. *Assessing multilingual children:*
 937 *Disentangling bilingualism from language impairment*, 13, 95-124.
- 938 Marinis, T., Chiat, S., Armon-Lotem, S., Piper, J., & Roy, P. (2011). *School-Age Sentence*
 939 *Imitation Test-English 32 (SASITE32)*. Unpublished test.
- 940 Martin, N. A., & Brownell, R. (2011). *Expressive one-word picture vocabulary test-4*
 941 *(EOWPVT-4)*. Academic Therapy Publications.
- 942 McGregor, K. K. (2020). How we fail children with developmental language disorder.
 943 *Language, speech, and hearing services in schools*, 51(4), 981-992.
 944 https://doi.org/10.1044/2020_LSHSS-20-00003
- 945 Norbury, C. (2022). *Surrey Communication and Language in Education Study: Intensive*
 946 *Data T2-T5, 2012-2020*. [data collection]. UK Data Service. SN: 8968.
 947 <http://doi.org/10.5255/UKDA-SN-8968-1>
- 948 Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., ... & Pickles, A.
 949 (2016). The impact of nonverbal ability on prevalence and clinical presentation of
 950 language disorder: Evidence from a population study. *Journal of child psychology*
 951 *and psychiatry*, 57(11), 1247-1257. <https://doi.org/10.1111/jcpp.12573>
- 952 Pawłowska, M. (2014). Evaluation of three proposed markers for language impairment in
 953 English: A meta-analysis of diagnostic accuracy studies. *Journal of Speech,*
 954 *Language, and Hearing Research*, 57(6), 2261-2273.
 955 https://doi.org/10.1044/2014_JSLHR-L-13-0189
- 956 Pham, G., & Ebert, K. D. (2020). Diagnostic accuracy of sentence repetition and nonword
 957 repetition for developmental language disorder in Vietnamese. *Journal of Speech,*
 958 *Language, and Hearing Research*, 63(5), 1521-1536.
 959 https://doi.org/10.1044/2020_JSLHR-19-00366

- 960 Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based
961 approach. *Language, Speech, and Hearing Services in Schools*, 25(1), 15-24.
962 <https://doi.org/10.1044/0161-1461.2501.15>
- 963 Poll, G. H., Miller, C. A., & Van Hell, J. G. (2016). Sentence repetition accuracy in adults with
964 developmental language impairment: Interactions of participant capacities and
965 sentence structures. *Journal of Speech, Language, and Hearing Research*, 59(2),
966 302-316.
- 967 R Core Team. (2023). R: A language and environment for statistical computing [Software]. R
968 Foundation for Statistical Computing. <https://www.R-project.org/>
- 969 Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011).
970 pROC: an open-source package for R and S+ to analyze and compare ROC curves.
971 *BMC Bioinformatics*, 12, 77.
- 972 Rujas, I., Mariscal, S., Murillo, E., & Lázaro, M. (2021). Sentence repetition tasks to detect
973 and prevent language difficulties: A scoping review. *Children*, 8(7), 578.
974 <https://doi.org/10.3390/children8070578>
- 975 Seeff-Gabriel, B., Chiat, S., & Dodd, B. (2010). Sentence imitation as a tool in identifying
976 expressive morphosyntactic difficulties in children with severe speech difficulties.
977 *International Journal of Language & Communication Disorders*, 45(6), 691-702.
978 <https://doi.org/10.3109/13682820903509432>
- 979 Taha, J., Stojanovik, V., & Pagnamenta, E. (2021). Sentence repetition as a clinical marker
980 of developmental language disorder: Evidence from Arabic. *Journal of Speech,*
981 *Language, and Hearing Research*, 64(12), 4876-4899.
982 https://doi.org/10.1044/2021_JSLHR-21-00244
- 983 Theodorou, E., Kambanaros, M., & Grohmann, K. K. (2017). Sentence repetition as a tool for
984 screening morphosyntactic abilities of bilingual children with SLI. *Frontiers in*
985 *psychology*, 8, 2104. <https://doi.org/10.3389/fpsyg.2017.02104>

- 986 Thiele, C., & Hirschfeld, G. (2021). cutpointr: Improved Estimation and Validation of Optimal
987 Cutpoints in R. *Journal of Statistical Software*, 98(11), 1–27.
988 <https://doi.org/10.18637/jss.v098.i11>.
- 989 Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prevost, P., ... & Zebib, R. (2018).
990 Identifying language impairment in bilingual children in France and in Germany.
991 *International journal of language & communication disorders*, 53(4), 888-904.
992 <https://doi.org/10.1111/1460-6984.12397>
- 993 Turner, H. M. I., & Bernard, R. M. (2006). Calculating and synthesizing effect sizes.
994 *Contemporary issues in communication science and disorders*, 33, 42-55.
995 https://doi.org/10.1044/cicsd_33_S_42
- 996 Vamvakas, G., Norbury, C. F., Vitoratou, S., Gooch, D., & Pickles, A. (2019). Standardizing
997 test scores for a target population: The LMS method illustrated using language
998 measures from the SCALES project. *PLoS One*, 14(3), e0213492.
999 <https://doi.org/10.1371/journal.pone.0213492>
- 1000 Vang Christensen, R. (2019). Sentence repetition: A clinical marker for developmental
1001 language disorder in Danish. *Journal of Speech, Language, and Hearing Research*,
1002 62(12), 4450-4463. https://doi.org/10.1044/2019_JSLHR-L-18-0327
- 1003 Wang, D., Zheng, L., Lin, Y., Zhang, Y., & Sheng, L. (2022). Sentence repetition as a clinical
1004 marker for Mandarin-speaking preschoolers with developmental language disorder.
1005 *Journal of Speech, Language, and Hearing Research*, 65(4), 1543-1560.
1006 https://doi.org/10.1044/2021_JSLHR-21-00401
- 1007 Ward, L., Polišenská, K., & Bannard, C. (2024). Sentence Repetition as a Diagnostic Tool for
1008 Developmental Language Disorder: A Systematic Review and Meta-Analysis. *Journal*
1009 *of Speech, Language, and Hearing Research*, 67(7), 2191-2221.
1010 https://doi.org/10.1044/2024_JSLHR-23-00490
- 1011 Wechsler, D. (2003). *Wechsler preschool and primary scales of intelligence – Third UK*
1012 *edition (3rd edn)*. Pearson Assessment.
- 1013

Appendix A

This appendix outlines the criteria set out by Lee and Wagenmakers (2013; adjusted from Jeffreys, 1961) for the interpretation of Bayes Factor values used within the current study.

Table A

Classification system from Lee and Wagenmakers (2013; adjusted from Jeffreys, 1961) for the interpretation of Bayes Factor values

BF 10 Value	Recommended Interpretation
> 100	Extreme evidence for H_1
30 – 100	Very strong evidence for H_1
10 – 30	Strong evidence for H_1
3 – 10	Moderate evidence for H_1
1 – 3	Anecdotal evidence for H_1
1	No evidence for either hypothesis
1/3 – 1	Anecdotal evidence for H_0
1/3 – 1/10	Moderate evidence for H_0
1/10 – 1/30	Strong evidence for H_0
1/30 – 1/100	Very strong evidence for H_0

Note. BF 10 = Bayes factor value giving the evidence for H_1 over H_0 , H_1 = Experimental hypothesis, H_0 = Null hypothesis

Appendix B

This appendix provides an overview of all the posterior parameters from each of the Bayesian beta regression models fit to compare SR performance between all the age-matched groups across the four scoring methods. Raw parameters are provided for reproducibility purposes.

Table B

Summary of raw posterior parameters outputted from Bayesian beta regression models fit to compare SR performance between all the age-matched groups across the four scoring methods

Parameter	Time-Point 2			
	Binary Scoring	Content Word Scoring	Function Word Scoring	Verb Inflection Scoring
Intercept	-0.93 [-1.19, -0.65]	1.45 [1.26, 1.64]	0.92 [0.72, 1.12]	0.18 [-0.03, 0.40]
diagnosisTD	1.06 [0.77, 1.34]	0.91 [0.69, 1.13]	1.10 [0.88, 1.31]	1.06 [0.83, 1.31]
age	0.23 [0.14, 0.32]	0.18 [0.10, 0.26]	0.19 [0.12, 0.26]	0.21 [0.14, 0.29]
zoi_Intercept	-2.04 [-2.82, -1.38]	-6.43 [-11.15, -3.45]	-7.14 [-12.93, -3.62]	-6.85 [-12.21, -3.56]
zoi_diagnosisTD	-2.55 [-3.90, -1.30]	3.90 [0.90, 8.73]	2.13 [-1.67, 7.91]	2.46 [-1.01, 7.87]
coi_Intercept	-10.64 [-27.57, -1.89]			
coi_diagnosisTD	10.62 [0.94, 28.34]			
phi_Intercept	1.23 [0.87, 1.56]	2.19 [1.84, 2.50]	1.82 [1.49, 2.13]	1.41 [1.09, 1.69]
phi_diagnosisTD	0.14 [-0.21, 0.52]	0.48 [0.12, 0.86]	2.13 [-1.67, 7.91]	0.62 [0.29, 0.97]
Parameter	Time-Point 3			
	Binary Scoring	Content Word Scoring	Function Word Scoring	Verb Inflection Scoring
Intercept	-0.17 [-0.40, 0.07]	2.19 [1.95, 2.41]	1.76 [1.56, 1.97]	1.04 [0.83, 1.23]
diagnosisTD	1.07 [0.82, 1.32]	0.93 [0.68, 1.20]	1.04 [0.82, 1.26]	1.03 [0.81, 1.25]
age	0.09 [0.01, 0.17]	0.10 [0.01, 0.19]	0.08 [0.01, 0.14]	0.07 [-0.01, 0.14]
zoi_Intercept	-7.17 [-12.83, -3.65]	-2.54 [-3.51, -1.72]	-7.02 [-12.51, -3.65]	-6.54 [-11.56, -3.50]
zoi_diagnosisTD	1.63 [-2.24, 7.16]	1.35 [0.46, 2.37]	2.04 [-1.52, 7.55]	3.58 [0.50, 8.57]

phi_Intercept	1.21 [0.90, 1.49]	2.54 [2.13, 2.90]	2.38 [2.02, 2.72]	1.81 [1.46, 2.12]
phi_diagnosisTD	0.66 [0.34, 1.02]	0.90 [0.49, 1.35]	1.22 [0.85, 1.63]	0.97 [0.60, 1.36]
Time-Point 4				
Parameter	Binary Scoring	Content Word Scoring	Function Word Scoring	Verb Inflection Scoring
Intercept	0.82 [0.58, 1.07]	2.62 [2.30, 2.93]	2.47 [2.22, 2.71]	1.78 [1.53, 2.02]
diagnosisTD	0.89 [0.63, 1.15]	1.09 [0.77, 1.44]	0.90 [0.65, 1.17]	0.88 [0.62, 1.14]
age	0.05 [-0.03, 0.13]	0.06 [-0.04, 0.15]	0.04 [-0.03, 0.10]	0.04 [-0.03, 0.12]
zoi_Intercept	-6.38 [-11.16, -3.21]	-1.69 [-2.47, -0.95]	-6.27 [-11.36, -3.13]	-4.07 [-6.47, -2.45]
zoi_diagnosisTD	3.02 [-0.24, 7.83]	1.15 [0.35, 2.02]	3.48 [0.32, 8.57]	1.82 [0.14, 4.25]
phi_Intercept	1.60 [1.22, 1.98]	2.69 [2.19, 3.15]	2.80 [2.36, 3.19]	2.32 [1.89, 2.71]
phi_diagnosisTD	0.92 [0.51, 1.34]	1.60 [1.08, 2.15]	1.81 [1.37, 2.27]	1.12 [0.70, 1.58]

1041 *Note.* Values show estimates of the parameters surrounded by 95% confidence intervals.

1042 The parameters resulting from a model depend on whether a beta regression was fit, or
 1043 whether a zero-, one- or zero-one- inflated beta regression was fit. Intercepts represent the
 1044 central tendencies (on the link scale) of the main response variable (proportion correct),
 1045 zero-one inflation (zoi), continuous-one inflation (coi), zero-inflation (zi), or precision (phi).
 1046 Covariate effects represent the effects of predictors (diagnosis (TD/DLD) or age) on the
 1047 corresponding intercepts. Rhat values for all parameters was 1.00.

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

SUPPLEMENTARY MATERIAL S1

This material shows the results of a sensitivity analysis. For the main analysis, children who were found to meet criteria for DLD at a timepoint (T2, T3, or T4), were included in the DLD grouping across each time point. This meaning that a child, for example, who met criteria for DLD at T4 only, would still be included in the DLD grouping for the T2 and T3 analysis.

In evaluating this research decision, we perform the analyses for answering RQ1-3 on age-matched TD and DLD groups, where the DLD grouping was defined by each time point separately. For example, a child meeting criterion for DLD at T4 would only be included in the DLD grouping at T4 only. This was to assess whether this DLD grouping resulted in different conclusions to those outlined in the main text. Based on the following analysis, we find no indication that this decision influenced the reported results.

Table S1-1

Table comparing the sample size of those identified as meeting criteria for DLD in the main analysis and in the sensitivity analysis

Time-Point	Main Analysis		Sensitivity analysis	
	DLD	TD	DLD	TD
T2	67	340	32	375
T3	66	325	44	347
T4	50	244	28	266

Table S1-2

Results of the Bayesian beta regression models fit to compare the difference in performance between the TD and DLD group

Time Point	Scoring	Estimate of Difference	BF 10
T2	Binary Scoring	-0.24	1.58×10^7
	Content Word Scoring	-0.13	3.16×10^6
	Function Word Scoring	-0.19	2.94×10^9
	Verb Inflection Scoring	-0.26	1.38×10^9
T3	Binary Scoring	-0.31	6.10×10^{18}
	Content Word Scoring	-0.06	1.03×10^{11}
	Function Word Scoring	-0.11	4.99×10^{20}
	Verb inflection Scoring	-0.18	1.91×10^{18}
T4	Binary Scoring	-0.16	1.06×10^6
	Content Word Scoring	-0.05	1.98×10^7
	Function Word Scoring	-0.06	9.62×10^{15}
	Verb inflection Scoring	-0.09	2.90×10^7

Notes. The estimate of difference refers to the average marginal effect on SR performance from being identified as DLD compared to TD. BF 10 = Bayes factor value giving the evidence for H_1 over H_0 .

1101 **Table S1-3**1102 *Diagnostic Accuracy Metrics for each Method of Scoring for the groups*

Time Point	Method of Scoring	Optimal Cut-Off	AUC	Sensitivity	Specificity	LR+	LR-
T2	Binary (/32)	10	0.81	0.75	0.75	2.99	0.33
	Content Word (/121)	110	0.81	0.88	0.61	2.23	0.21
	Function Word (/176)	127	0.81	0.62	0.89	5.86	0.42
	Verb Inflection (/44)	30	0.82	0.81	0.72	2.87	0.26
T3	Binary (/32)	20	0.85	0.89	0.66	2.61	0.17
	Content Word (/121)	116	0.81	0.82	0.66	2.41	0.28
	Function Word (/176)	160	0.85	0.77	0.81	4.10	0.28
	Verb Inflection (/44)	35	0.83	0.66	0.84	4.06	0.41
T4	Binary (/32)	25	0.74	0.75	0.70	2.48	0.36
	Content Word (/121)	117	0.74	0.57	0.82	3.14	0.52
	Function Word (/176)	166	0.75	0.61	0.88	5.22	0.44
	Verb Inflection (/44)	39	0.74	0.61	0.79	2.85	0.50

1103 *Note.* Optimal cut-off points refer to raw scores of performances. AUC = Area under the
1104 curve (calculated from receiver operator curves), LR+ = positive likelihood ratio, LR- =
1105 negative likelihood ratio.

1106

1107

1108

1109

1110

1111

1112

1113

1114

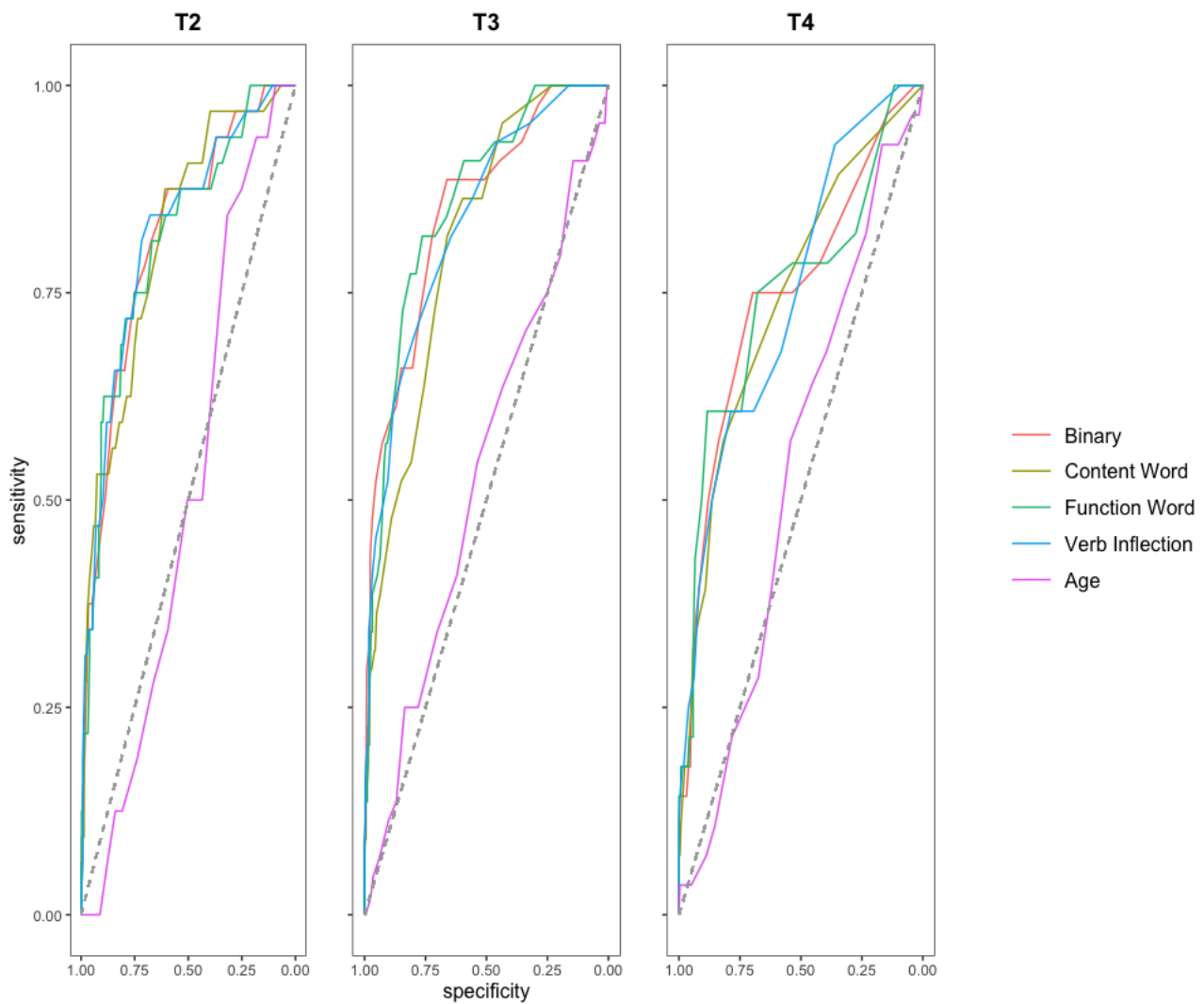
1115

1116

1117

Figure S1-1

ROC curves plotting sensitivity by specificity values for the diagnostic accuracy of the SR task for each group and for each method of scoring



Note. Each column represents each time-point, being those matched at time point 2, 3, and 4 respectively.

1130 **Supplementary Material S2**

1131 This material provides an overview of parameter convergence and posterior predictive
1132 checks of the Bayesian beta regression models.

1133

1134 **Figure S2-1**

1135 *Time Series Plots of Markov Chains showing model parameter convergence for models fit*
1136 *comparing groups at T2 for each method of scoring*

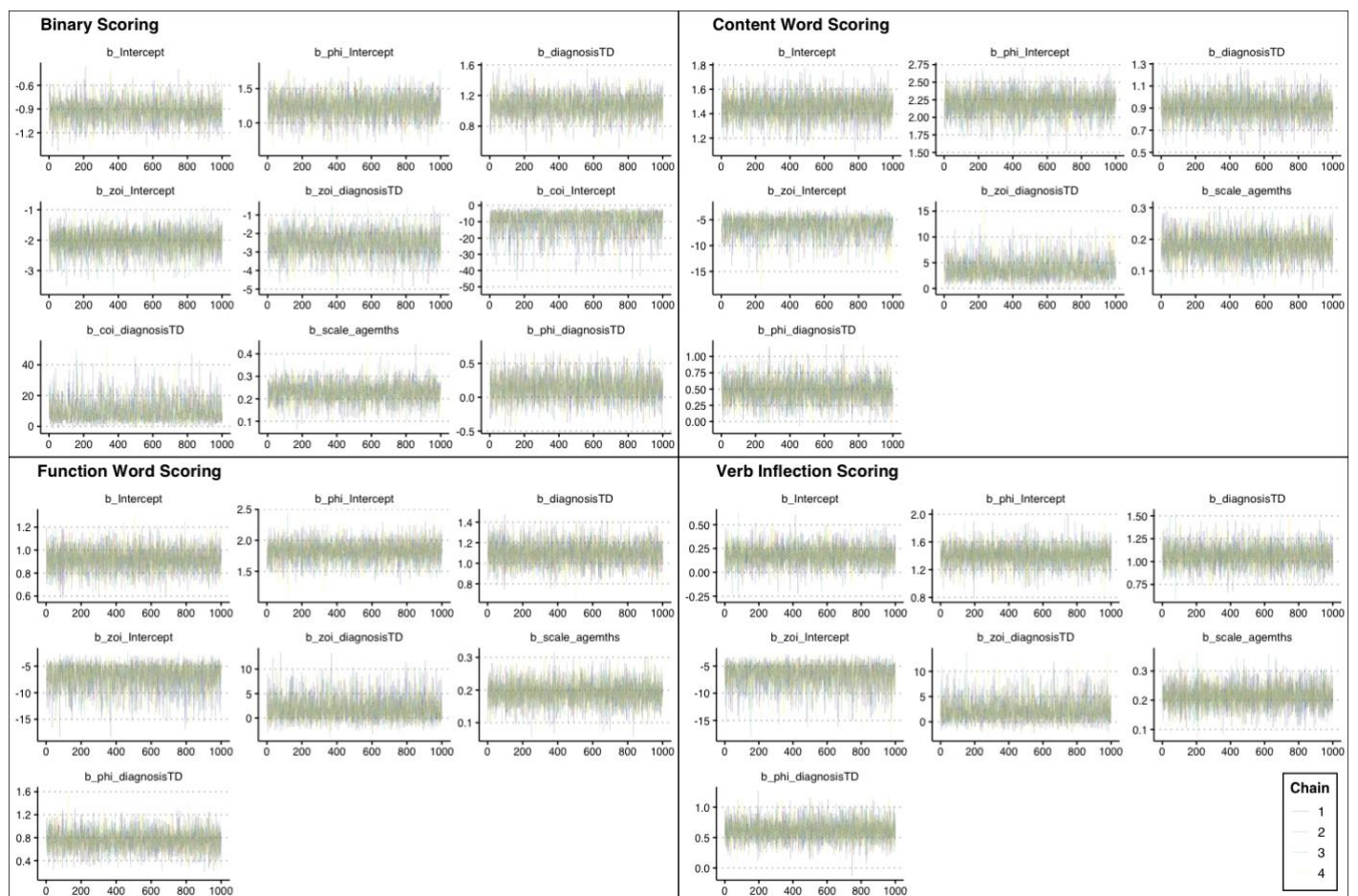
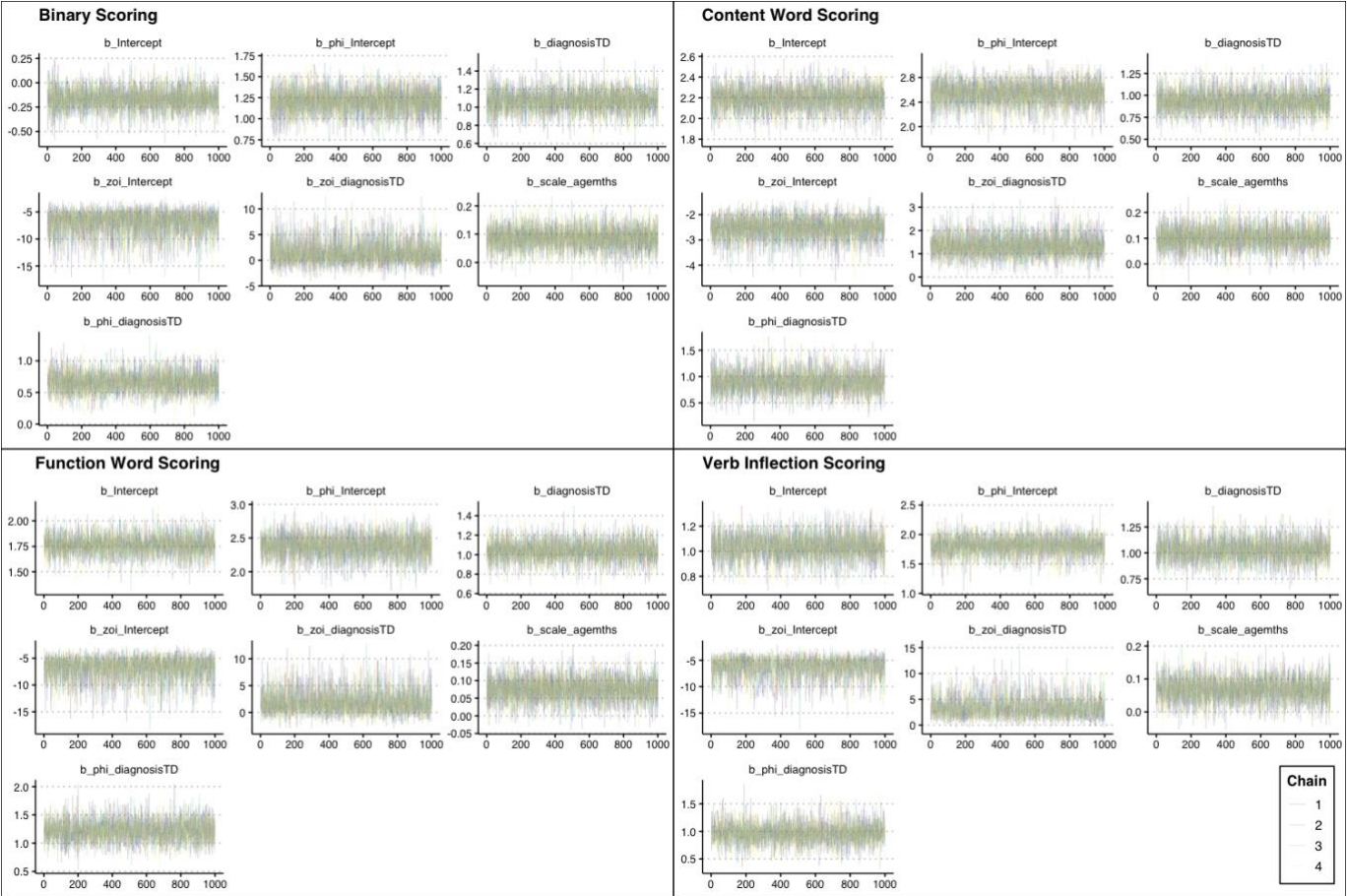


Figure S2-2

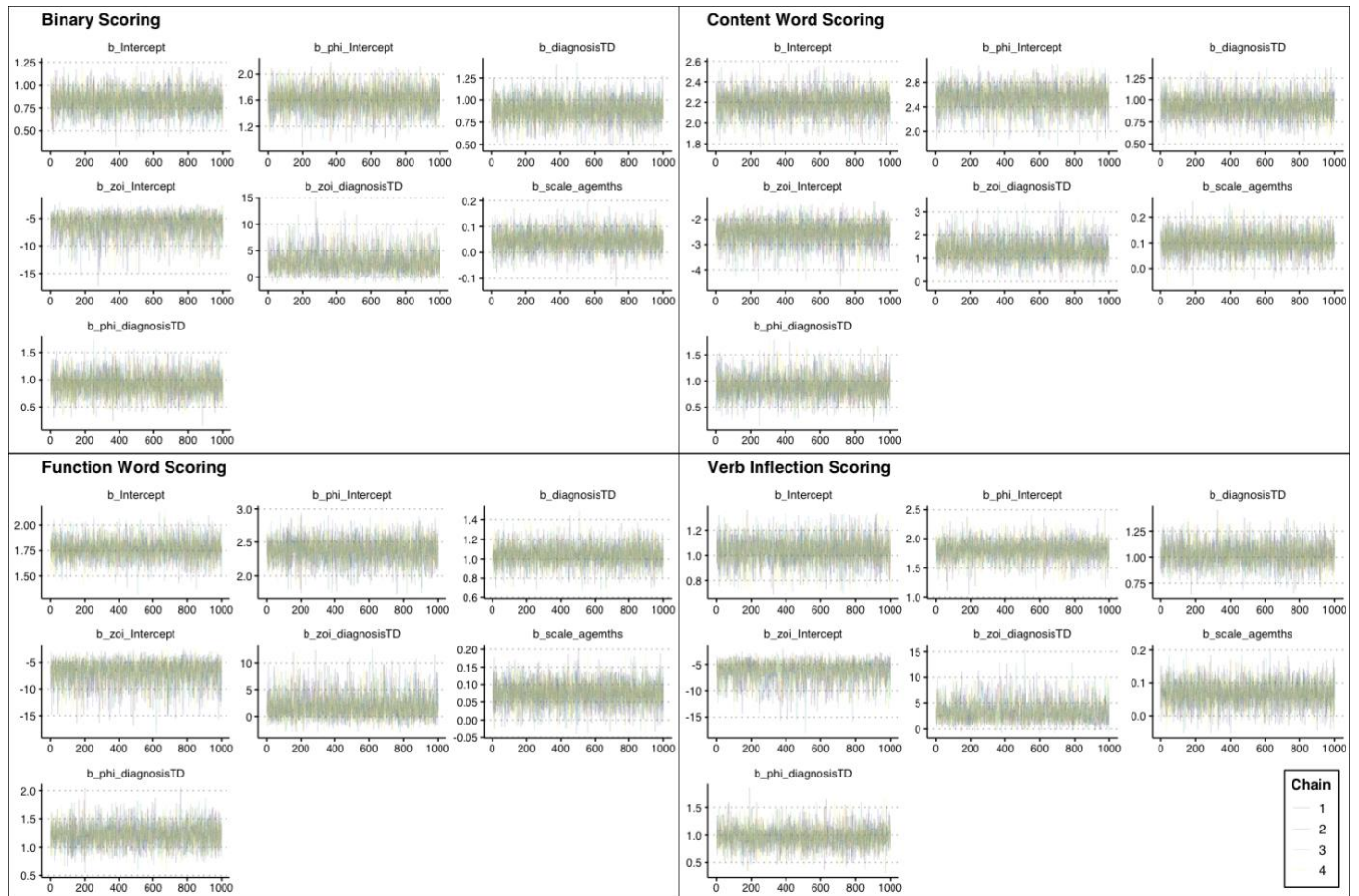
Time Series Plots of Markov Chains showing model parameter convergence for models fit comparing groups at T3 for each method of scoring



1158 **Figure S2-3**

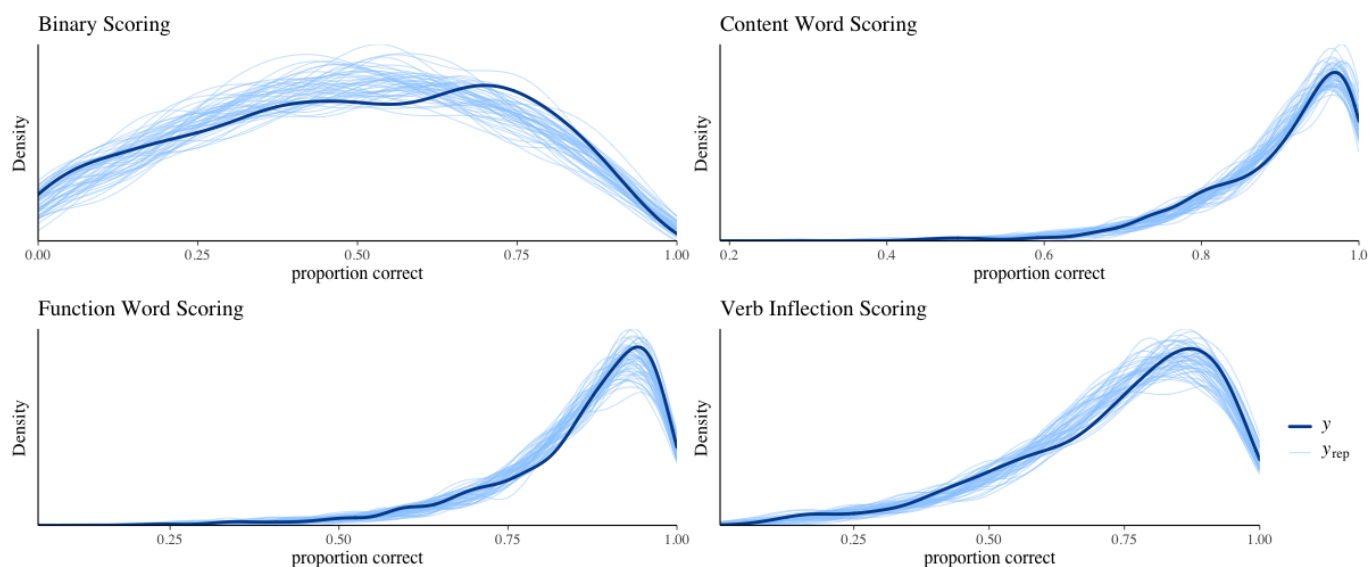
1159 *Time Series Plots of Markov Chains showing model parameter convergence for models fit*

1160 *comparing groups at T4 for each method of scoring*



1172 **Figure S2-7**

1173 *Posterior Predictive Checks comparing the distribution of the dependent variable (proportion*
 1174 *correct) to the distribution of simulated datasets for models comparing performance of*
 1175 *groups at T2 for each method of scoring SR productions.*

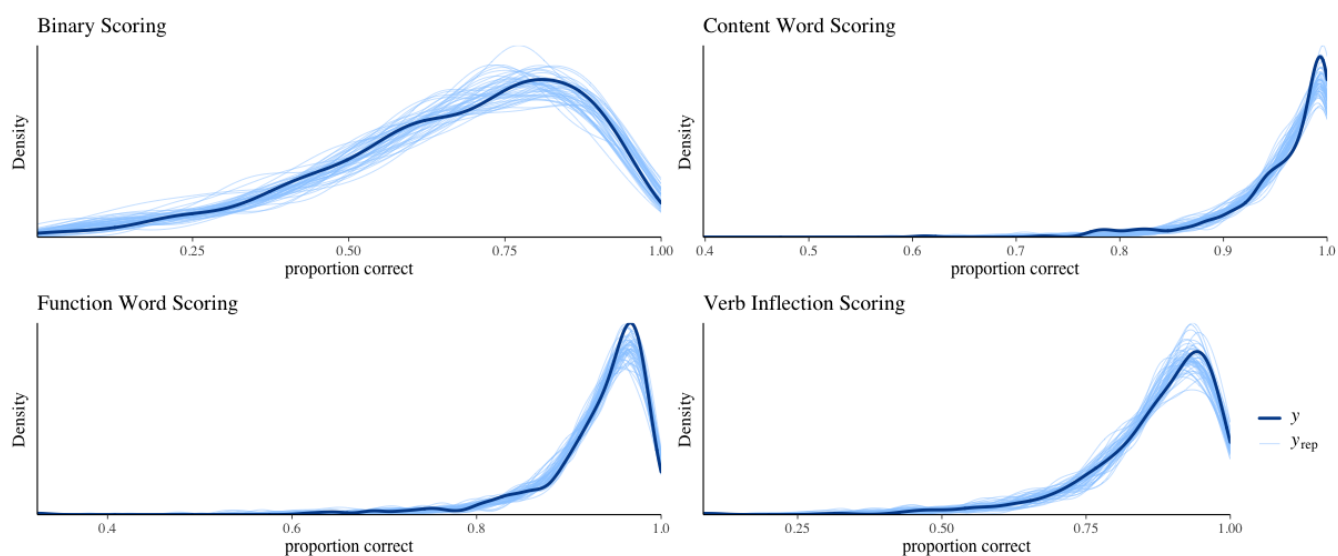


1176

1177

1178 **Figure S2-8**

1179 *Posterior Predictive Checks comparing the distribution of the dependent variable (proportion*
 1180 *correct) to the distribution of simulated datasets for models comparing performance of*
 1181 *groups at T3 for each method of scoring SR productions.*



1182

Figure S2-9

Posterior Predictive Checks comparing the distribution of the dependent variable (proportion correct) to the distribution of simulated datasets for models comparing performance of groups at T4 for each method of scoring SR productions.

