



Sentence Repetition as a Diagnostic Tool for Developmental Language Disorder: A Systematic Review and Meta-Analysis

DOI:

[10.1044/2024_JSLHR-23-00490](https://doi.org/10.1044/2024_JSLHR-23-00490)

Document Version

Accepted author manuscript

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

Citation for published version (APA):

Ward, L., Polisenska, K., & Bannard, C. (2024). Sentence Repetition as a Diagnostic Tool for Developmental Language Disorder: A Systematic Review and Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, 1-31. Advance online publication. https://doi.org/10.1044/2024_JSLHR-23-00490

Published in:

Journal of Speech, Language, and Hearing Research

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact openresearch@manchester.ac.uk providing relevant details, so we can investigate your claim.



1
2
3 **Sentence Repetition as a Diagnostic Tool for Developmental Language Disorder: A**
4 **Systematic Review and Meta-Analysis**

5
6 Leah Ward¹, Kamila Poliřenská^{1, 2} and Colin Bannard³

7 ¹Division of Psychology, Communication and Human Neuroscience,

8 The University of Manchester

9 ²Department of Language and Communication Science,

10 City, University of London

11 ³Department of Linguistics and English Language,

12 The University of Manchester

13
14
15
16 **Author Note**

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

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

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

20 Kamila Poliřenská is now at the Department of Language and Communication
21 Science, City, University of London.

22 This study was registered with the PROSPERO International Prospective Register of
23 Systematic Reviews and Meta-Analyses (Identifier CRD42022303100). The authors have no
24 conflicts of interest to disclose.

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

28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55

Abstract

Purpose: This systematic review and multilevel meta-analysis examines the accuracy of Sentence Repetition (SR) tasks in distinguishing between typically developing (TD) children and children with developmental language disorder (DLD). It explores variation in the way that SR tasks are administered and/or evaluated and examines whether variability in the reported ability of SR to detect DLD is related to these differences.

Method: Four databases were searched to identify studies which had used a SR task on groups of monolingual children with DLD and TD children. Searches produced 3,459 articles of which, after screening, 66 were included in the systematic review. A multilevel meta-analysis was then conducted using 46 of these studies. Multiple preregistered subgroup analyses were conducted in order to explore the sources of heterogeneity.

Results: The systematic review found a great deal of methodological variation, with studies spanning 19 languages, 39 SR tasks, and four main methods of production scoring. There was also variation in study design, with different sampling (clinical and population sampling) and matching methods (age- and language-matching). The overall meta-analysis found that on average TD children outperformed children with DLD on the SR tasks by 2.08 SDs. Subgroup analyses found that effect size only varied as a function of matching method and language of task.

Conclusions: Our results indicate that SR tasks can distinguish children with DLD from both age- and language- matched samples of TD children. The usefulness of SR appears robust to most kinds of task and study variation.

56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82

Introduction

Sentence repetition (SR) tasks have become popular for use in language assessment and research. They are quick and easy to administer, while also providing insightful information on a participant's language abilities. Indeed, performance on a SR task is viewed as a promising clinical marker of developmental language disorder (DLD; formerly referred to as specific language impairment or SLI; Bishop et al., 2017) (Conti-Ramsden et al., 2001; Archibald & Joanisse, 2009). To comprehensively assess the utility of SR as a screener for DLD, we conduct a systematic review and meta-analysis assessing the accuracy of SR tasks in distinguishing between typically developing (TD) children and children with DLD. Critically we explore how variability in task and study design affects the reported ability of SR to detect DLD.

SR tasks involve participants listening to a sentence and being asked to repeat it verbatim with no delay. Over the course of the task, these sentences will often differ in length and/or complexity. This verbal recall requires the processing, storage, and regeneration of the sentence, all of which are thought to involve not only short-term memory, but also prior language knowledge and the ability to form a conceptual representation of the sentences recalled (Klem et al., 2015; Potter, 2012). Performance on these repetition tasks therefore provides a reflection of a participant's linguistic knowledge and working memory ability (Nag et al., 2018; Poll et al., 2016; Riches, 2012). Polišenská et al. (2015) provided evidence that performance on SR tasks is dependent on the linguistic structure of the stimuli, providing primarily a test of lexical phonology and morphosyntax. It is for these reasons that SR is a widely used method of assessing language ability and impairment. There is ongoing debate to how each factor is theoretically involved in SR performance, but this is not the focus of the current research. A recent scoping review by Rujas et al. (2021) identified 203 studies which used SR in their methods between 2010 and 2021. Of these, 62% used SR to assess different language abilities and 18% of them used SR as a clinical marker of language impairment.

83 DLD is a neurodevelopmental language condition characterised by impairments in
84 learning, using, and understanding spoken language (Bishop et al., 2017). The reported
85 prevalence of DLD varies in the literature, likely reflecting the lack of a 'gold standard'
86 method of diagnosis. Studies which have estimated prevalence based upon their own direct
87 use of language assessments have reported values of 7.58% in the UK (Norbury et al.,
88 2016), 7.4% in the USA (Tomblin et al., 1997), 6.4% in Australia (Calder et al., 2022), and
89 8.5% in China (Wu et al., 2023). Studies which have estimated prevalence through more
90 indirect methods have reported a lower prevalence, including an estimate of <1% in Finland
91 (Hannus et al., 2009) based upon a retrospective analysis of records from speech and
92 language therapists (SLTs), and 3.36%–3.70% in Denmark (Nudel et al., 2023) based upon
93 self-report questionnaires given to adults. Interpretation of these prevalence values can
94 therefore be difficult, with the values likely being heavily reliant on the chosen methodology.
95 Following from this, McGregor (2020) argues that the problems children with DLD face are
96 often ignored, with the area being under researched and ultimately the children not receiving
97 the support they need. There is a need for reliable screening tools that can identify those
98 who are showing signs of having DLD and need to undergo further diagnostic testing – SR is
99 one such promising tool.

100 Measures of diagnostic accuracy are a useful tool in assessing the diagnostic utility
101 of a task. The most commonly used measures of diagnostic accuracy are sensitivity and
102 specificity (for an introduction see Chu, 1999). Sensitivity is a measure of a task's ability to
103 identify disorder, here the proportion of children with DLD correctly identified by the SR task.
104 Specificity is a measure of a task's ability to reject the presence of disorder, being the
105 proportion of those without DLD correctly identified by the task. Determining the most
106 effective cut-off point (see Yang & Berdine, 2017) for a SR task helps to reduce the two
107 types of classification error caused by low sensitivity (under-diagnosis from false negatives)
108 or low specificity (over-diagnosis from false positives), both of which are harmful in clinical
109 contexts. Other measures of diagnostic accuracy involve likelihood ratios (LR) which directly
110 link to the pre-test and post-test probability of the disorder (Deeks & Altman, 2004). The LR

111 for positive test results (LR+) refers to how likely the positive result (identification of DLD) is
112 to occur in those with the disorder (children who have DLD) compared to without (TD
113 children). In contrast, LR for negative test results (LR-) refers to how likely the negative
114 result is to occur (identification of not having DLD) in those with the disorder (children who
115 have DLD) compared to without (TD children).

116 In their influential work, Conti-Ramsden et al. (2001) administered different clinical
117 marker tasks to 11-year-old children either with or without a history of DLD (then termed
118 SLI). SR was found to deliver high levels of sensitivity and specificity for identifying SLI in
119 English-speaking children. These levels were higher than those yielded from other tasks,
120 including those that tested third person singular or past tense production, and nonword
121 repetition. It is thought that the combined involvement of wider language systems and short-
122 term memory in SR sets it apart from these other tasks. In the years since the publication of
123 Conti-Ramsden et al. (2001), these types of tasks have become commonplace in the
124 research and diagnosis of DLD. In a similar vein, Archibald and Joanisse (2009) found SR
125 tasks to be a better clinical marker of DLD in school-aged children (aged between five and
126 ten years of age) than nonword repetition. More recently, Redmond et al. (2019) reported
127 further evidence of their usefulness in screening for language impairment in children of
128 seven or eight years of age.

129 Pawlowska (2014) conducted a meta-analysis comparing 13 studies and three
130 proposed markers of language impairment – verb tense (seen in 8 of the studies), nonword
131 repetition (seen in 9 of the studies), and SR (seen in 4 of the studies). Each of the studies
132 had to have reported the number of true and false positives and negatives found by the
133 marker tasks in distinguishing between language impairment and TD age-matched groups.
134 SR was found to be the better marker of the three tests, achieving the most promising
135 likelihood ratios across the analysed studies. However, it was concluded that their results
136 were “at best suggestive” (p.2271) of SR as a diagnostic tool for language impairment. It was
137 proposed that existing marker tasks needed refining and validating in future studies to
138 increase their clinical utility.

139 In their scoping review of 203 studies, Rujas et al. (2021) highlight how across the
140 literature they reviewed, the reported evidence of SR as a clinical marker of DLD appeared
141 positive. While no direct quantitative analysis is provided in their paper, they do describe SR
142 as a “suitable” task for detecting DLD. However, their review highlights that there is in fact
143 much variation in the individual uses of tasks, for example surrounding language, stimuli,
144 and scoring. Of the reviewed studies, 65% administered a SR task as part of a wider battery
145 assessment. For example, a popular battery assessment seen was the Recalling Sentences
146 subtest from the Clinical Evaluation of Language Fundamentals (CELF; e.g., the CELF-5;
147 Wiig et al., 2013). There were also at least 50 original tasks used. In addition, Leclercq et al.
148 (2014) highlight that the scoring of productions often differs across the use of SR tasks, and
149 this may impact diagnostic accuracy.

150 **Objectives**

151 Given the current wide-spread use of SR tasks and need for a reliable screener
152 which can aid in the identification of those with DLD, this review aims to synthesise available
153 evidence on the use of SR tasks on monolingual groups of TD children and children with
154 DLD and assess the reported performance differences between the two groups. A
155 systematic review will explore the variation seen in the administration of the SR tasks and
156 the diagnostic accuracy reported in studies. A meta-analysis will then aim to quantify and
157 explore the differences in performance between the TD and DLD groups and how
158 differences in performance may be influenced by task and study variation. While useful
159 reviews have been conducted in relation to DLD for nonword repetition (see Estes et al.,
160 2007 and Schwob et al., 2021), and narrative performance (see Winters et al., 2022), none
161 has previously focused specifically on SR.

162 ***Systematic review***

163 The systematic review will involve a narrative synthesis of the following questions –

- 164 1) What diagnostic accuracy has been reported for SR in distinguishing between
165 children with DLD and TD children?
- 166 2) What kinds of SR tasks have been used?

167 3) What methods are used to score children's productions on the task?

168 4) What levels of reliability does the task achieve?

169 5) What languages are the tasks administered in?

170 6) How has DLD and TD been defined in the sample?

171 **Meta-analysis**

172 A multilevel meta-analysis will calculate an overall effect size of the standardised difference
173 in performance between the DLD and TD groups across the studies. Subgroup analyses will
174 then build upon the some of the variations identified in the systematic review to see how
175 different factors may influence the size of the difference (effect size) in performance between
176 DLD and TD groups. As such, the meta-analysis will focus on answering the following
177 questions –

178 7) Do SR tasks reveal significant performance differences between groups of TD
179 children and children with DLD? What is the main effect size of the studies?

180 8) How does variability in study design and SR administration influence the effect size
181 across the studies? More specifically does effect size vary as a function of the
182 following factors:

183 a. Task choice (standardised/norm references or unstandardised)

184 b. Stimuli presentation (pre-recorded or produced live)

185 c. Time of scoring (live or offline)

186 d. Type of scoring (sentence binary, sub-sentence binary, target binary or
187 error scoring)

188 e. Language of the task

189 f. DLD sample recruited (clinical or population)

190 g. Matching of TD children (age- or language-matched)

191 **Methods**

192 This review was conducted and reported in line with the Preferred Reporting Items
193 for Systematic Review (PRISMA) guidelines (Page et al., 2021). The review is registered
194 with the PROSPERO international prospective register of systematic reviews and meta-

195 analyses, accessible at
196 https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022303100.

197 Ethical approval was not required as data was only retrieved and synthesised from
198 studies already published.

199 **Eligibility Criteria**

200 The following inclusion criteria were used to identify studies for both the systematic review
201 and meta-analysis:

- 202 • Participants needed to be children, defined as subgroups having a mean age of 18
203 years or below.
- 204 • Participants needed to be assumed to be monolingual. Allowances were made for
205 cases where the language status of the children was not specified. No exclusions
206 were made based on dialect spoken, as long as the children could be presumed to
207 be monolingual speakers of the language of the test used.
- 208 • The article must have been published in English.
- 209 • At least two definable sub-groups had to be included (with allocation being
210 independent of the SR task) involving:
 - 211 - A language-impaired group meeting general criteria for DLD or date-appropriate
212 alternative. This meeting of criteria must have been determined through language
213 assessments either as part of the current study, a previous study, or through
214 SLTs (or equivalent).
 - 215 - A typically developing group/control. To be defined as typically developing there
216 must be no concern expressed for the children in terms of a diagnosis of a
217 biomedical condition (such as autism spectrum disorder or hearing loss).
- 218 • A SR task must have been completed by both groups either in isolation or as part of
219 a wider language battery of tests. This must have involved children having to listen
220 to, and verbally repeat, sequences of at least two words in length. This repetition
221 must have been immediate.

- 222 • Studies must have reported some indication of performance of the groups on the SR
223 task. Studies must have included at least one of the following – mean or median
224 performance of groups, statistical tests comparing group performance, graphs or
225 figures visualising performance, or reported measures of achieved diagnostic
226 accuracy of the SR task.

227 Studies were excluded if they: (a) were systematic reviews or meta-analyses, (b)
228 involved solely adult or bilingual populations, (c) involved DLD populations consisting of
229 children with associated disorders or who did not fit the criteria (e.g., delayed speech), (d)
230 did not involve a SR task or involved a task with visual or delayed stimuli, or (e) were not
231 published in English.

232 To be included in the meta-analysis, studies had to meet the above criteria and include
233 the information needed for effect size calculation. This being the performance mean and
234 standard deviation for each group of children with DLD and TD children.

235 **Information Sources**

236 A search of four databases was conducted: PsycINFO, MEDLINE, Scopus, and Web
237 of Science. The search was conducted in April 2022 by the first author. The PsycINFO and
238 MEDLINE databases were searched via the Ovid platform. This was complemented by a
239 Google Scholar search, to ensure that any relevant publications not on the databases (such
240 as recent publications) were considered.

241 **Search Strategy**

242 Databases were searched using a comprehensive list of keywords relating to the two
243 core themes of the review: Developmental language disorder (DLD; 17 terms) and SR tasks
244 (9 terms). The specific search terms used were:

245 **Search 1** - Communica* concern* OR communica* delay* OR communica* disorder* OR
246 communica* impairment* OR delayed language OR developmental language disorder OR
247 DLD OR developmental dysphasia OR impaired language OR language concern* OR
248 language delay* OR language deficit* OR language disorder* OR language impairment* OR
249 primary language impairment* OR specific language impairment OR SLI

250 **Search 2** - elicited imitation OR imitation of sentences OR recalling sentences OR recall of
251 sentences OR repetition of sentences OR repeating sentences OR sentence imitation OR
252 sentence repetition OR sentence recall*

253 **Search 3** – 1 AND 2

254 An exhaustive list of keywords was used to account for previous changes in the use
255 of the diagnostic labels over time (the main change being the transition from ‘specific
256 language impairment’/SLI to ‘developmental language disorder’/DLD (Bishop et al., 2017))
257 and range of wordings which have been used to describe a SR procedure (e.g., recall or
258 imitation).

259 Each term was separated with the Boolean operator OR, and the two themes
260 combined using the operator AND. Searches were limited to published journal articles and
261 there was no restriction placed on publication date.

262 **Selection Process**

263 From those articles identified through the database search, duplicates were removed.
264 The titles and abstracts of all studies were reviewed by the first author to determine whether
265 they included (a) children, (b) a DLD group, and (c) a SR task. This was irrespective of the
266 overall inclusion criteria. If compliance with any one of the three criteria here was unclear,
267 the study was included for full text review. Articles meeting these criteria then underwent a
268 full text screening by the first author. This used the previously set out eligibility criteria to
269 assess whether the article was relevant for the review. In both abstract screening and full
270 text screening stages, 10% of articles were independently screened by the second author to
271 calculate interrater agreement (the outcome of this is included in the results).

272 **Data Collection Process**

273 Following the selection process and in answering the research questions, information
274 was obtained from the included studies using a data extraction form created by the authors
275 (see <https://osf.io/usw2k/>). Data was collected into this spreadsheet by the first author, with
276 the second author independently extracting data from 10% of included articles using the
277 same form. The third author was responsible for assessing agreement, individually

278 comparing each input across forms, and judging whether the information recorded reflected
279 the same level of information. The outcome of this agreement is included in the results.

280 **Data Items**

281 The information that was obtained from the studies included:

282 Author and year of publication; sample size; mean age of samples; language studied;
283 how DLD was defined; how TD was defined; how the groups were matched; the
284 origin of the SR task; the number, length and type of stimuli; how the task was
285 administered; where and how child performance was scored/measured; reliability
286 measures in scoring; and performance outcomes, including raw performance
287 measures, statistical tests and evaluations of diagnostic accuracy.

288 If a study failed to include any of this information, the corresponding cell was left blank.

289 **Type of Scoring**

290 It was expected that articles would use a wide range of methods to score and measure
291 accuracy in children's SR productions. To allow for more meaningful between-study
292 comparisons, each study's method of scoring was assigned one of these four grouping
293 labels during the data extraction process:

- 294 • Sentence Binary – the whole sentence production by a child was recorded as either
295 correct (1) or incorrect (0).
- 296 • Sub-Sentence Binary – the whole sentence production is scored but the score
297 reflects performance on subsequences. For example, each word or syllable within a
298 sentence is scored as either correct (1) or incorrect (0).
- 299 • Target Binary – Only specific elements within the sentence were scored for being
300 correct (1) or incorrect (0).
- 301 • Error Scoring – The score reflects the number of errors made in the production.

302 Full details of these categories are provided in the data extraction guide

303 [<https://osf.io/usw2k/>].

304 **Risk of Bias Assessment**

305 The quality of the included papers was assessed using the Standard Quality
306 Assessment Criteria for quantitative studies (Kmet et al., 2004). The assessment involves 14
307 criteria items relating to all aspects of a study's design ranging from the clarity of its research
308 questions to appropriateness of conclusions drawn (see Kmet et al., 2004 for a full list of
309 criteria). Three of these criteria (points 5, 6, and 7 as numbered in Kmet et al, 2004) were
310 omitted as they were not applicable to the studies analysed here (they relate instead to
311 interventional designs). Each included article was rated against each of the 11 remaining
312 criteria on a scale of 0-2 based on whether they fulfilled the specific criteria: yes (2), partially
313 (1), and no (0). These were summed and proportional quality scores (score / total possible
314 score) calculated for each, with a higher proportion indicating better research quality.
315 The papers were rated relative to the criteria of the current research project. So, for
316 example, point 13 ('results reported in sufficient detail') was scored in relation to SR
317 performance outcomes being reported and not any possible wider results reported. The
318 quality calculated for each study therefore is specific to the quality of evidence contributed to
319 this research and does not hold meaning outside of it. The second author independently
320 assessed the quality of 10% of the articles, with agreement assessed by the third author (the
321 outcome of this is included in the results).

322 **Effect Measures**

323 The primary outcome of the systematic review is a summary of the ways SR tasks
324 are used and assessed in groups of children with DLD and TD children and of the reported
325 performance differences between the groups. Regarding diagnostic accuracy (research
326 question one (RQ1)), this involved looking to any common diagnostic accuracy metrics
327 reported in the papers which quantify the power of the SR tasks in detecting the presence or
328 absence of DLD. This includes the reporting of sensitivity, specificity, and likelihood ratios.

329 **Effect Size Calculations**

330 The meta-analysis (RQ7) builds upon this outcome with the calculation of an overall
331 effect size for the studies. For this, standardised mean difference (SMD) was calculated to
332 quantify the difference in performance between groups of children with DLD and TD children.

333 SMD was calculated using the measure of Hedges' g (Hedges, 1981) due to the small
334 sample size that was expected for some of the studies. A negative effect size indicates that
335 the children with DLD performed with less accuracy (lower score) on the task than those who
336 are TD (higher score). Studies which scored in the opposite direction, with higher scores
337 indicating lower accuracy, had their effect sizes flipped to allow for consistency. For
338 example, in research conducted by Smolík and Vávrů (2014), SR performance was
339 measured by number of inaccurate imitations, which resulted in those with DLD achieving
340 higher scores than comparative TD children. The effect size for this would be positive and
341 inconsistent with our interpretation of effect size. For this reason, the effect size for this (and
342 similar studies) would be flipped and reported as negative to allow for correct interpretation.

343 **Synthesis Methods**

344 ***Systematic Review***

345 The systematic review (concerning RQ1 to RQ6) involves a narrative synthesis and
346 includes every study identified as relevant for the review.

347 ***Meta-analysis***

348 The meta-analysis includes only those relevant studies which also included the
349 information needed for effect size calculations to occur. The calculated effect size estimates
350 are used in the meta-analysis to estimate the overall effect size of the difference between
351 the performance of groups of TD children and children with DLD on SR tasks (RQ7). To
352 overcome dependencies that existed within the data, a multilevel meta-analysis model was
353 fitted in R (using the {metafor} package (Viechtbauer, 2010)). These dependencies occurred
354 on two levels – (1) multiple effect size estimates concerned the same sample's SR
355 productions, for example in comparing measures of scoring productions; and (2) some
356 studies reported effect size estimates for multiple groups of participants. Therefore, a
357 multilevel meta-analysis model was fit to account for effect measures being nested within
358 samples and in turn, samples nested within study/publication. This full meta-analysis model
359 was then compared to models including just one of these levels of nesting individually, with
360 likelihood ratio tests used to compare which model best represents the variability in the data.

361 Heterogeneity was assessed with Cochran's Q statistic (Cochran, 1954) and the I² Index
362 (Higgins & Thompson, 2002). Because heterogeneity was expected between studies,
363 random-effects modelling was used.

364 In exploring the sources of heterogeneity and determining which specific factors
365 influence the power of SR tasks in discriminating groups of TD children and children with
366 DLD, multiple subgroup analyses were conducted in line with RQ8. As part of the subgroup
367 analysis, an overall effect size was calculated for each categorised subgroup, along with the
368 same heterogeneity measures as the main analysis. Omnibus tests calculated as part of the
369 model were used to identify whether there was a moderating effect of one or more of the
370 variables included. The number of studies included in each subgroup analysis varied due to
371 some studies not including the information needed to classify their subgroup.

372 A subgroup analysis will be conducted for each of the following:

373 **Q8a. Task Choice** – This analysis involved grouping and comparing studies as to
374 whether they use a standardised SR task or whether the task they used was
375 unstandardised.

376 **Q8b. Stimuli presentation** – Studies were grouped and compared according to
377 whether sentences were pre-recorded and played to the children, either over
378 speakers or headphones, or whether sentences were produced live to the children.

379 **Q8c. Time of scoring** – Studies were grouped and compared according to whether
380 child SR productions were scored for accuracy live (with the experimenter scoring
381 productions during the session) or offline (with productions recorded and scored after
382 the session).

383 **Q8d. Type of scoring** – Studies were grouped according to how they scored and
384 measured accuracy in SR productions. This involved the four grouping labels
385 described previously: sentence binary, sub-sentence binary, target binary, and error
386 scoring. Two separate subgroup analyses were run using this information. The first
387 comparing each of these four categories. The second compares just sentence binary

388 scoring against each other type of scoring as a single subgroup (a collapsed group
389 containing parts binary scoring, target binary scoring and error scoring).

390 **Q8e. Language of the task** – Studies were grouped and compared according to the
391 language the task was conducted in.

392 **Q8f. DLD sample recruited** – Studies are grouped and compared by the
393 classification of their DLD group of children. The two groups being – those which
394 have a clinical sample of children with DLD (i.e., DLD inclusion is dependent on
395 having a clinical diagnosis of language disorder), and those which performed or
396 gained their DLD sample from a population study.

397 **Q8g. Matching of TD children** – Studies are grouped and compared according to
398 whether the TD control group was formed by matching the DLD sample by
399 chronological age, or by language ability.

400 **Reporting Bias Assessment**

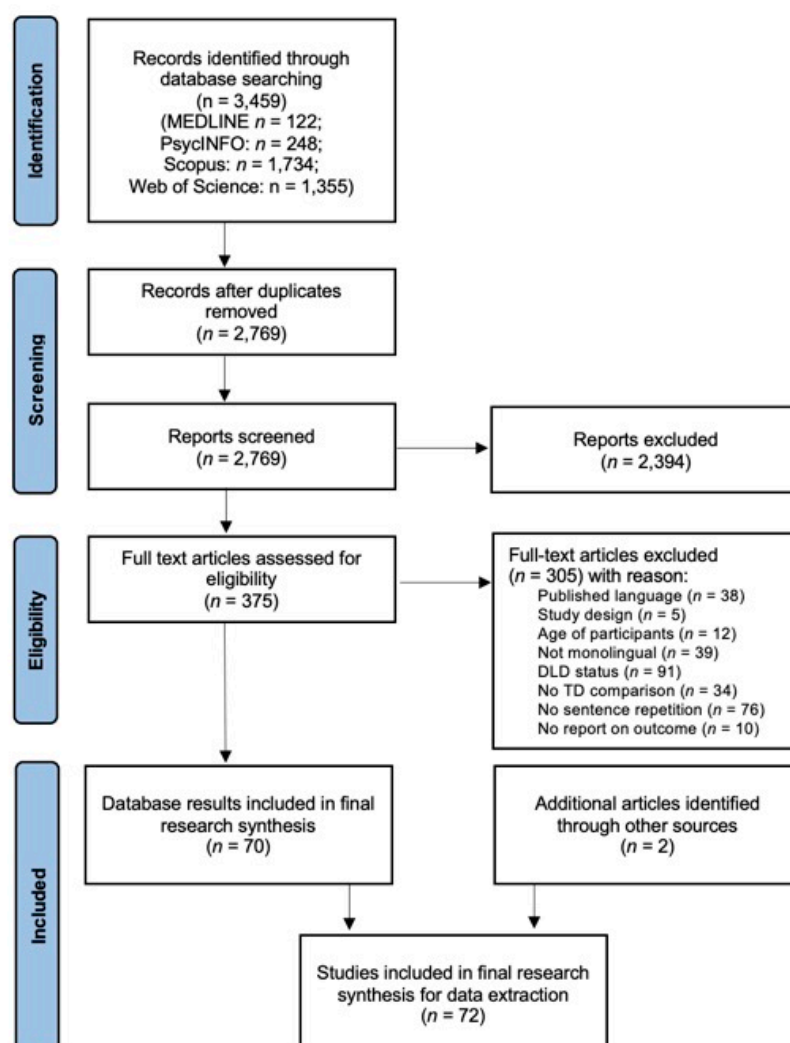
401 The possible presence of publication bias in our data was assessed. A funnel plot
402 was created to visualise any asymmetry that may have been present in effect sizes, which
403 could indicate possible selective reporting. Funnel plot asymmetry was also quantified using
404 Egger's regression test.

405 **Results**

406 **Study Selection**

407 Figure 1 provides a PRISMA flow diagram of the completed selection process. Of the
408 72 articles deemed relevant for the review, a further six were excluded during data extraction
409 and therefore outside of the main selection process. Four of these articles were deemed not
410 to meet the eligibility criteria despite making it through full text review. All of these four
411 articles had group allocation not independent to SR performance. Two articles were
412 duplicate articles which had not been identified in the deduplication phase. This resulted in
413 66 papers included in the final review. These 66 papers were included in the systematic
414 review, and 46 of these were also included in the meta-analysis.

415

416 **Figure 1**417 *PRISMA Flow Diagram Outlining the Selection Process*

418

419 *Note.* A further 6 studies were excluded after the selection process had occurred (during the
 420 data extraction phase) resulting in 66 studies included in the systematic review (and 46 in
 421 the meta-analysis)

422

423 In both the abstract screening and full text screening stages, 10% of articles (277 and
 424 38 randomly selected articles respectively) were independently screened by the second
 425 author. Interrater agreement was 93.86% for the abstract screening stage, with agreement
 426 on the relevance (include or exclude) of 260 out of the 277 articles. Of the 17 disagreed
 427 upon, a separate screening of their full texts identified that none met the eligibility criteria for

428 inclusion, if hypothetically, they had all made it to the next stage. Following on from this, in
429 the full text screening phase, interrater agreement was 97.37%, with agreement on 37 of the
430 38 articles. The second author also independently extracted data from 10% of the articles
431 (seven studies) included in the final review. The overall percentage of interrater agreement
432 (as determined by the third author) for this was 80.26%.

433 **Study Characteristics**

434 The 66 included studies were published between 1986 and 2022. Table 1 provides a
435 summary of the general characteristics of the included studies.

436 There were 75 unique samples of children with DLD, comprising a total of 1675
437 children with DLD (an average of 22.3 children per sample). The ages of children with DLD
438 ranged from 3-years (the specific age in terms of months was not provided) to 16;7 (age;
439 months).

440 There were also 84 unique samples of TD children, comprising 2772 TD children (an
441 average of 33 children per sample). Of these 84 samples of TD children, 64 were matched to
442 the DLD groups based on age (3-years to 13;4), 14 were matched on language level (2;7 to
443 10;1), three were not specifically matched and rather included a younger group of TD
444 children compared to the children with DLD (3;4 to 5;6), and three studies did not specify
445 what matching had taken place (4-years to 14;11).

446 **Risk of Bias Assessment**

447 The quality of included studies, as assessed using the Standard Quality Assessment
448 Criteria for quantitative studies (Kmet et al., 2004), was found to range from 53.38% to
449 95.45%, with a mean quality rating of 78.76% per study. Appendix A breaks down the quality
450 of the studies by criteria. The second author independently assessed the quality of 10% of
451 the articles. The overall percentage of interrater agreement was 72.72%. Looking specifically
452 to the individual points of disagreement, all but one concerned difference in opinion on the
453 assessment of “yes” (2) versus “partial” (1) as to the fulfilment of specific criteri

454 **Table 1**455 *Summary of Study Characteristics*

Study	Language	DLD n	Age	TD n	Age	Matching	Task
Abel, Rice & Bontempo (2015)*	English	20	4;11 - 6;1 (5;5)	23	5;0 - 5;11 (5;5)	Age	Original
				16	3;2 - 3;11 (3;7)	Language	
Acosta-Rodriguez et al. (2020)*	Spanish	25	5;2 - 6;3 (5;6)	25	5;2 - 6;3 (5;7)	Age	CELF-4 (Semel et al., 2006)
		25	5;3 - 6;2 (5;7)	24	5;2 - 6;3 (5;8)		
Alsiddiqi et al. (2021)*	Arabic	24	4;0– 6;11 (5;3)	40	4;0– 6;11 (5;5)	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
Armon-Lotem & Meir (2016)*	Hebrew	14	mean = 6;1	38	mean = 6;0	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
	Russian	14	mean = 5;10	20	mean = 6;1	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
Benavides et al. (2018)*	Spanish	73	4-year-olds	189	4-year-olds	Age	TPL SCREENER
		63	5-year-olds	245	5-year-olds		
		48	6-year-olds	152	6-year-olds		
Blom & Boerma (2019)*	Dutch	78	5;0-6;11 (5;11)	39	5;0-6;10 (5;10)	Age	TAK Sentence Formation
Caselli et al. (2008)*	Italian	16	3;6 - 5;8 (4;8)	32	4;0 - 5;8 (4;8)	Age	Phrase Repetition Test (PRT; Devescovi & Caselli, 2001)
Christensen & Hansson (2012)	Danish	11	5;2 - 7;11 (6;4)	11	5;2 - 7;9 (6;4)	Age	Not specified
				11	3;6 - 5;7 (4;3)	Language	
Coady et al. (2010)	English	18	7;3 - 10;6 (9;0)	18	7;4 - 10;0 (8;10)	Age	Sentences drawn from Hearing In Noise Test
Conti-Ramsden et al. (2001)	English	160	mean = 10;9	100	mean = 10;9	Age	CELF-R (Semel et al., 1994)
De Almeida et al. (2021)	French	17	6;3–8;7 (7;6)	37	5;6–8;4 (7;0)	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
Delage & Frauenfelder (2020)*	French	28	5;0-14;6 (8;10)	48	5;2-12;9 (9;0)	Age	Repetition of Complex Sentences (Delage & Frauenfelder, 2012)

Delage et al. (2021)*	French	52	6;0 - 12;5	36	6;0 - 12;7	Age	Repetition of Complex Sentences (Delage & Frauenfelder, 2012)
Delcenserie et al. (2019)*	French	15	5;2 - 7;0 (6;2)	15	5;0-7;1 (6;2)	Age	CELF-R (Semel et al., 1987)
Duman et al. (2015)*	Turkish	13	5;6-9;1 (6;9)	13	6;3-8;11 (6;9)	Age	Original
Eadie et al. (2002)	English	9	mean = 5;3	10	mean = 3;3	Language	WPPSI-R (Wechsler, 1989) supplementary substest, Sentences
Engberg-Pedersen & Christensen (2017)*	Danish	12	11;1 - 14;0 (12;5)	30	10;10 - 13;4 (12;1)	Age	Test of sentence repetition (Christensen et al., 2012)
Fleckstein et al. (2018)	French	13	6;11 - 8;04 (7;06)	37	5;07 to 6;05 (7;0)	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
Foltz et al. (2015)*	German	8	4;0 - 5;9 (4;10)	8	4;2 - 5;7 (4;10)	Age	Not specified
Frizelle & Fletcher (2014a)*	English	32	6;0 - 7;11 (6;10)	32	6;0 - 7;11 (6;11)	Age	Original
				20	4;7 - 4;11 (4;9)	Not matched	
Frizelle & Fletcher (2014b)	English	32	6;0 - 7;11 (6;10)	32	6;0 - 7;11 (6;11)	Age	Original
				20	4;7 - 4;11 (4;9)	Not matched	
Gagiano & Southwood (2015)	Afrikaans	5	5;3 - 5;10 (5;6)	20	5;3 - 5;11 (5;7)	Age	Original
	English	5	5;2 - 5;11 (5;8)	20	5;4 - 5;11 (5;8)	Age	
Garraffa et al. (2015)	Italian	19	4;3 - 6;3 (5;6)	19	4;2 - 6;5 (5;1)	Age	Original
Georgiou & Spanoudis (2021)*	Greek	24	6;0 - 16;1 (8;1)	39	6;0 - 12;0 (8;10)	Not specified	EREL (Spanoudis & Pahiti, 2014), Sentence Repetition Task
Hakansson & Hansson (2000)	Swedish	10	4;0 - 6;3	10	3;1 - 3;7	Language	Original
Hamaan & Abed Ibrahim (2017)	German	12	5;8 - 9;4 (6;10)	10	5;6 - 7;8 (6;4)	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
Hutchinson et al. (2012)*	English	18	6;1-9;4 (7;9)	24	6;5-9;0 (7;8)	Age	TOLD-P:3 (Newcomer & Hammill, 1997), Sentence Imitation substest
Kamhi & Catts (1986)*	English	12	6;11 - 9;2	12	6;2 - 8;5	Age	Not specified
Kueser & Leonard (2020)	English	17	4;2 - 5;10 (4;11)	19	4;2 - 5;11 (5;0)	Age	Original

				17	2;7 - 3;11 (3;3)	Language	
Lalioti et al. (2016)*	Greek	10	mean = 8;5	24	mean = 4;11	Not matched	DVIQ (Stavrakaki & Tsimpli, 2000), Sentence Repetition Subtest
Leclercq et al. (2014)*	French	34	mean = 9;1	34	mean = 10;2	Age	L2MA2, Sentence Repetition Task
Leroy et al. (2013)*	French	14	6;6 - 11;7 (8;11)	14	5;0 - 10;1 (7;4)	Language	Original
Lukacs et al. (2013)	Hungarian	17	4;10 - 7;2 (6;0)	17	3;3 - 6;2 (5;1)	Language	Original
		29	7;11 - 11;4 (9;10)	29	4;4 - 8;2 (6;3)	Language	
Nash et al. (2013)*	English	32	3-4 (3;8)	69	3-4 (3;9)	Age	SIT-16 (Seeff-Gabriel et al., 2008)
Oetting et al. (2016)*	English	35	5;1 - 6;2 (5;7)	35	5;0 - 5;11 (5;6)	Age	Original
		18	5;0 - 5;11 (5;6)	18	4;11 - 6;2 (5;7)	Age	
Orsolini et al. (2001)*	Italian	10	4;0 - 6;0 (5;1)	20	3;11 - 6;0 (5;1)	Age	Sentence recall task (adapted from Devescovi et al. 1992)
				12	3;4 - 5;6 (4;4)	Not matched	
Peristeri et al. (2021)*	Greek	30	6;0–8;1 (6;9)	30	6;1–7;9 (6;9)	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
Petrucelli et al. (2012)*	English	24	mean = 5;3	32	mean = 5;3	Age	CELF-4 (Semel et al., 2003)
Pham & Ebert (2020)	Vietnamese	10	mean = 5;5	94	mean = 5;8	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
Redmond (2005)*	English	10	5;0 - 8;2 (6;7)	13	5;0 - 8;2 (6;7)	Age	TOLD-P:3 (Newcomer & Hammill, 1997), Sentence Imitation subtest
							Redmond (2005) sentence recall (RSR) task
Redmond & Ash (2017)*	English	29	5;6–11;0 (7;8)	76	5;6–10;10 (7;8)	Age	Redmond (2005) sentence recall (RSR) task
Redmond et al. (2015)*	English	19	mean = 8;3	19	mean = 8;2	Age	Redmond (2005) sentence recall (RSR) task
Redmond et al. (2011)*	English	20	7;0 - 8;11 (7;10)	20	7;1 - 8;11 (7;10)	Age	Redmond (2005) sentence recall (RSR) task
Riches (2015)*	English	17	6;0–7;2 (6;7)	17	4;4–4;9 (4;8)	Language	Original

Riches et al. (2010)*	English	14	14;5 – 16;7 (15;3)	17	14;0 – 14;11 (14;4)	Not specified	Original
Riches (2012)*	English	23	6;0–7;3 (6;7)	19	mean = 6;5	Age	Original
				21	mean = 4;8	Language	
Riches (2017)*	English	17	6;0 - 7;3 (6;7)	17	mean = 6;5	Age	Original
				21	mean = 4;8	Language	
Seeff-Gabriel et al. (2010)*	English	13	4;0 – 6;0 (4;9)	33	4;0 – 6;3 (4;10)	Age	SIT-61 (Seeff-Gabriel et al., 2010)
Smolík et al. (2021)*	Czech	17	5;1–7;6 (6;6)	17	3;8–4;11 (4;3)	Language	Original
Smolik & Vavru (2014)*	Czech	19	4;10 - 7;6 (6.13)	19	4;11 - 7;8, (6.31)	Age	Original
				19	2;09 - 5;8, (4.25)	Language	
Stokes & Fletcher (2003)	Cantonese	13	3;8 - 5;11 (4;6)	14	4;0 - 4;11 (4;5)	Age	Not specified
Stokes et al. (2006)*	Cantonese	14	4;2 - 5;7 (4;11)	15	4;1 - 6;9, (5;0)	Age	Not specified
				15	2;11 - 3;6, (3;3)	Language	
Taha et al. (2021)*	Arabic	30	4;0 - 6;10 (5;2)	60	4;0–6;8 (5;4)	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
Talli & Stavrakaki (2020)*	Greek	16	mean = 8;11	20	mean = 9;0	Age	DVIQ (Stavrakaki & Tsimpli, 2000), Sentence Repetition Subtest
Taylor et al. (2014)*	English	19	5;3 - 12;1 (8;3)	61	5;0–12;1 (8;10)	Age	NEPSY-II; (Korkman et al. 2007), Sentence Repetition Task (SNRep)
Theodorou et al. (2016)*	Greek	16	4;11 - 8;1 (6;2)	22	4;5 - 8;7 (6;10)	Age	DVIQ (Stavrakaki & Tsimpli, 2000), Sentence Repetition Subtest
Theodorou et al. (2017)*	Greek	9	4;11–5;11 (5;6)	10	4;5–6;6 (5;8)	Age	Original
		7	6.7–8.1 (7;8)	12	6;7–8;7 (7;10)	Age	
Thordardottir & Brandeker (2013)	French	14	mean = 5;2	14	mean = 5;0	Age	French adaptation by Royle and Elin Thordardottir (2003) of the Recalling Sentences in Context subtest of the CELF-Preschool

Thordardottir et al. (2011)*	French	14	4;6 - 5;11 (5;1)	78	4;1–5;11 (4;11)	Age	French adaptation by Royle and Elin Thordardottir (2003) of the Recalling Sentences in Context subtest of the CELF-Preschool DVIQ (Stavrakaki & Tsimpli, 2000), Sentence Repetition Subtest
Tsimpli et al. (2016)*	Greek	21	5;5–11;6 (9;3)	21	5;2–11;5 (9;0)	Age	
Tuller et al. (2018)	French	17	6;3 - 8;8 (7;7)	37	5;7–8;5 (7;0)	Age	LITMUS-SRep (Marinis and Armon-Lotem 2015)
	German	12	5;8 - 9;4 (7;0)	10	5;6 - 7;8 (6;4)	Age	
Van Der Meulen et al. (1997)	Dutch	30	4;4 - 6;11	30	Not Specified	Age	Original
Vang Christensen (2019)*	Danish	16	5;10–9;11 (7;9)	37	5;3–10;4 (7;9)	Age	Original
		11	11;1–14;1 (12;3)	50	10;10–13;4 (12;5)	Age	
Ziethe et al. (2013)	German	19	5+-6;0	25	4+-6;0	Not specified	HSET (Grimm & Schöler, 1991) – The Imitation of Grammatical Structure Forms (IGS) subtest
Wang et al. (2022)*	Mandarin	16	4;2 - 5;10 (5;0)	16	4;2 - 5;11 (5;1)	Age	Original
Dosi (2019)	Greek	10	mean = 8;11	10	mean = 8;11	Age	DVIQ (Stavrakaki & Tsimpli, 2000), Sentence Repetition Subtest

456 *Note.* References marked with an asterix (*) were included in the meta-analysis. Ages are presented in years;months format. DLD = children
457 with developmental language disorder; TD = typically developing children; CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth
458 Edition; LITMUS-SRep = LITMUS Sentence Repetition task; TPL = Tamiz de Problemas de Lenguaje; TAK = Taaltest Alle Kinderen; CELF-R =
459 Clinical Evaluation of Language Fundamentals–Revised; WPPSI-R = Wechsler Preschool and Primary Scale of Intelligence–Revised; EREL =
460 Expressive and Receptive Language Evaluation; TOLD-P:3 = Test of Language Development–Primary: Third Edition; DVIQ = Diagnostic
461 Verbal IQ Test; L2MA2 = Language Oral, Language Écrit, Mémoire et Attention 2; SIT = Sentence Imitation Test; NEPSY-II =
462 Neuropsychological Assessment–Second Edition; HSET = Heidelberger Sprachentwicklungstest.

463 **Systematic Review**464 **RQ1 What is the diagnostic accuracy of SR in distinguishing between children with**
465 **DLD and typically developing children (TD)?**

466 In assessing the discriminative power of SR in indicating the presence or absence of
467 DLD, diagnostic accuracy metrics were reported in 18 of the studies (for 21 different
468 samples). These values are outlined in Table 2.

469

470 **Table 2**471 *Diagnostic accuracy metrics for SR in indicating the presence or absence of DLD*

Study	Language	Scoring	Matching	Cut Off	Sen	Spe	LR+	LR-
Armon-Lotem & Meir (2016)	Hebrew	Target Binary	Age	0.86	1	0.87	7.6	0
Armon-Lotem & Meir (2016)	Russian	Target Binary	Age	0.88	0.86	0.9	8.57	0.16
Christensen & Hansson (2012) ^a	English	Target Binary	Age and Language	62%-77%	1	1	-	0
De Almeida et al. (2021)	French	Sentence Binary	Age	80%	0.94	0.92	8.54	0.07
Fleckstein et al. (2018)	French	Sentence Binary	Age	80%	0.92	0.92	11.4	0.08
Hamaan & Abed Ibrahim (2017)	German	Sentence Binary	Age	63.33%	1	1	-	0
Hamaan & Abed Ibrahim (2017)	German	Target Binary	Age	77.78%	1	1	-	0
Leclercq et al. (2014)	French	Sentence Binary	Age	-1.31SD	0.97	0.91	10.78	0.03
Oetting et al. (2016)	African American English (AAE)	Error Scoring	Age	40	0.89	0.86	6.36	0.13
Oetting et al. (2016)	Southern White English (SWE)	Error Scoring	Age	40	0.94	0.83	5.53	0.07
Pham & Ebert (2020)	Vietnamese	Sentence Binary	Age	0.85	1	0.57	2.33	0
Pham & Ebert (2020)	Vietnamese	Error Scoring	Age	0.89	0.9	0.71	3.13	0.14

Pham & Ebert (2020)	Vietnamese	Target Binary	Age	0.89	0.89	0.71	2.76	0.28
Redmond et al. (2011)	English	Error Scoring	Age	14.5	0.9	0.9	9	0.11
Stokes et al. (2006)	Cantonese	Error Scoring	Age	67	0.77	0.97	25.66	0.24
Taha et al. (2021)	Arabic	Sentence Binary	Age	70.14	0.93	0.93	13.94	0.07
Taha et al. (2021)	Arabic	Error Scoring	Age	79.4	0.93	0.98	54.88	0.07
Taha et al. (2021)	Arabic	Target Binary	Age	90.97	0.97	0.92	11.27	0.07
Theodorou et al. (2016)	Greek	Error Scoring	Age	43	0.88	0.91	9.62	0.14
Theodorou et al. (2017)	Greek	Sentence Binary	Age		0.75	0.82	4.12	0.31
Theodorou et al. (2017)	Greek	Error Scoring	Age		0.75	0.77	3.3	0.32
Thordardottir & Brandeker (2013)	French	Sub- Sentence Binary	Age	74%	0.93	0.86	6.64	0.08
Thordardottir et al. (2011)	French	Sub- Sentence	Age	0.74	0.93	0.79	4.36	0.09
Tuller et al. (2018)	French	Sentence Binary	Age	78.3%	0.93	0.92	11.52	0.07
Tuller et al. (2018)	German	Sentence Binary	Age	63.3%	1	1	-	0
Vang Christensen (2019)	Danish – Younger sample	Sentence Binary	Age	17	0.94	0.97	34.7	0.06
Vang Christensen (2019)	Danish – Older sample	Sentence Binary	Age	31	0.91	0.98	45.5	0.09
Wang et al. (2022)	Mandarin	Error Scoring	Age	63%	1	1	-	0
Wang et al. (2022)	Mandarin	Sentence Binary	Age	41%	1	0.88	8	0

472 *Note.* Only studies that provided values based on a calculated cutoff point of performance

473 are included. Sen = sensitivity; Spe = specificity; LR+ = positive likelihood ratio,

474 calculated using the following formula: $\text{sensitivity}/(1 - \text{specificity})$; LR- = negative likelihood
475 ratio, calculated using the following formula: $(1 - \text{sensitivity})/\text{specificity}$. '—'
476 indicates that LR+ could not be calculated.

477 ^a Christensen and Hansson (2012) reported that for any cutoff score between 62% and 77%,
478 100% sensitivity and specificity was achieved for both age- and language-matched
479 groups.

480

481 Looking first to sensitivity (the proportion of children with DLD being correctly
482 identified) and specificity (the proportion of children without DLD correctly identified) – values
483 ranged from 75% to 100% sensitivity and 57% to 100% specificity. According to Plante and
484 Vance's (1994) recommendations, studies ranged from having unacceptably low levels of
485 discriminative accuracy to good accuracy. 51.72% of the pairs of sensitivity and specificity
486 values indicated good classification accuracy of the SR test (above 90% for both values).
487 24.14% indicated fair accuracy (both values above 80%), and 24.14% indicated poor
488 discrimination (one or both values below 80%).

489 These sensitivity and specificity values also allowed for the calculation of Likelihood
490 ratios to further assess the utility of these tasks. These values are also shown in Table 2.
491 The further the likelihood ratio is from one, the better the discriminative ability of the task,
492 with a stronger association between task performance and the presence or absence of DLD.
493 Looking to general guidelines for Likelihood ratio interpretation (Deeks & Altman, 2004), all
494 studies showed $\text{LR+} > 1$ and $\text{LR-} < 1$, indicating that test results on SR tasks are associated
495 with both the presence and absence of DLD. Across the sets of values, 44.83% of the pairs
496 of likelihood ratios showed $\text{LR+} > 10$ and $\text{LR-} < 0.1$, and the SR tasks can be considered to
497 show strong evidence of detecting the presence and absence of DLD.

498 ***RQ2 What kinds of SR tasks have been used?***

499 Looking back to Table 1, a variety of different SR tasks were utilised by the studies.
500 Table 3 provides a summary of the specific tasks used. As can be seen, 18 tasks were
501 original and created by their authors. The remaining 41 tasks seen involved using or

502 adapting a pre-existing task or principles. Of these, 22 tasks were standardised/norm
503 referenced tasks.

504

505 **Table 3**

506 *Specific SR Tasks Administered*

Task	Frequency
Original Task	18
Using or adapting LITMUS-SRep (Marinis & Armon-Lotem 2015)	8
Using or adapting the Recalling sentences subtest of the CELF*	6
Sentence repetition subsection of the DVIQ (Stavrakaki & Tsimpli, 2000)*	5
Sentence recall (RSR) task (Redmond, 2005)*	4
Sentence Imitation (SI) subtest from the TOLD-P3 (Newcomer & Hammill, 1997)*	2
Repetition of Complex Sentences (Delage & Frauenfelder, 2012)	2
Sentence repetition subtest of the TPL screener tool (Benavides et al., 2018)	1
Sentence formation test from the TAK Language Proficiency Test (Verhoeven & Vermeer, 2001)*	1
Phrase Repetition Test (PRT; Devescovi & Caselli, 2001)	1
Sentence recall task (adapted from Devescovi et al. 1992)	1
Sentences extracted from Hearing in Noise Test (HINT; Nilsson et al., 1994)*	1
Sentences supplementary subtest of WPPSI-R (Wechsler, 1989)*	1
Test of sentence repetition (Christensen et al., 2012)	1
Recalling Sentences subtest of EREL (Spanoudis & Pahiti, 2014)*	1
Sentence repetition task of the L2MA2 (Chevrie-Muller et al., 2010)*	1
SIT-16 (Seeff-Gabriel et al., 2008)	1
SIT-61 (Seeff-Gabriel et al., 2010)	1
Sentence Repetition Task (SNRep) from the NEPSY-II; (Korkman et al. 2007)*	1
The Imitation of Grammatical Structure Forms (IGS) subtest from the HSET (Grimm & Schöler, 1991)*	1

507 *Note.* SR tasks marked with an asterisk (*) are classified as standardized/norm-referenced
508 tasks. LITMUS-SRep = LITMUS Sentence Repetition task; CELF-5 = Clinical Evaluation of
509 Language Fundamentals–Fifth Edition; DVIQ = Diagnostic Verbal IQ Test; TOLD-P:3 = Test
510 of Language Development–Primary: Third Edition; TPL = Tamiz de Problemas de Lenguaje;
511 TAK = Taaltest Alle Kinderen; WPPSI-R = Wechsler Preschool and Primary Scale of
512 Intelligence–Revised; EREL = Expressive and Receptive Language Evaluation; L2MA2 =

513 Language Oral, Language Écrit, Mémoire et Attention 2; SIT = Sentence Imitation Test;
514 NEPSY-II = Neuropsychological Assessment–Second Edition; HSET = Heidelberger
515 Sprachentwicklungstest.

516

517 The most common sentence repetition task seen (aside from those which were
518 completely original tasks) was those created based upon the principles of the LITMUS-SRep
519 task (Marinis & Armon-Lotem 2015), initially created as part of COST Action IS0804
520 'Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to
521 Assessment'. While the primary intentions of COST Action IS0804 was to identify bilingual
522 DLD, the principles set out by LITMUS-SRep have been applied in the creation of sentence
523 repetition tests for a diverse set of languages, with its application here being seen in: Arabic,
524 French, German, Greek, Hebrew, Russian, and Vietnamese.

525 Not all the tasks used or adapted were originally designed for language assessment.
526 For example, sentences were seen from the Wechsler Preschool and Primary Scale of
527 Intelligence-Revised (WPPSI-R; Wechsler, 1989; used by Eadie et al., 2002) which is an
528 assessment of child intelligence. Sentences were also seen extracted from the Hearing in
529 Noise Test (HINT; Nilsson et al., 1994; used by Coady et al., 2010) which is generally used
530 within audiology. In the context of the reviewed studies, these tests were used for language
531 assessment to evaluate the difference in performance between children with DLD and TD
532 children.

533 Turning more closely to how these tasks were administered, 30% of studies involved
534 the individual presenting the task reading the sentences live for the children to repeat, 43%
535 had children listen to pre-recorded sentences and 27% did not specify. Of those that
536 presented pre-recorded sentences, 10 were played over headphones and four over a
537 speaker without headphones. The rest again did not specify.

538 In terms of specific methods of administration, seven were presented using
539 PowerPoint slides (Armon-Lotem & Meir, 2016; De Almeida et al., 2021; Fleckstein et al.,
540 2018; Oetting et al., 2016; Pham & Ebert, 2020; Theodorou et al., 2017; Wang et al., 2022),

541 six were presented in a task involving a puppet producing the sentences to repeat
 542 (Christensen & Hansson, 2012; Frizelle & Fletcher, 2014a, 2014b; Riches, 2012, 2015,
 543 2017), four were presented with accompanying figures or pictures (Caselli et al., 2008;
 544 Garraffa et al., 2015; Orsolini et al., 2001; Stokes & Fletcher, 2003), and three involved
 545 sentences embedded within stories (Leroy et al., 2013; Thordardottir & Brandeker, 2013;
 546 Thordardottir et al., 2011).

547

548 ***RQ3 What methods are used to score children's productions on the task?***

549 Around half of studies scored children's productions in the SR tasks offline, meaning
 550 that children's productions were audio-recorded in the session to be later transcribed and/or
 551 scored. 20% were scored online, with children's productions being scored for accuracy as
 552 the session was taking place. The remaining 32% did not specify where scoring had taken
 553 place. There was a range of methods of scoring seen in the SR tasks. Across the studies
 554 this was broken down into 4 main categories. Table 4 describes these categories and the
 555 frequency in which each was seen across the studies.

556

557 **Table 4**

558 *Four defined categories for scoring SR productions*

Category of Scoring	Description	Frequency
Sentence Binary	Tasks where the whole sentence production by a child was recorded as either correct (1) or incorrect (0).	24
Sub-Sentence Binary	Tasks which scored the whole sentence production on a closer level. Each word or syllable (etc.) within a sentence are scored as either correct (1) or incorrect (0).	12
Target Binary	Tasks where only specific elements of productions are scored. These specific elements are scored as either correct (1) or incorrect (0).	26
Error Scoring	Each sentence is scored on a scale as to how many errors are produced in the production.	28

559 *Note.* Full details of these categories are provided in the data extraction guide

560 (<https://osf.io/usw2k/>).

561

562 The target structure specifically looked to as part of “target binary scoring” varied per
563 study. For example, Christensen and Hansson (2012) created an original task looking at
564 past tense verb position and only scored productions for whether the target verbs were
565 produced correctly or not. Other examples of target structures included object-relatives
566 (Delage et al., 2021), lexicalized and non-lexicalized forms (Leroy et al., 2013), and suffixes
567 on nouns (Lukacs et al., 2013). In a similar light, there was variation as to the specific type of
568 error scoring seen. One popular method of error scoring was that used with the recalling
569 sentences subtest of the CELF (Wiig et al., 2013) which involves scoring responses in
570 relation to the number of errors in the production on a scale of 0 to 3 — 3 points were given
571 to productions identical to the target sentence, 2 points were given to productions with one
572 error/deviation from the target, 1 point was given to productions with two or three errors, and
573 0 was given to those with four or more errors. This method was seen not only in those
574 studies using the CELF recalling sentences subtest, but also in many studies which used
575 different or original SR tasks. Another popular method of error scoring was on a scale of 0 to
576 2 (developed by Archibald & Joanisse, 2009). Other methods of error scoring involved 5-
577 point (Duman et al., 2015) and 10-point (Frizelle & Fletcher, 2014a; 2014b) scales with
578 scores reflecting the specific type of error made, and Levenshtein Distance calculated for
579 words (Riches, 2012) or morphemes (Riches, 2017).

580 For these described scoring methods, phonological deviations were generally
581 disregarded. This was with the exception of Delage et al. (2021), Kamhi and Catts (1986),
582 Kueser and Leonard (2020), Taylor et al. (2014) and Wang et al. (2022), who all classified
583 phonological errors as causing an incorrect production. Some studies made further
584 allowances, for example Armon-Lotem and Meir (2016) allowed for lexical substitutions (e.g.
585 son/boy) in their binary target structure scoring, as did Duman et al. (2015) and Garraffa et
586 al. (2015) in their respective scoring systems.

587

588

589

590 **RQ4 What levels of reliability does the task achieve?**

591 Only 51.52% of studies reported that the transcriptions or scorings of children's
 592 productions were verified, and the reliability assessed. The depth of this ranged from
 593 assessing a subsample of 5% of the sample to looking to the whole sample.
 594 For studies which provided specific measures of reliability, levels were generally high. Inter-
 595 transcriber agreement was reported in eight studies (12.12%) and ranged from 92.5% to
 596 99.6%. Inter-scorer agreement was reported in 27 studies (40.91%) and ranged from 86.5%
 597 to 100%.

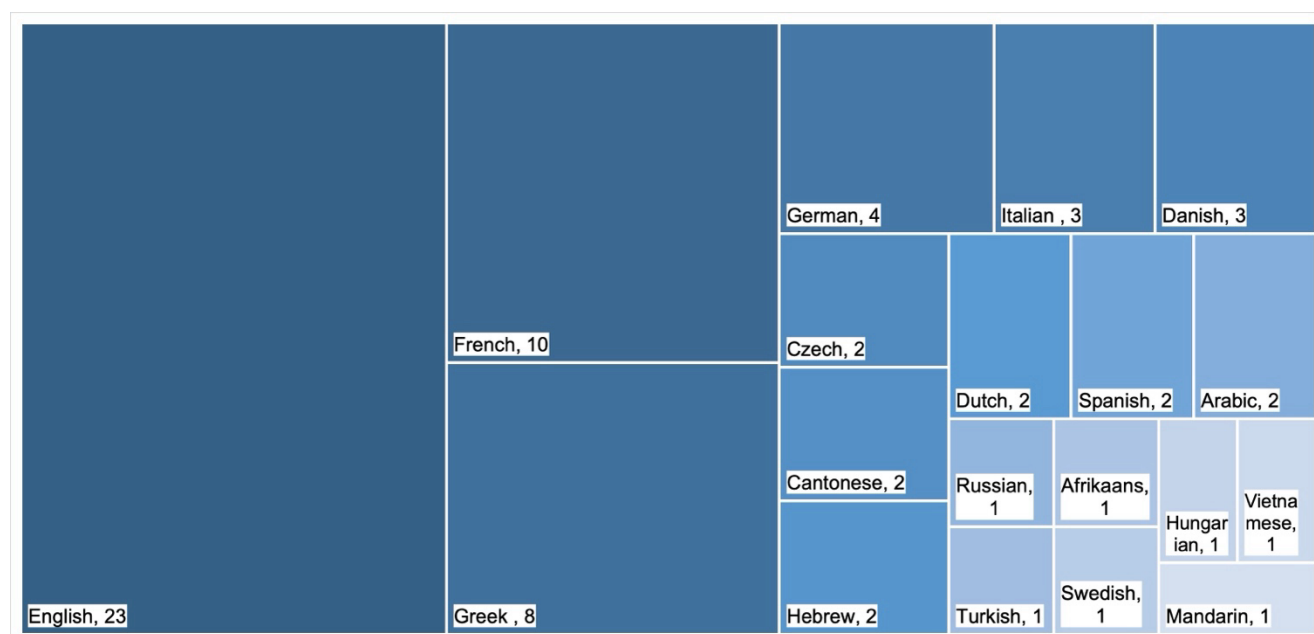
598 **RQ5 What languages are the tasks administered in?**

599 Tasks were conducted in 19 different languages, visualised in Figure 2. The most
 600 common language spoken was English (32.86% of samples), followed by French (14.29%)
 601 and Greek (11.43%).

602

603 **Figure 2**

604 *Tree map of the languages the tasks were conducted in across studies*



605

606 *Note.* Areas of the tree map are in proportion to the frequencies of studies seen in the

607 systematic review.

608

609 ***RQ6 How has DLD and TD been defined in the samples?***

610 Across the studies there were 52 clinical samples of children with DLD, with children
611 being recruited because of a prior referral or diagnosis, for example from speech and
612 language clinics or hospitals. Of these, 38 (of the 52) clinical samples also underwent
613 additional testing as part of the study to verify the children's language status. There were
614 also 11 population samples of children with DLD, with the grouping determined by the
615 studies own or a prior study's testing alone. Children in these studies were generally
616 recruited from schools.

617 For defining TD children, 50 samples involved children undergoing the same testing as DLD
618 children to determine TD status. As previously outlined, of the 84 samples of TD children, 64
619 were matched to the DLD groups based on age, and 14 were matched on language level.
620 Age matching generally occurred on a group-level and involved children being matched
621 because they are in the same school year. The methods of language matching varied across
622 the studies: seven samples of TD children were matched to DLD groups based on their
623 mean length utterance (either in words or morphemes), five TD samples were matched on
624 measures of receptive vocabulary/grammar, one was matched on a measure of productive
625 vocabulary, and one was matched on sentence comprehension abilities.

626 **Meta-Analysis**

627 The meta-analysis involved the inclusion of 46 studies, all of which reported the
628 means and standard deviations of DLD and TD group performance. From these 46 studies,
629 there were 103 effect sizes calculated for use in the meta-analysis.

630 Because multiple effect sizes sometimes came from a single sample, and in turn a
631 single study, a multilevel meta-analysis was fit. Model fit was compared using likelihood ratio
632 tests for different levels of nesting (see appendix B). From this, it was determined that the
633 model which best represented the variance in the data was one where effect sizes were
634 nested within study. Nesting within sample in addition to or in place of nesting by study did
635 not allow for a better fit.

636 ***RQ7 Do SR tasks reveal significant performance differences between groups of TD***
637 ***children and children with DLD? What is the main effect size of studies?***

638 Figure 3 shows a forest plot showing the effect sizes from each included study. The
639 overall meta-analysis found an average effect size of $g = -2.08$ (95% CI [-2.32, -1.84]). On
640 average, TD children outperformed children with DLD on the SR tasks by 2.08 SDs.
641 Heterogeneity was found across effect sizes, $Q(102) = 635.40$, $p < .001$, with a between-study
642 I^2 value of 52.67%, and a within-study I^2 value of 30.57%.

643 ***RQ8 How does variability in study design and SR administration influence the effect***
644 ***size across the studies? More specifically does effect size vary as a function of the***
645 ***following factors:***

646 In exploring the sources of heterogeneity and what specific factors influence the
647 power of SR tasks in discriminating groups of TD children and children with DLD, multiple
648 subgroup analyses were conducted. Forest plots showing the overall results of these
649 subgroup analyses are shown in Figures 4 and 5.

650 **RQ8a. Task choice (standardised/norm references or unstandardised)** – This
651 subgroup analysis involved 83 effect sizes (20 were excluded for not including the
652 necessary information). The test for subgroup difference showed there was no
653 significant subgroup effect ($p = .81$). This indicates that there is no evidence that the
654 SR task used (in terms of being a standardised test or unstandardised test)
655 influenced the size of the difference between groups of DLD and TD children.

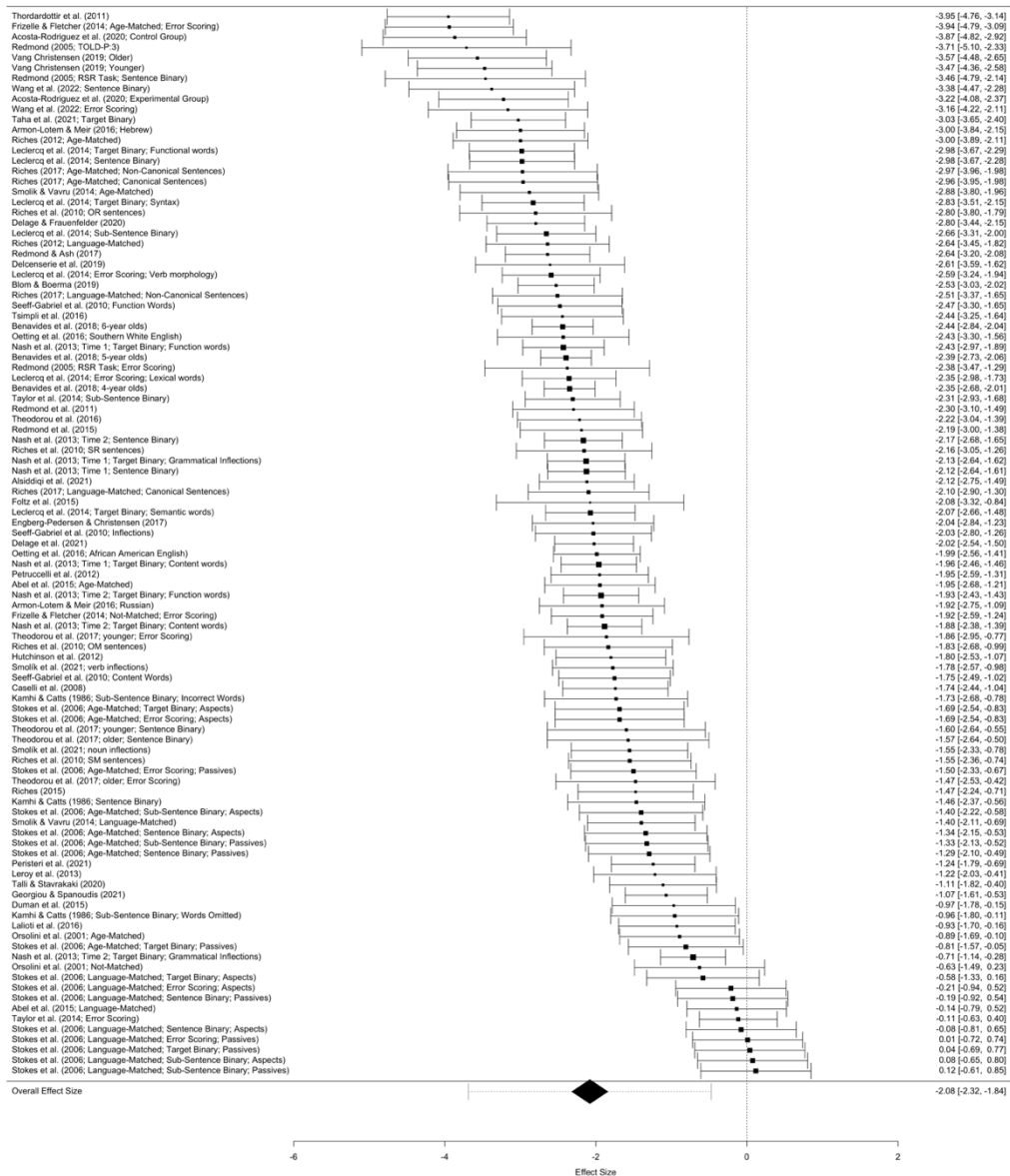
656 **RQ8b. Stimuli presentation (pre-recorded or produced live)** – This subgroup
657 analysis involved 69 effect sizes (34 were excluded for not including the necessary
658 information). The test for subgroup difference showed no significant subgroup effect
659 ($p = .55$), indicating that there was no evidence that the difference in performance
660 between groups of DLD and TD children was affected by the sentences in the tasks
661 being pre-recorded or produced live.

662

663

664 **Figure 3**

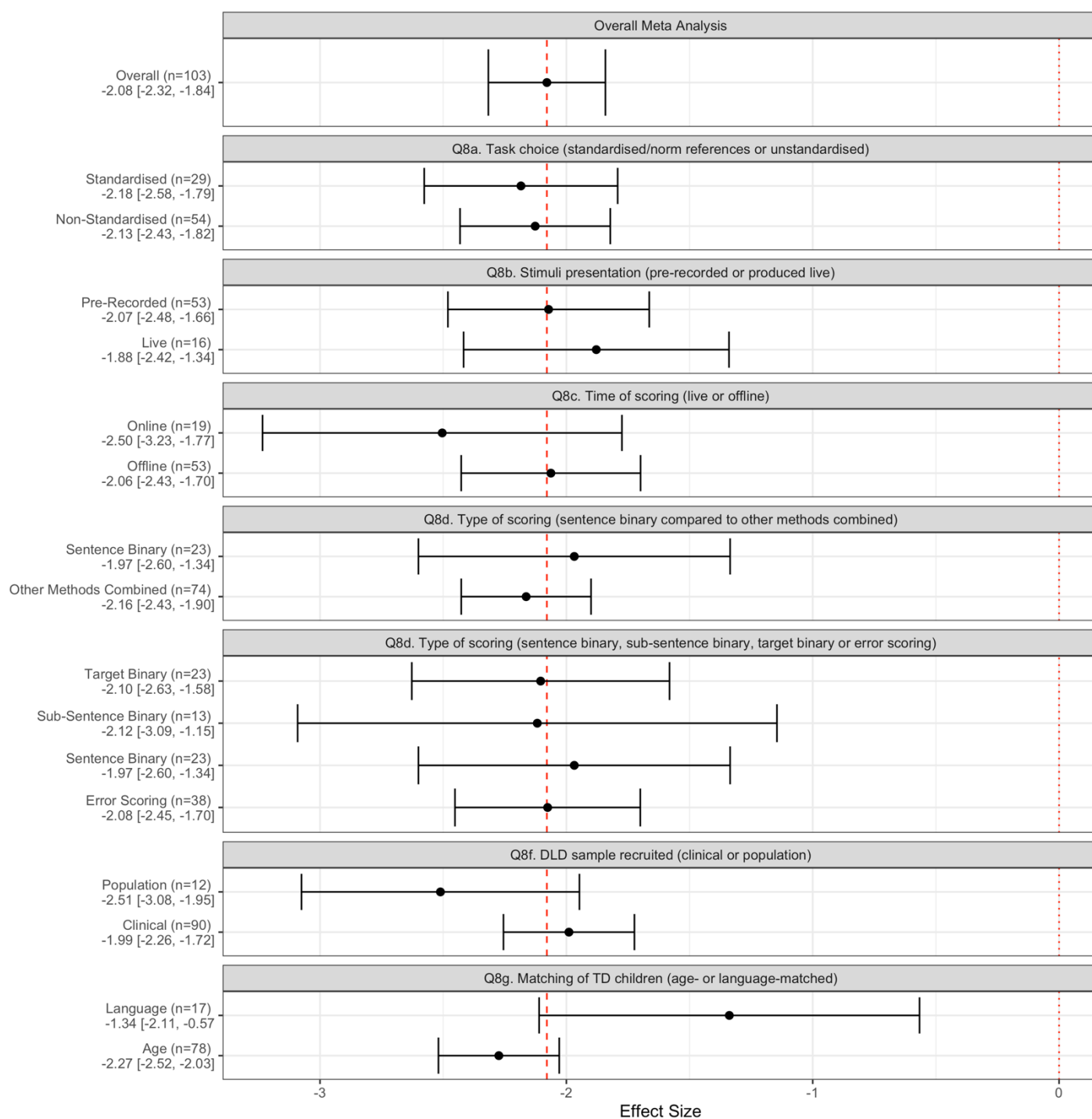
665 *Forest plot showing the effect size of each included study and calculated pooled effect size*



666 *Note.* Data points are presented in order of effect. Points represent a calculation of standardized
 667 mean difference using Hedges' g (Hedges, 1981) surrounded by 95% confidence intervals (the
 668 values to which are on the right-hand side). The "overall effect size" displays the result of the
 669 multilevel meta-analysis. The size of points is proportional to the weight of the point in relation to
 670 the pooled estimate (overall effect size). TOLD-P:3 = Test of Language Development–Primary:
 671 Third Edition; RSR = Redmond sentence recall; SM = subject relative sentences with adjectives
 672 in the main clause; SR = subject relative sentences with adjectives in the relative clause; OM =
 673 object relative sentences with adjectives in the main clause; OR = object relative sentences with
 674 adjectives in the relative clause.

675 **Figure 4**

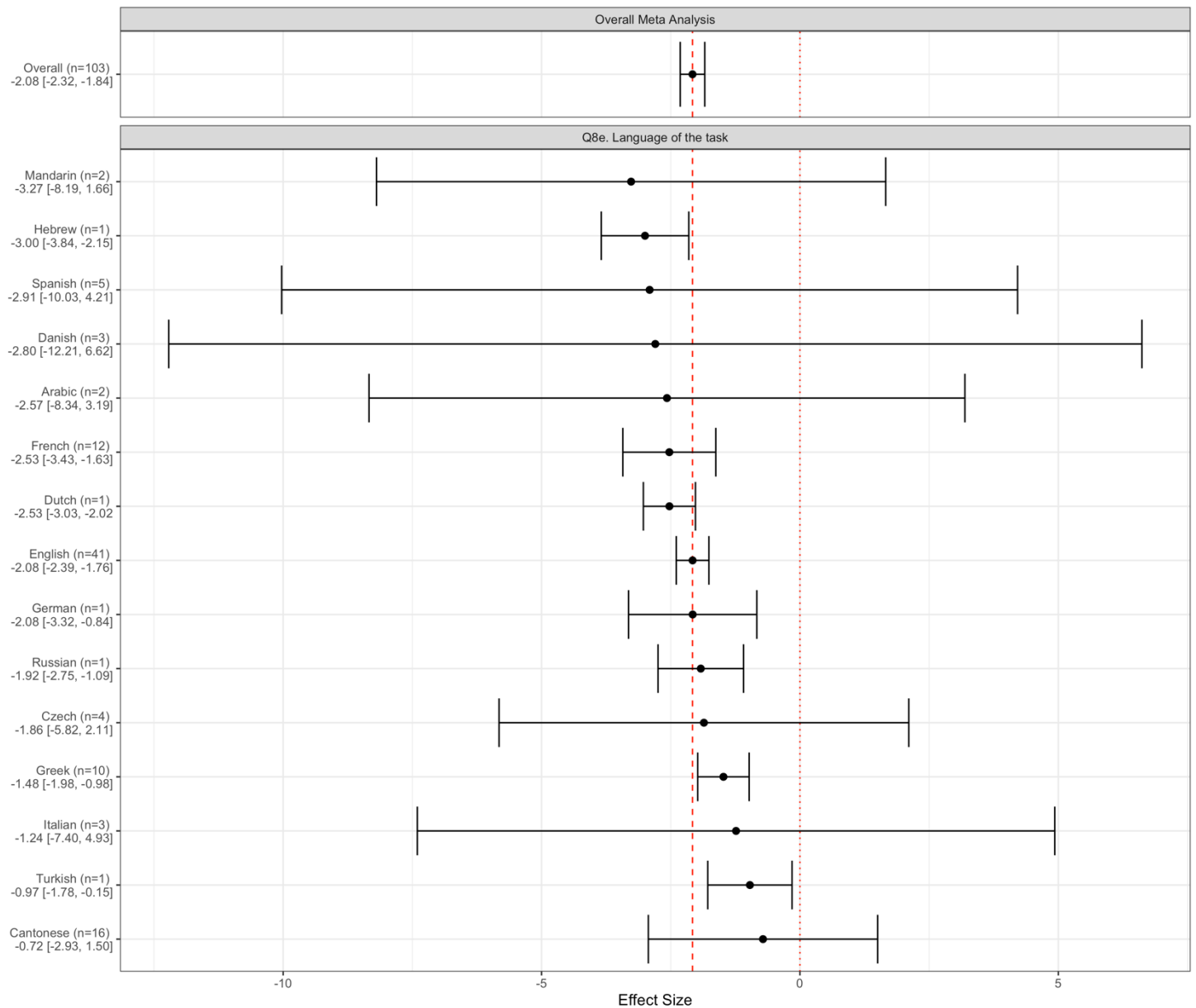
676 *Summary forest plot showing the pooled effect size for each subgroup analysis*



677 *Note.* Points represent a calculation of the pooled estimate of effect size (Hedges' g) from a
 678 multilevel meta-analysis for each defined subgroup (surrounded by 95% confidence
 679 intervals). In the image, 'n' refers to the number of datapoints included in each analysis (not
 680 number of studies). DLD = developmental language disorder; TD = typically developing.

681 **Figure 5**

682 *Summary forest plot showing the pooled effect size for each subgroup as part of the*
 683 *subgroup analysis ran for language of task*



684 *Note.* Datapoints are presented in order of effect. Points represent a calculation of the
 685 pooled estimate of effect size (Hedges' g) from a multilevel meta-analysis for each defined
 686 subgroup (surrounded by 95% confidence intervals). In the image, ' n ' refers to the number of
 687 datapoints included in each analysis (not number of studies)

688

689 **RQ8c. Time of scoring (live or offline)** – This subgroup analysis involved 72 effect
690 sizes (31 were excluded for not including the necessary information). The test for
691 subgroup difference showed no significant subgroup effect ($p = .20$), indicating that
692 there is no evidence that the difference in performance between groups of DLD and
693 TD children was influenced by children's productions being scored during the task or
694 after the session.

695 **RQ8d. Type of scoring (sentence binary, sub-sentence binary, target binary or
696 error scoring)** – There were two separate subgroup analyses run here, both
697 including 97 effect sizes (6 were excluded for not including the necessary
698 information) . The first compared each of the four categorised types of scoring
699 (sentence binary, sub-sentence binary, target scoring, and error scoring). This test
700 for subgroup difference did not reveal a significant subgroup effect ($p = .92$). The
701 second subgroup analysis compared sentence binary scoring to the three other types
702 of scoring combined. This analysis also did not show a significant subgroup effect (p
703 $= .88$). These results suggest that the type of scoring used on SR tasks does not
704 influence the size of the difference in performance between groups of DLD and TD
705 children.

706 **RQ8e. Language of the task** – This subgroup analysis involved all 103 effect sizes.
707 The test for subgroup differences showed a significant subgroup effect of language
708 ($p = .014$), suggesting that the language of the SR task did influence the size of the
709 difference in performance between groups of children with DLD and TD children.

710 **RQ8f. DLD sample recruited (clinical or population)** – This subgroup analysis
711 involved 102 effect sizes (1 was excluded for not including the necessary
712 information). The test for subgroup difference showed there was no significant
713 subgroup effect ($p = .09$), indicating that there is no evidence that difference in
714 performance between groups of DLD and TD children was affected by the sample of
715 children with DLD being of clinical or population origin.

716 **RQ8g. Matching of TD children (age- or language-matched)** – This subgroup
717 analysis involved 95 effect sizes (8 were excluded for not including the necessary
718 information). The test for subgroup differences revealed a significant subgroup effect
719 of matching ($p < .0001$). This suggests that the size of the difference in performance
720 between groups of children with DLD and TD children is influenced by the type of
721 matching, with the size of the effect being larger when TD children were matched by
722 age ($g = -2.27$) compared to when children were matched for language level ($g = -$
723 1.34). It is important to note that the average effect size for language matched
724 studies was still significantly greater than 0 ($p = .0046$).

725 The significant subgroup effect found in Q8g (for age versus language matching) was
726 further assessed with an exploratory analysis. The same subgroup analyses conducted for
727 RQ8a-f were run again separately for studies which just used age-matching. This was to
728 assess whether any of the previous analyses were influenced by matching as a confounding
729 variable. The results of this exploratory analysis can be found in supplementary materials
730 S1. None of the subgroup analyses run with the age-matched studies revealed a significant
731 subgroup difference. Unlike for the composite data, a subgroup analysis ran for language did
732 not reveal a significant effect ($p = .11$), suggesting that the significant effect found before
733 was confounded by the type of matching used in the studies.

734 This was not run for language-matched studies as it was deemed that there was not
735 enough variation among studies, with only eight studies using language matching which
736 were included in the meta-analysis. Table 5 summarises the key features and effect sizes of
737 these eight studies.

738 **Reporting Bias Assessment**

739 The possible presence of publication bias in our data was assessed. A funnel plot
740 and the results of Egger's regression test can be found in appendix C. There was some
741 evidence of asymmetry in our data which was further investigated by assessing the impact of
742 potential outliers and small-study effect. Due to the multilevel structure of the meta-analysis
743 and the high level of heterogeneity found, other methods of assessment (e.g., the 'trim-and-

744 **Table 5**

745 *Summary of study characteristics for studies included in the meta-analysis which involved*
 746 *TD children matched to children with DLD by language level*

Study	Type of Language Matching	Language of SR task	Type of Task	Type of Scoring	Effect Size
Riches (2012)	MLU in words	English	Original	Error Scoring	-2.64
Riches (2017)	MLU in words	English	Original	Error Scoring (Non-canonical)	-2.51
				Error Scoring (Canonical)	-2.1
Smolík et al. (2021)	Receptive vocabulary scores	Czech	Original	Error Scoring (Verb inflection)	-1.78
				Error Scoring (Noun inflection)	-1.55
Riches (2015)	MLU in words	English	Original	Target Binary	-1.47
Smolik & Vavru (2014)	Receptive vocabulary score and verbal memory	Czech	Original	Sentence Binary	-1.4
Leroy et al. (2013)	Sentence comprehension abilities	French	Original	Sentence Binary	-1.22
Abel et al. (2015)	MLU in morphemes	English	Original	Sentence Binary	-0.14
Stokes et al. (2006)	Receptive Grammar Scores	Cantonese	Original	Target Binary (Aspect)	-0.58
				Error Scoring (Aspect)	-0.21
				Sentence Binary (Passive)	-0.19
				Sentence Binary (Aspect)	-0.1
				Error Scoring (Passive)	0.01
				Target Binary (Passive)	0.04
				Sub-sentence Binary (Aspect)	0.08
				Sub-sentence Binary (Passive)	0.12

747 *Note.* Effect size here is a calculation of SMD using Hedges' g (Hedges, 1981). A negative
 748 effect size indicates that the children with DLD performed with less accuracy (lower score)
 749 on the task than those who are TD (higher score). SR = sentence repetition; MLU = mean
 750 length of utterance.

751 fill' method) were not looked to. To detect potential outliers, Cook's distances (Cook &
752 Weisberg, 1982) were calculated for each datapoint. Studies with the highest Cook's
753 distance were removed until the asymmetry (as calculated through Egger's regression) was
754 no longer evident, which resulted in four effect sizes being removed. This removal of outliers
755 resulted in an updated average effect size of $g = -2.03$ (95% CI [-2.26, -1.81] and a funnel
756 plot and Egger's regression test also shown in appendix C. The change in effect is minimal
757 when compared to our original effect size of $g = -2.08$, showing the effect size to be
758 insensitive to the influence of small study effect. Because of this, these potential outliers
759 remained in our final reported analyses.

760

Discussion

761 This article has explored the differences in performance on SR tasks by groups of
762 DLD and TD monolingual children in a systematic review of 66 studies and a multilevel
763 meta-analysis of 46 studies. Substantial methodological diversity was observed. Studies in
764 the review spanned 19 languages, 37 tasks (18 of which were original to their research
765 studies) and an age range of 14 years (with children aged between 2;7 to 16;7). Despite
766 these variations, the finding across the studies was that there is a robust difference in the
767 performance of children with DLD in comparison to TD children on SR tasks. Our meta-
768 analysis revealed this to be a large effect, insensitive to potential small study effects, with TD
769 children across the studies outperforming children with DLD on the tasks by 2.08 SDs. This
770 was while accounting for the dependencies which may have occurred due to some studies
771 contributing multiple effect sizes to the meta-analysis in our multilevel model.

772 As McGregor (2020) points out, to be of clinical use a tool must be able to detect
773 cases of disorder (sensitivity) and its absence (specificity). Diagnostic accuracy metrics were
774 reported in 18 of the studies. Of the values provided, the majority (75.86%) indicated
775 acceptable levels of sensitivity and specificity (above 80% for both values) and can be
776 viewed as having fair (to good) diagnostic accuracy when following the recommendations set
777 out by Plante and Vance (1994). On the other hand, this meant that 24.14% of the values
778 reported in the studies included show poor sensitivity and specificity (under 80% on at least

779 one value). If applied in a clinical context this could cause harm by either misdiagnosing a
780 child with DLD (false positive) or missing a diagnosis (false negative). The authors of a
781 particularly low specificity (57%) study (Pham & Ebert, 2020) suggest that their SR task (with
782 binary scoring) could present a quick and effective screening tool to identify those in need of
783 further testing, rather than acting as a diagnostic test.

784 Across the studies, 44.83% of diagnostic values reported showed $LR+ > 10$ and $LR-$
785 < 0.1 , suggesting that in these cases a child with DLD was more than ten times as likely to
786 score below the specified cut off on the task than a TD child and less than 0.1 times as likely
787 to score above the cut off than a TD child. In these cases, SR shows strong evidence
788 (Deeks et al., 2004) of identifying those with and without DLD and can be considered to have
789 good discriminative ability. All likelihood ratios reported showed an association between
790 productions and the presence or absence of DLD. Our observations therefore show that
791 while SR cannot be recommended as a stand-alone task and tool in DLD diagnosis (though
792 note that no single task should be used to confirm a diagnosis), SR tasks can effectively
793 contribute to a decision on diagnosis in combination with other assessments.

794 Our multilevel meta-analysis found a very large effect size of $g = -2.08$ (95% CI [-
795 2.32, -1.84]) for the difference in performance between groups of children with DLD and
796 groups of TD children. This is a larger effect size than reported for meta-analyses looking at
797 other methods of identifying children with DLD when compared with subgroups of TD
798 children. A meta-analysis by Winters et al. (2022) looked at narrative performance ($g = -0.82$
799 (95% CI [-0.99, -0.66])), and there have been two meta-analyses to date looking specifically
800 at nonword repetition (Schwob et al., 2021 and Estes et al., 2007; $g = 1.57$ (95% CI [1.37,
801 1.72]) and $d = 1.27$ (95% CI [1.15, 1.39]) respectively). Note however, that Winters et al.
802 (2022), and Schwob et al. (2021) did not exclude studies and results from bilingual
803 populations, whereas our review and meta-analysis did. From the available evidence, SR
804 appears to be the best available means of discriminating children with DLD from typically
805 developing children. SR provides a test of lexical phonology and morphosyntax (Polišenská
806 et al., 2015), with each repetition requiring short-term memory and prior language knowledge

807 to process, store and regenerate the sentences. This overall reflection of language ability is
808 likely what sets SR apart from alternative methods, as it targets areas in which those with
809 DLD are impaired.

810 Multiple subgroup analyses were run to look at the influence of different factors on
811 the size of this effect. No difference was found based on a number of these factors –
812 whether tasks were standardised, whether sentences were pre-recorded or produced live,
813 whether scoring was online or offline, the type of scoring used, use of a clinical or a
814 population sample of children with DLD. This lack of systematic variability suggests SR to be
815 a robust tool, strong enough to differentiate the performance of those with and without DLD
816 despite methodological and sample differences. It is important that to be of use clinically, SR
817 tasks must be able to accurately detect language disorder, while also being simple enough
818 in design and application to provide an efficient and reliable process. As such, this improves
819 the practicality of SR tasks as they can be adapted to the needs of the specific sample and
820 situation with minimal risk of reduced discriminative value.

821 In looking at variation in how SR tasks were administrated, no meaningful difference
822 was found in performance as a function of stimuli delivery – sentences being pre-recorded or
823 produced live by the task administrator. Delivery can therefore be adapted to the sample
824 based on factors such as age (there is evidence that presenting sentences in a live voice
825 aids in engaging children with repetition tasks; Frizelle et al., 2017). By contrast, pre-
826 recording stimuli and presenting them over headphones might be preferred where possible
827 as it allows for consistency and better quality of input (Armon-Lotem et al., 2015).

828 The review also saw a variety of scoring methods used in the evaluation of SR
829 performance, encompassing four categories. These methods can be divided into four
830 classes – sentence binary, sub-sentence binary, target binary, and error scoring. Again, as
831 part of the meta-analysis no significant difference was found in effect when comparing all
832 four types. For clinical use, arguably the most efficient way of scoring is the sentence binary
833 method (Hamaan & Abed Ibrahim, 2017), allowing for quick and easy assessments of
834 performance. It is also likely to be the most reliable in implementation, with Ebert et al.

835 (2019) finding that even those without a background in language assessment could reliably
836 score SR performance if a binary scoring system was used. Target scoring on the other
837 hand, can provide the most detail (Komeli & Marshall, 2013) and can be used to gain further
838 insight into the specific language struggles a child may have. There is some discussion of
839 the relative value of the different methods in the literature. Hamaan and Abed Ibrahim
840 (2017), Taha et al. (2021) and Theodorou et al. (2017) found little to no difference in the
841 sensitivity and specificity values achieved across scoring methods. Pham and Ebert (2020),
842 and Wang et al. (2022) found better specificity when productions are scored using error
843 scoring rather than binary scoring, with Wang et al. concluding that the error method of
844 scoring provides in-depth information on children's language ability and, due to its efficiency,
845 binary scoring should only be used when time is a factor in evaluating performance.
846 However, we found no meaningful difference between scoring method in our meta-analysis,
847 indicating that scoring can be adapted to the needs and information required from the task.

848 A significant subgroup effect was found for how DLD and TD groups were matched –
849 the size of the difference between groups was significantly larger when TD children were
850 matched to those with DLD by age, by comparison to when TD children were matched to
851 those with DLD by language ability. However, while the effect was smaller, the overall effect
852 size across studies which compared DLD performance to language-matched TD groups
853 remained large ($g = -1.31$ (95% CI [-2.0360, -0.5918]), with children with DLD showing less
854 accurate SR performance in comparison to younger, language-matched children. In a clinical
855 context children would be compared to those of a similar age, with standardised SR tasks
856 such as the CELF (Wiig et al., 2013) having norm-referenced comparisons for age.
857 However, this remains an important finding because a task distinguishing age-matched DLD
858 and TD children may just target general language properties that a child with other
859 impairments including language delay would perform poorly on when compared to children
860 of the same age (Van der Lely & Howard, 1993). In distinguishing between those with DLD
861 and language-matched, younger, TD children, a SR task is likely to be targeting the specific
862 structures which cause low performance in DLD specifically, leading to more crude individual

863 differences. Riches (2012) concluded that their finding that children with DLD perform
864 significantly worse than language-matched controls is indeed strong validating evidence of
865 the use of SR as a clinical marker.

866 There were a limited number of studies that used language-matched control groups,
867 with only eight studies contributing data for the meta-analysis with language-matched
868 groups, coming from five independent research teams. This highlights a key area of future
869 research in looking to SR tasks in relation to children with DLD and language-matched TD
870 children to further explore differences in SR performance and perhaps even shed more light
871 on the nature of DLD itself.

872 It is also important to note that while age matching is simple to perform — across the
873 studies this was generally performed on a group-level and involved, for example, children in
874 the same school year — language matching is less than straightforward, in part because
875 language is a multidimensional skill. Of the eight studies included in the meta-analysis which
876 used language-matching, four matched for language based upon mean length of utterance
877 (MLU) either in words or morphemes (all English tasks). Two matched on receptive
878 vocabulary (both Czech tasks), one on receptive grammar (Cantonese task), and one on
879 sentence comprehension (French task). There is limited evidence present to consider the
880 influence that the language profiles of children may have on SR performance differences
881 between groups of children with DLD and TD children. Indeed, type of language matching for
882 these eight studies appears confounded by language of task, with all four of the studies
883 matching by MLU being conducted in English. Considering the different types of language
884 matching with the same sample of children with DLD presents an interesting avenue of
885 research.

886 The language of the SR task also was found to significantly impact the size of the
887 effect in a subgroup analysis. However, when this analysis was rerun with just studies who
888 used age-matched TD groups, this effect disappears, suggesting that there was a
889 confounding effect of how TD children were matched. It can therefore be tentatively
890 concluded that SR tasks reliably result in a difference in performance between DLD and TD

891 groups across different languages (in monolingual children), even those with a vastly
892 different morphosyntactic structure to English, such as Arabic (Alsiddiqi et al., 2021; Taha et
893 al., 2021). This may be in part due to the standardising influence of COST Action IS0804
894 “Language Impairment in a Multilingual Setting: Linguistic Pattern and the Road to
895 Assessment”, and the LITMUS-SRep task (Marinis & Armon-Lotem, 2015) that was
896 developed as part of the project. LITMUS was a collaborative effort to develop methods of
897 language assessment (including a SR task) which can identify DLD within a bilingual setting.
898 The most frequent task seen was those developed following LITMUS-SRep principles (and
899 used here in a monolingual setting) which propose that the sentences used should differ in
900 the grammatical structures known to be difficult to those with DLD across languages (e.g.,
901 relative clauses) as well as the language specific to the task. Global collaborations such as
902 this may be important for the development of SR tasks in the future and to promote a more
903 standardised use across clinical and research contexts.

904 **Limitations**

905 While reliability in terms of transcription and scoring appeared high across studies, a
906 paper included in our systematic review was roughly only as likely to have reported on
907 reliability (51.52%) as it was to have not broached the topic at all. This is surprising given
908 that the transcription and scoring of the children’s responses relied entirely on judgement by
909 coders. Indeed, transcriptions for speech produced by children generally shows lower levels
910 of inter-transcriber agreement compared to the transcription of adult speech (Stoel-
911 Gammon, 2001). While included studies generally focused on language and not speech,
912 accuracy in transcriptions/scoring cannot be assumed.

913 This lack of detail was a consistent challenge when addressing our research
914 questions. Many studies were unable to be included in some of the subgroup analyses ran
915 due to inconsistent reporting of key methodological features. For example, some included
916 studies failed to specify the origin of the SR task used, and others failed to describe
917 methodological factors such as where and how productions were scored.

918 Looking to study quality and scores on the Standard Quality Assessment Criteria for
919 quantitative studies (Kmet et al., 2004), many studies scored poorly on points relating to
920 sample size and estimates of variance being reported in the results. Indeed, 49 of the 66
921 studies involved at least one participant group with under 20 children in it. While limitations
922 such as low sample size are to be expected with clinical samples of DLD, it is important to
923 note that many of the included studies were likely underpowered.

924 Further to this, high heterogeneity was seen across the studies. While the subgroup
925 analyses were conducted to explore differences across the studies, it is likely there was
926 some residual confounding. As previously explored, this was seen with type of matching
927 (age vs. language) and the language of the task. There were likely confounding influences
928 occurring in addition to this. For example: studies included in the meta-analysis with
929 language-matched TD groups only used original non-standardised tasks; studies which used
930 standardised tasks were more likely to score productions online; countries have different
931 agreed clinical definitions of DLD, and this may have been reflected in the results by
932 language of the task.

933 **Conclusions and Clinical Implications**

934 This study examined the literature on the use of SR tasks in identifying monolingual
935 children with DLD in a systematic review and novel multilevel meta-analysis which
936 accounted for dependencies from studies contributing multiple effect sizes. The review
937 identified a number of key points of variation in the application of SR tasks relating to the
938 types of tasks used, types of scoring used and languages the task is seen in. Nonetheless,
939 our meta-analysis indicated that SR tasks can discriminate between children with DLD and
940 both age- and language-matched TD children. The effect was large across the studies and
941 appears robust to most sample and study variation. There is evidence therefore, that within a
942 clinical setting, SR tasks can be adapted to practical constraints, while still accurately
943 discriminating performance between monolingual children with DLD and TD children.

944

945

946

Acknowledgements

947 This systematic review and meta-analysis were completed as parts of Leah Ward's PhD
948 research at The University of Manchester. Sponsorship and funding for the PhD and all
949 resulting publications is provided by the North West Social Science Doctoral Training
950 Partnership Economic and Social Research Council.

Data Availability Statement

952 Complete data extraction files used for both the systematic review and meta-analysis, and R
953 code used to conduct the meta-analysis are available on the Open Science Framework at
954 <https://osf.io/usw2k/>.

955

References

- 956 Abel, A. D., Rice, M. L., & Bontempo, D. E. (2015). Effects of verb familiarity on finiteness
957 marking in children with specific language impairment. *Journal of Speech, Language,
958 and Hearing Research, 58*(2), 360–372. https://doi.org/10.1044/2015_jslhr-l-14-0003
- 959 Acosta-Rodríguez, V. M., Ramírez-Santana, G. M., Hernández Expósito, S., & Axpe
960 Caballero, Á. (2020). Intervention in syntactic skills in pupils with developmental
961 language disorder. *Psicothema, 32*(4), 541–548.
962 <https://doi.org/10.7334/psicothema2020.160>
- 963 Alsiddiqi, Z. A., Stojanovik, V., & Pagnamenta, E. (2021). Emergent literacy skills of Saudi
964 Arabic speaking children with and without developmental language disorder. *Clinical
965 Linguistics & Phonetics, 36*(4–5), 301–318.
966 <https://doi.org/10.1080/02699206.2021.1955299>
- 967 Archibald, L. M. D., & Joanisse, M. F. (2009). On the Sensitivity and Specificity of Nonword
968 Repetition and Sentence Recall to Language and Memory Impairments in Children.
969 *Journal of Speech, Language, and Hearing Research, 52*(4), 899-914.
970 <https://doi.org/10.1044/1092-4388>
- 971 Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.). (2015). *Assessing multilingual children:
972 Disentangling bilingualism from language impairment* (Vol. 13). Multilingual matters.

- 973 Armon-Lotem, S., & Meir, N. (2016). Diagnostic accuracy of repetition tasks for the
974 identification of specific language impairment (SLI) in bilingual children: evidence
975 from Russian and Hebrew. *International Journal of Language & Communication*
976 *Disorders*, 51(6), 715-731. <https://doi.org/10.1111/1460-6984.12242>
- 977 Benavides, A. A., Kapantzoglou, M., & Murata, C. (2018). Two grammatical tasks for
978 screening language abilities in Spanish-speaking children. *American Journal of*
979 *Speech Language Pathology*, 27(2), 690–705.
980 https://doi.org/10.1044/2017_AJSLP17-0052
- 981 Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T. & the CATALISE-2
982 Consortium. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary
983 Delphi consensus study of problems with language development: Terminology.
984 *Journal of Child Psychology and Psychiatry*, 58(10), 1068-1080.
985 <https://doi.org/10.1111/jcpp.12721>
- 986 Blom, E., & Boerma, T. (2019). Reciprocal relationships between lexical and syntactic skills
987 of children with Developmental Language Disorder and the role of executive
988 functions. *Autism & Developmental Language Impairments*, 4.
989 <https://doi.org/10.1177/2396941519863984>
- 990 Calder, S. D., Brennan-Jones, C. G., Robinson, M., Whitehouse, A., & Hill, E. (2022). The
991 prevalence of and potential risk factors for Developmental Language Disorder at 10
992 years in the Raine Study. *Journal of Paediatrics and Child Health*, 58(11),
993 2044-2050. <https://doi.org/10.1111/jpc.16149>
- 994 Caselli, M. C., Monaco, L., Trasciani, M., & Vicari, S. (2008). Language in Italian Children
995 with Down Syndrome and with Specific Language Impairment. *Neuropsychology*,
996 22(1), 27-35. <https://doi.org/10.1037/0894-4105.22.1.27>
- 997 Chevrie-Muller, C., Simon, A. M., Fournier, S., & Brochet, M. O. (2010). *Batterie langage*
998 *oral-langage écrit, mémoire-attention: L2ma* (2nd ed.). ECPA.
- 999 Christensen, R. V., & Hansson, K. (2012). The Use and Productivity of Past Tense
1000 Morphology in Specific Language Impairment: An Examination of Danish. *Journal of*

- 1001 *Speech Language and Hearing Research*, 55(6), 1671-1689.
1002 [https://doi.org/10.1044/1092-4388\(2012/10-0350\)](https://doi.org/10.1044/1092-4388(2012/10-0350))
- 1003 Christensen, R. V., Jensen, S. T. & Nielsen, I. I. (2012). *Sætningsgentagelsestesten [The*
1004 *Sentence Repetition Test]*. Institut for Nordiske Studier og Sprogvidenska
1005 [Department of Nordic Studies and Linguistics], Københavns Universitet [University of
1006 Copenhagen].
- 1007 Chu, K. (1999). An introduction to sensitivity, specificity, predictive values and likelihood
1008 ratios. *Emergency Medicine*, 11(3), 175-181.
1009 <https://doi.org/10.1046/j.14422026.1999.00041.x>
- 1010 Coady, J. A., Evans, J. L., & Kluender, K. R. (2010). The Role of Phonotactic Frequency in
1011 Sentence Repetition by Children With Specific Language Impairment. *Journal of*
1012 *Speech Language and Hearing Research*, 53(5), 1401-1415.
1013 [https://doi.org/10.1044/1092-4388\(2010/07-0264\)](https://doi.org/10.1044/1092-4388(2010/07-0264))
- 1014 Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*,
1015 10(4), 417-451. <https://doi.org/10.2307/3001616>
- 1016 Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific
1017 language impairment (SLI). *Journal of Child Psychology and Psychiatry*, 42(6), 741-
1018 748. <https://doi.org/10.1111/1469-7610.00770>
- 1019 Cook, R. D., & Weisberg, S. (1982). Criticism and Influence Analysis in Regression.
1020 *Sociological Methodology*, 13, 313. <https://doi.org/10.2307/270724>
- 1021 de Almeida, L. de, Ferré, S., Morin, E., Prévost, P., Santos, C. dos, Tuller, L., Zebib, R., &
1022 Barthez, M. A. (2017). Identification of bilingual children with Specific Language
1023 Impairment in France. *Language Impairment in Bilingual Children*, 7(3-4), 331-358.
1024 <https://doi.org/10.1075/lab.15019.alm>
- 1025 Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *Bmj*, 329(7458),
1026 168-169. <https://doi.org/10.1136/bmj.329.7458.168>

- 1027 Delage, H., & Frauenfelder, U. (2012). Développement de la mémoire de travail et traitement
1028 des phrases complexes: Quelle relation? *SHS Web of Conferences*, 1, 1555-1573.
1029 <https://doi.org/10.1051/shsconf/20120100141>
- 1030 Delage, H., & Frauenfelder, U. H. (2020). Relationship between working memory and
1031 complex syntax in children with developmental language disorder. *Journal of Child*
1032 *Language*, 47(3), 600–632. <https://doi.org/10.1017/s0305000919000722>
- 1033 Delage, H., Stanford, E., & Durrleman, S. (2021). Working memory training enhances
1034 complex syntax in children with Developmental Language Disorder. *Applied*
1035 *Psycholinguistics*, 42(5), 1341-1375, [http://doi.org/ 10.1017/S0142716421000369](http://doi.org/10.1017/S0142716421000369)
- 1036 Delcenserie, A., Genesee, F., Trudeau, N., & Champoux, F. (2019). A multi-group approach
1037 to examining language development in at-risk learners. *Journal of Child Language*,
1038 46(1), 51–79. <https://doi.org/10.1017/S030500091800034X>
- 1039 Devescovi, A., & Caselli, M. C. (2001). Una prova di ripetizione di frasi per la valutazione del
1040 primo sviluppo grammaticale. *Psicologia clinica dello sviluppo*, (3), 341-364.
- 1041 Devescovi, A., Caselli, M. C. and Ossella, T. (1992). Rilevazione delle prime fasi dello
1042 sviluppo morfosintattico attraverso una prova di ripetizione. *Rassegna di Psicologia*,
1043 2, 25–42.
- 1044 Dosi, I., & Koutsipetsidou, E.-C. (2019). Measuring linguistic and cognitive abilities by means
1045 of a sentence repetition task in children with developmental dyslexia and
1046 developmental language disorder. *European Journal of Research in Social Sciences*,
1047 7(4), 10–19.
- 1048 Duman, T. Y., Blom, E., & Topbas, S. (2015). At the Intersection of Cognition and Grammar:
1049 Deficits Comprehending Counterfactuals in Turkish Children With Specific Language
1050 Impairment. *Journal of Speech Language and Hearing Research*, 58(2), 410-421.
1051 https://doi.org/10.1044/2015_jslhr-l-14-0054
- 1052 Eadie, P. A., Fey, M. E., Douglas, J. M., & Parsons, C. L. (2002). Profiles of grammatical
1053 morphology and sentence imitation in children with specific language impairment and

- 1054 Down syndrome. *Journal of Speech, Language, and Hearing Research*, 45(4), 720–
1055 732. [https://doi.org/10.1044/1092-4388\(2002/058\)](https://doi.org/10.1044/1092-4388(2002/058))
- 1056 Ebert, K. D., Rak, D., Slawny, C. M., & Fogg, L. (2019). Attention in bilingual children with
1057 developmental language disorder. *Journal of Speech, Language, and Hearing
1058 Research*, 62(4), 979-992. https://doi.org/10.1044/2018_JSLHR-L-18-0221
- 1059 Engberg-Pedersen, E., & Christensen, R. V. (2017). Mental states and activities in Danish
1060 narratives: Children with autism and children with language impairment. *Journal of
1061 Child Language*, 44(5), 1192-1217. <https://doi.org/10.1017/S0305000916000507>
- 1062 Estes, K. G., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition
1063 performance of children with and without specific language impairment: A meta-
1064 analysis. *Journal of Speech, Language, and Hearing Research*, 50(1), 177-195.
1065 [https://doi.org/10.1044/1092-4388\(2007/015\)](https://doi.org/10.1044/1092-4388(2007/015))
- 1066 Fleckstein, A., Prevost, P., Tuller, L., Sizaret, E., & Zebib, R. (2018). How to identify SLI in
1067 bilingual children: A study on sentence repetition in French. *Language Acquisition*,
1068 25(1), 85-101. <https://doi.org/10.1080/10489223.2016.1192635>
- 1069 Foltz, A., Thiele, K., Kahsnitz, D., & Stenneken, P. (2015). Children's syntactic-priming
1070 magnitude: Lexical factors and participant characteristics. *Journal of Child Language*,
1071 42(4), 932-945. <https://doi.org/https://dx.doi.org/10.1017/S0305000914000488>
- 1072 Frizelle, P., & Fletcher, P. (2014a). Relative clause constructions in children with specific
1073 language impairment. *International Journal of Language & Communication
1074 Disorders*, 49(2), 255-264. <https://doi.org/10.1111/1460-6984.12070>
- 1075 Frizelle, P., & Fletcher, P. (2014b). Profiling relative clause constructions in children with
1076 specific language impairment. *Clinical Linguistics & Phonetics*, 28(6), 437-449.
1077 <https://doi.org/10.3109/02699206.2014.882991>
- 1078 Frizelle, P., O'Neill, C., & Bishop, D. V. (2017). Assessing understanding of relative clauses:
1079 A comparison of multiple-choice comprehension versus sentence repetition. *Journal
1080 of Child Language*, 44(6), 1435-1457. <https://doi.org/10.1017/S0305000916000635>

- 1081 Gagiano, S., & Southwood, F. (2015). The use of digit and sentence repetition in the
1082 identification of language impairment: The case of child speakers of Afrikaans and
1083 South African English. *Stellenbosch Papers in Linguistics*, 44, 37–60.
1084 <https://doi.org/10.5774/44-0-187>
- 1085 Garraffa, M., Coco, M. I., & Branigan, H. P. (2015). Effects of Immediate and Cumulative
1086 Syntactic Experience in Language Impairment: Evidence from Priming of Subject
1087 Relatives in Children with SLI. *Language Learning and Development*, 11(1), 18-40.
1088 <https://doi.org/10.1080/15475441.2013.876277>
- 1089 Georgiou, N., & Spanoudis, G. (2021). Developmental language disorder and autism:
1090 Commonalities and differences on language. *Brain Sciences*, 11(5), Article 589.
1091 <https://doi.org/10.3390/brainsci11050589>
- 1092 Grimm, H., & Schöler, H. (1977). *Heidelberger Sprachentwicklungstest (HSET)*. Hogrefe.
1093 Grimm, H., & Schöler, H. (1991). *Heidelberger Sprachentwicklungstest (H-S-E-T)*. Hans
1094 Huber Verlag.
- 1095 Håkansson, G., & Hansson, K. (2000). Comprehension and production of relative clauses: A
1096 comparison between Swedish impaired and unimpaired children. *Journal of Child*
1097 *Language*, 27(2), 313–333. <https://doi.org/10.1017/s0305000900004128>
- 1098 Hamann, C., & Abed Ibrahim, L. (2017). Methods for identifying specific language
1099 impairment in bilingual populations in Germany. *Frontiers in Communication*, 2,
1100 Article 16. <https://doi.org/10.3389/fcomm.2017.00016>
- 1101 Hannus, S., Kauppila, T., & Launonen, K. (2009). Increasing prevalence of specific
1102 language impairment (SLI) in primary healthcare of a Finnish town, 1989–99.
1103 *International journal of language & communication disorders*, 44(1), 79-97.
1104 <https://doi.org/10.1080/13682820801903310>
- 1105 Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related
1106 estimators. *Journal of Educational Statistics*, 6(2), 107-128.
1107 <https://doi.org/10.3102/10769986006002107>

- 1108 Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis.
1109 *Statistics in medicine*, 21(11), 1539-1558. <https://doi.org/10.1002/sim.1186>
- 1110 Hutchinson, E., Bavin, E., Efron, D., & Sciberras, E. (2012). A comparison of working
1111 memory profiles in school-aged children with specific language impairment, attention
1112 deficit/hyperactivity disorder, comorbid SLI and ADHD and their typically developing
1113 peers. *Child Neuropsychology*, 18(2), 190–207.
1114 <https://doi.org/10.1080/09297049.2011.601288>
- 1115 Kamhi, A. G., & Catts, H. W. (1986). Toward an understanding of developmental language
1116 and reading disorders. *The Journal of speech and hearing disorders*, 51(4), 337-347.
1117 <https://doi.org/10.1044/jshd.5104.337>
- 1118 Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C.
1119 (2015). Sentence repetition is a measure of children's language skills rather than
1120 working memory limitations. *Developmental science*, 18(1), 146-154.
1121 <https://doi.org/10.1111/desc.12202>
- 1122 Kmet, L. M., Lee, R. C., & Cook, L. S. (2004) Standard quality assessment criteria for
1123 evaluating primary research papers from a variety of fields. *Alberta Heritage*
1124 *Foundation for Medical Research*, 13, 1–22.
- 1125 Komeili, M., & Marshall, C. R. (2013). Sentence repetition as a measure of morphosyntax in
1126 monolingual and bilingual children. *Clinical Linguistics & Phonetics*, 27(2), 152-162.
1127 <https://doi.org/10.3109/02699206.2012.751625>
- 1128 Korkman, M., Kirk, U., & Kemp, S. L. (2007). *NEPSY-II: A developmental*
1129 *neuropsychological assessment*. The Psychological Corporation.
- 1130 Kueser, J. B., & Leonard, L. B. (2020). The Effects of Frequency and Predictability on
1131 Repetition in Children With Developmental Language Disorder. *Journal of Speech*
1132 *Language and Hearing Research*, 63(4), 1165-1180.
1133 https://doi.org/10.1044/2019_jslhr-19-00155
- 1134 Lalioti, M., Stavrakaki, S., Manouilidou, C., & Talli, I. (2016). Subject–verb agreement and
1135 verbal short-term memory: A perspective from Greek children with specific language

- 1136 impairment. *First Language*, 36(3), 279–294.
1137 <https://doi.org/10.1177/0142723716648844>
- 1138 Leclercq, A. L., Quemart, P., Magis, D., & Maillart, C. (2014). The sentence repetition task: A
1139 powerful diagnostic tool for French children with specific language impairment.
1140 *Research in Developmental Disabilities*, 35(12), 3423-3430.
1141 <https://doi.org/https://dx.doi.org/10.1016/j.ridd.2014.08.026>
- 1142 Leroy, S., Parrisé, C., & Maillart, C. (2013). The influence of the frequency of functional
1143 markers on repetitive imitation of syntactic constructions in children with specific
1144 language impairment, from their own language productions. *Clinical Linguistics &*
1145 *Phonetics*, 27(6-7), 508-520. <https://doi.org/10.3109/02699206.2013.787546>
- 1146 Lukacs, A., Kas, B., & Leonard, L. B. (2013). Case marking in Hungarian children with
1147 specific language impairment. *First Language*, 33(4), 331-353.
1148 <https://doi.org/10.1177/0142723713490601>
- 1149 Marinis, T., & Armon-Lotem, S. (2015). Sentence Repetition. In Armon-Lotem, S., de Jong,
1150 J. & Meir, N. (Eds.). *Methods for assessing multilingual children: disentangling*
1151 *bilingualism from Language Impairment*. Multilingual Matters.
- 1152 McGregor, K. K. (2020). How we fail children with developmental language disorder.
1153 *Language, speech, and hearing services in schools*, 51(4), 981-992.
1154 https://doi.org/10.1044/2020_LSHSS-20-00003
- 1155 Nag, S., Snowling, M. J., & Mirković, J. (2018). The role of language production mechanisms
1156 in children's sentence repetition: Evidence from an inflectionally rich language.
1157 *Applied Psycholinguistics*, 39(2), 303-325.
1158 <https://doi.org/10.1017/S0142716417000200>
- 1159 Nash, H. M., Hulme, C., Gooch, D., & Snowling, M. J. (2013). Preschool language profiles of
1160 children at family risk of dyslexia: Continuities with specific language impairment. The
1161 *Journal of Child Psychology and Psychiatry*, 54(9), 958–968.
1162 <https://doi.org/10.1111/jcpp.12091>

- 1163 Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development-Primary* (3rd
1164 edition). Pro-Ed.
- 1165 Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing In Noise Test
1166 for the measurement of speech reception thresholds in quiet and in noise. *The*
1167 *Journal of the Acoustical Society of America*, 95(2), 1085-1099.
1168 <https://doi.org/10.1121/1.408469>
- 1169 Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., &
1170 Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical
1171 presentation of language disorder: Evidence from a population study. *Journal of child*
1172 *psychology and psychiatry*, 57(11), 1247-1257. <https://doi.org/10.1111/jcpp.12573>
- 1173 Nudel, R., Christensen, R. V., Kalnak, N., Schwinn, M., Banasik, K., Dinh, K. M., ... & DBDS
1174 Genomic Consortium. (2023). Developmental language disorder—a comprehensive
1175 study of more than 46,000 individuals. *Psychiatry Research*, 323, 115171.
1176 <https://doi.org/10.1016/j.psychres.2023.115171>
- 1177 Oetting, J. B., McDonald, J. L., Seidel,
1178 C. M., & Hegarty, M. (2016). Sentence Recall by Children With SLI Across Two
1179 Nonmainstream Dialects of English. *Journal of Speech Language and Hearing*
Research, 59(1), 183-194. https://doi.org/10.1044/2015_jslhr-l-15-0036
- 1180 Orsolini, M., Sechi, E., Maronato, C., Bonvino, E., & Corcelli, A. (2001). Nature of
1181 phonological delay in children with specific language impairment. *International*
1182 *Journal of Language & Communication Disorders*, 36(1), 63-90.
1183 <https://doi.org/10.1080/13682820150217572>
- 1184 Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ...
1185 & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting
1186 systematic reviews. *Journal of Clinical Epidemiology*, 134, 178-189.
1187 <https://doi.org/10.1016/j.jclinepi.2021.03.001>
- 1188 Pawłowska, M. (2014). Evaluation of three proposed markers for language impairment in
1189 English: A meta-analysis of diagnostic accuracy studies. *Journal of Speech*,

- 1190 *Language, and Hearing Research*, 57(6), 2261-2273.
1191 https://doi.org/10.1044/2014_JSLHR-L-13-0189
- 1192 Peristeri, E., Andreou, M., Tsimpli, I. M., & Durrleman, S. (2021). Bilingualism effects in the
1193 narrative comprehension of children with Developmental Language Disorder and L2-
1194 Greek. In U. Bohnacker & N. Gagarina (Eds.), *Studies in bilingualism* (Vol. 61, pp.
1195 297–330). John Benjamins.
- 1196 Petruccelli, N., Bavin, E. L., & Bretherton, L. (2012). Children with specific language
1197 impairment and resolved late talkers: Working memory profiles at 5 years. *Journal of*
1198 *Speech, Language, and Hearing Research*, 55(6), 1690–1703.
1199 [https://doi.org/10.1044/1092-4388\(2012/11-0288\)](https://doi.org/10.1044/1092-4388(2012/11-0288))
- 1200 Pham, G., & Ebert, K. D. (2020). Diagnostic Accuracy of Sentence Repetition and Nonword
1201 Repetition for Developmental Language Disorder in Vietnamese. *Journal of Speech*
1202 *Language and Hearing Research*, 63(5), 1521-1536.
1203 https://doi.org/10.1044/2020_jslhr-19-00366
- 1204 Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based
1205 approach. *Language, Speech, and Hearing Services in Schools*, 25(1), 15-24.
1206 <https://doi.org/10.1044/0161-1461.2501.15>
- 1207 Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task
1208 measure? *International Journal of Language & Communication Disorders*, 50(1),
1209 106-118. <https://doi.org/10.1111/1460-6984.12126>
- 1210 Poll, G. H., Miller, C. A., & Van Hell, J. G. (2016). Sentence repetition accuracy in adults with
1211 developmental language impairment: Interactions of participant capacities and
1212 sentence structures. *Journal of Speech, Language, and Hearing Research*, 59(2),
1213 302-316. https://doi.org/10.1044/2015_JSLHR-L-15-0020
- 1214 Potter, M. C. (2012). Conceptual short term memory in perception and thought. *Frontiers in*
1215 *Psychology*, 3, 113. <https://doi.org/10.3389/fpsyg.2012.00113>

- 1216 Redmond, S. M. (2005). Differentiating SLI from ADHD using children's sentence recall and
1217 production of past tense morphology. *Clinical Linguistics & Phonetics*, 19(2), 109–
1218 127. <https://doi.org/10.1080/02699200410001669870>
- 1219 Redmond, S. M., & Ash, A. C. (2017). Associations between the 2D:4D proxy biomarker for
1220 prenatal hormone exposures and symptoms of developmental language disorder.
1221 *Journal of Speech, Language, and Hearing Research*, 60(11), 3226–3236.
1222 https://doi.org/10.1044/2017_jslhr-l-17-0143
- 1223 Redmond, S. M., Ash, A. C., Christopoulos, T. T., & Pfaff, T. (2019). Diagnostic accuracy of
1224 sentence recall and past tense measures for identifying children's language
1225 impairments. *Journal of Speech, Language, and Hearing Research*, 62(7), 2438–
1226 2454. https://doi.org/10.1044/2019_JSLHR-L-18-0388
- 1227 Redmond, S. M., Ash, A. C., & Hogan, T. P. (2015). Consequences of co-occurring
1228 attention-deficit/hyperactivity disorder on children's language impairments.
1229 *Language, Speech, and Hearing Services in Schools*, 46(2), 68–80.
1230 https://doi.org/10.1044/2014_lshss-14-0045
- 1231 Redmond, S. M., Thompson, H. L., & Goldstein, S. (2011). Psycholinguistic profiling
1232 differentiates specific language impairment from typical development and from
1233 attention deficit/hyperactivity disorder. *Journal of Speech, Language, and Hearing
1234 Research*, 54(1), 99–117. [https://doi.org/10.1044/1092-4388\(2010/10-0010\)](https://doi.org/10.1044/1092-4388(2010/10-0010))
- 1235 Riches, N. (2015). Past tense -ed omissions by children with specific language impairment:
1236 The role of sonority and phonotactics. *Clinical Linguistics & Phonetics*, 29(6), 482–
1237 497. <https://doi.org/10.3109/02699206.2015.1027832>
- 1238 Riches, N. G. (2012). Sentence repetition in children with specific language impairment: an
1239 investigation of underlying mechanisms. *International Journal of Language &
1240 Communication Disorders*, 47(5), 499-510. [https://doi.org/10.1111/j.1460-
1241 6984.2012.00158.x](https://doi.org/10.1111/j.1460-6984.2012.00158.x)

- 1242 Riches, N. G. (2017). Complex sentence profiles in children with Specific Language
1243 Impairment: Are they really atypical? *Journal of Child Language*, 44(2), 269-296.
1244 <https://doi.org/10.1017/s0305000915000847>
- 1245 Riches, N. G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2010). Sentence
1246 repetition in adolescents with specific language impairments and autism: An
1247 investigation of complex syntax. *International Journal of Language & Communication*
1248 *Disorders*, 45(1), 47–60. <https://doi.org/10.3109/13682820802647676>
- 1249 Royle, P., & Thordardottir, E. (2003). *Le grand déménagement [French adaptation of the*
1250 *Recalling Sentences in Context subtest of the CELF–P]*. Unpublished research tool,
1251 McGill University, Montreal, Quebec, Canada.
- 1252 Rujas, I., Mariscal, S., Murillo, E., & Lázaro, M. (2021). Sentence repetition tasks to detect
1253 and prevent language difficulties: A scoping review. *Children*, 8(7), 578.
1254 <https://doi.org/10.3390/children8070578>
- 1255 Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa,
1256 K. (2021). Using nonword repetition to identify developmental language disorder in
1257 monolingual and bilingual children: A systematic review and meta-analysis. *Journal*
1258 *of Speech, Language, and Hearing Research*, 64(9), 3578-3593.
1259 https://doi.org/10.1044/2021_JSLHR-20-00552
- 1260 Seeff-Gabriel, B., Chiat, S., & Dodd, B. (2010). Sentence imitation as a tool in identifying
1261 expressive morphosyntactic difficulties in children with severe speech difficulties.
1262 *International Journal of Language & Communication Disorders*, 45(6), 691–702.
1263 <https://doi.org/10.3109/13682820903509432>
- 1264 Seeff-Gabriel, B., Chiat, S., & Roy, P. (2008). *The early repetition battery*. Pearson
1265 Assessment.
- 1266 Semel, E., Wiig, E., & Secord, W. (1994). *Clinical Evaluation of Language Fundamentals*
1267 *Revised*. The Psychological Corporation.
- 1268 Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language*
1269 *Fundamentals, Fourth Edition (CELF–4)*. The Psychological Corporation.

- 1270 Semel, E., Wiig, H., Secord, W., & Langdon, W. (2006). *CELF 4: Clinical evaluation of*
1271 *language fundamentals 4: Spanish edition*. Psychological Corporation.
- 1272 Semel, E., Wiig, H., Secord, W., & Sabers, D. (1987). *CELF-R: Clinical Evaluation of*
1273 *Language Fundamentals – Revised (technical manual)*. Psychological Corporation.
- 1274 Smolík, F., Matiasovitsová, K., & Camarata, S. M. (2021). Sentence imitation with masked
1275 morphemes in Czech: Memory, morpheme frequency, and morphological richness.
1276 *Journal of Speech, Language, and Hearing Research*, 64(1), 105–120.
1277 https://doi.org/10.1044/2020_JSLHR-20-00370
- 1278 Smolík, F., & Vávrů, P. (2014). Sentence Imitation as a Marker of SLI in Czech:
1279 Disproportionate Impairment of Verbs and Clitics. *Journal of Speech Language and*
1280 *Hearing Research*, 57(3), 837-849. <https://doi.org/10.1044/2014>
- 1281 Spanoudis, G., & Pahiti, J. (2014). *Expressive and Receptive Language Evaluation: 5–12*
1282 *Years of Age*. Department of Psychology, University of Cyprus
- 1283 Stavrakaki, S., & Tsimpli, I. M. (2000). Diagnostic verbal IQ test for Greek preschool and
1284 school age children: Standardization, statistical analysis, psychometric properties. In
1285 *Proceedings of the 8th Symposium of the Panhellenic Association of Logopedists*
1286 (pp.95-106). Ellinika Grammata.
- 1287 Stoel-Gammon, C. (2001). Transcribing the speech of young children. *Topics in language*
1288 *disorders*, 21(4), 12-21.
- 1289 Stokes, S., & Fletcher, P. (2003). Aspectual forms in Cantonese children with specific
1290 language impairment. *Linguistics*, 41(2), 381-405.
1291 <https://doi.org/10.1515/ling.2003.013>
- 1292 Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and
1293 sentence repetition as clinical markers of specific language impairment: The case of
1294 Cantonese. *Journal of Speech, Language, and Hearing Research*, 49(2), 219–236.
1295 [https://doi.org/10.1044/1092-4388\(2006/019\)](https://doi.org/10.1044/1092-4388(2006/019))
- 1296 Taha, J., Stojanovik, V., & Pagnamenta, E. (2021). Sentence Repetition as a Clinical Marker
1297 of Developmental Language Disorder: Evidence From Arabic. *Journal of Speech*

- 1298 *Language and Hearing Research*, 64(12), 4876-4899.
- 1299 https://doi.org/10.1044/2021_jslhr-21-00244
- 1300 Talli, I., & Stavrakaki, S. (2020). Short-term memory, working memory and linguistic abilities
1301 in bilingual children with Developmental Language Disorder. *First Language*, 40(4),
1302 437–460. <https://doi.org/10.1177/0142723719886954>
- 1303 Taylor, L. J., Maybery, M. T., Grayndler, L., & Whitehouse, A. J. (2014). Evidence for distinct
1304 cognitive profiles in autism spectrum disorders and specific language impairment.
1305 *Journal of Autism and Developmental Disorders*, 44(1), 19-30.
1306 <https://doi.org/https://dx.doi.org/10.1007/s10803-013-1847-2>
- 1307 Theodorou, E., Kambanaros, M., & Grohmann, K. K. (2016). Diagnosing bilingual children
1308 with SLI: Determination of identification accuracy. *Clinical Linguistics & Phonetics*,
1309 30(12), 925–943. <https://doi.org/10.1080/02699206.2016.1182591>
- 1310 Theodorou, E., Kambanaros, M., & Grohmann, K. K. (2017). Sentence Repetition as a Tool
1311 for Screening Morphosyntactic Abilities of Bilingual Children with SLI. *Frontiers in*
1312 *Psychology*, 8, 2104. <https://doi.org/10.3389/fpsyg.2017.02104>
- 1313 Thordardottir, E., & Brandeker, M. (2013). The effect of bilingual exposure versus language
1314 impairment on nonword repetition and sentence imitation scores. *Journal of*
1315 *Communication Disorders*, 46(1), 1–16. <https://doi.org/10.1016/j.jcomdis.2012.08.002>
- 1316 Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., Trudeau,
1317 N., & Chilingaryan, G. (2011). Sensitivity and specificity of French language and
1318 processing measures for the identification of primary language impairment at age 5.
1319 *Journal of Speech, Language, and Hearing Research*, 54(2), 580–597.
1320 [https://doi.org/10.1044/1092-4388\(2010/09-0196\)](https://doi.org/10.1044/1092-4388(2010/09-0196))
- 1321 Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997).
1322 Prevalence of specific language impairment in kindergarten children. *Journal of*
1323 *speech, language, and hearing research*, 40(6), 1245-1260.
1324 <https://doi.org/10.1044/jslhr.4006.1245>
- 1325

- 1326 Tsimpli, I. M., Peristeri, E., & Andreou, M. (2016). Narrative production in monolingual and
1327 bilingual children with specific language impairment. *Applied Psycholinguistics*, 37(1),
1328 195–216. <https://doi.org/10.1017/S0142716415000478>
- 1329 Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prevost, P., dos Santos, C., Abed
1330 Ibrahim, L., & Zebib, R. (2018). Identifying language impairment in bilingual children
1331 in France and in Germany. *International Journal of Language & Communication*
1332 *Disorders*, 53(4), 888–904. <https://doi.org/10.1111/1460-6984.12397>
- 1333 Van der Lely, H. K., & Howard, D. (1993). Children With Specific Language Impairment:
1334 Linguistic Impairment or Short-Term Memory Deficit? *Journal of Speech, Language,*
1335 *and Hearing Research*, 36(6), 1193-1207. <https://doi.org/10.1044/jshr.3606.1193>
- 1336 Van Der Meulen, S., Janssen, P., & Os, E. D. (1997). Prosodic abilities in children with
1337 specific language impairment. *Journal of Communication Disorders*, 30(3), 155–170.
1338 [https://doi.org/10.1016/S0021-9924\(96\)00059-7](https://doi.org/10.1016/S0021-9924(96)00059-7)
- 1339 Vang Christensen, R. (2019). Sentence repetition: A clinical marker for developmental
1340 language disorder in Danish. *Journal of Speech, Language, and Hearing Research*,
1341 62(12), 4450–4463. https://doi.org/10.1044/2019_JSLHR-L-18-0327
- 1342 Verhoeven, L., & Vermeer, A. (2001). *Taaltoets Alle Kinderen (TAK)*. Cito.
- 1343 Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal*
1344 *of Statistical Software*, 36(3), 1-48. <https://doi.org/10.18637/jss.v036.i03>
- 1345 Wang, D., Zheng, L., Lin, Y., Zhang, Y., & Sheng, L. (2022). Sentence Repetition as a
1346 Clinical Marker for Mandarin-Speaking Preschoolers with Developmental Language
1347 Disorder. *Journal of Speech, Language, and Hearing Research*, 65(4), 1543-1560.
1348 https://doi.org/10.1044/2021_JSLHR-21-00401
- 1349 Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence-Revised*. The
1350 Psychological Corporation.
- 1351 Wiig E. H., Semel E., Secord W. A. (2013). *Clinical Evaluation of Language Fundamentals–*
1352 *Fifth Edition (CELF-5)*. Pearson.

- 1353 Winters, K. L., Jasso, J., Pustejovsky, J. E., & Byrd, C. T. (2022). Investigating narrative
1354 performance in children with developmental language disorder: A systematic review
1355 and meta-analysis. *Journal of Speech, Language, and Hearing Research*, 65(10),
1356 3908-3929. https://doi.org/10.1044/2022_JSLHR-22-00017
- 1357 Wu, S., Zhao, J., de Villiers, J., Liu, X. L., Rolfhus, E., Sun, X., ... & Jiang, F. (2023).
1358 Prevalence, co-occurring difficulties, and risk factors of developmental language
1359 disorder: first evidence for Mandarin-speaking children in a population-based study.
1360 *The Lancet Regional Health Western Pacific*, 34, 1-11.
1361 <http://doi.org/10.1016/j.lanwpc.2023.100713>
- 1362 Yang, S., & Berdine, G. (2017). The receiver operating characteristic (ROC) curve. *The*
1363 *Southwest Respiratory and Critical Care Chronicles*, 5(19), 34-36.
- 1364 Ziethe, A., Eysholdt, U., & Doellinger, M. (2013). Sentence repetition and digit span:
1365 Potential markers of bilingual children with suspected SLI? *Logopedics Phoniatics*
1366 *Vocology*, 38(1), 1–10. <https://doi.org/10.3109/14015439.2012.664652>
- 1367
- 1368
- 1369
- 1370
- 1371
- 1372
- 1373
- 1374
- 1375
- 1376
- 1377
- 1378
- 1379
- 1380

Appendix A

1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391

Table A

Quality of Included Studies Assessed using the Standard Quality Assessment Criteria for Quantitative Studies (Kmet et al., 2004)

Criteria	Yes (2)	Partial (1)	No (0)	NA
1 Question / objective sufficiently described?	54	12	0	0
2 Study design evident and appropriate?	65	1	0	0
3 Method of subject/comparison group selection or source of information/input variables described and appropriate?	38	27	1	0
4 Subject (and comparison group, if applicable) characteristics sufficiently described?	47	17	2	0
8 Outcome and (if applicable) exposure measure(s) well defined and robust to measurement / misclassification bias? Means of assessment reported?	42	17	7	0
9 Sample size appropriate?	21	42	3	0
10 Analytic methods described/justified and appropriate?	48	16	2	0
11 Some estimate of variance is reported for the main results?	22	37	7	0
12 Controlled for confounding?	39	23	4	0
13 Results reported in sufficient detail?	38	26	2	0
14 Conclusions supported by the results?	36	5	0	25

Note. Those marked with NA for criteria 14 had made no mention of sentence repetition performance in their conclusions/discussions. Three of these criteria (points 5, 6, and 7) were omitted as they were not applicable to the studies analysed here (they relate instead to interventional designs).

1392 **Appendix B**

1393

1394 **Table B**

1395 *Table showing likelihood ratio tests comparing model fit*

Model	AIC	<i>pval</i>
Effect sizes nested by sample and study	245.08	NA
Effect sizes nested by sample	248.17	0.024
Effect sizes nested by study ^a	243.08	1.000
No added nesting	282.06	<.001

1396

1397 *Note.* The table shows the results of likelihood ratio tests used to compare different
 1398 multilevel meta-analysis models. The model chosen for the meta-analysis was based upon
 1399 the Akaike information criterion (AIC) and resulting statistical significance. The three-level
 1400 model where effect sizes are nested within studies was deemed most appropriate as its AIC
 1401 value was the lowest, and it did not differ significantly from the full four-level model.
 1402 Therefore, it provided the least complex way (in comparison to the four-level model) of
 1403 representing the variability in our data.

1404 ^a The multilevel meta-analysis model where effect sizes are nested by study is the model
 1405 chosen for the final analysis.

1406

1407

1408

1409

1410

1411

1412

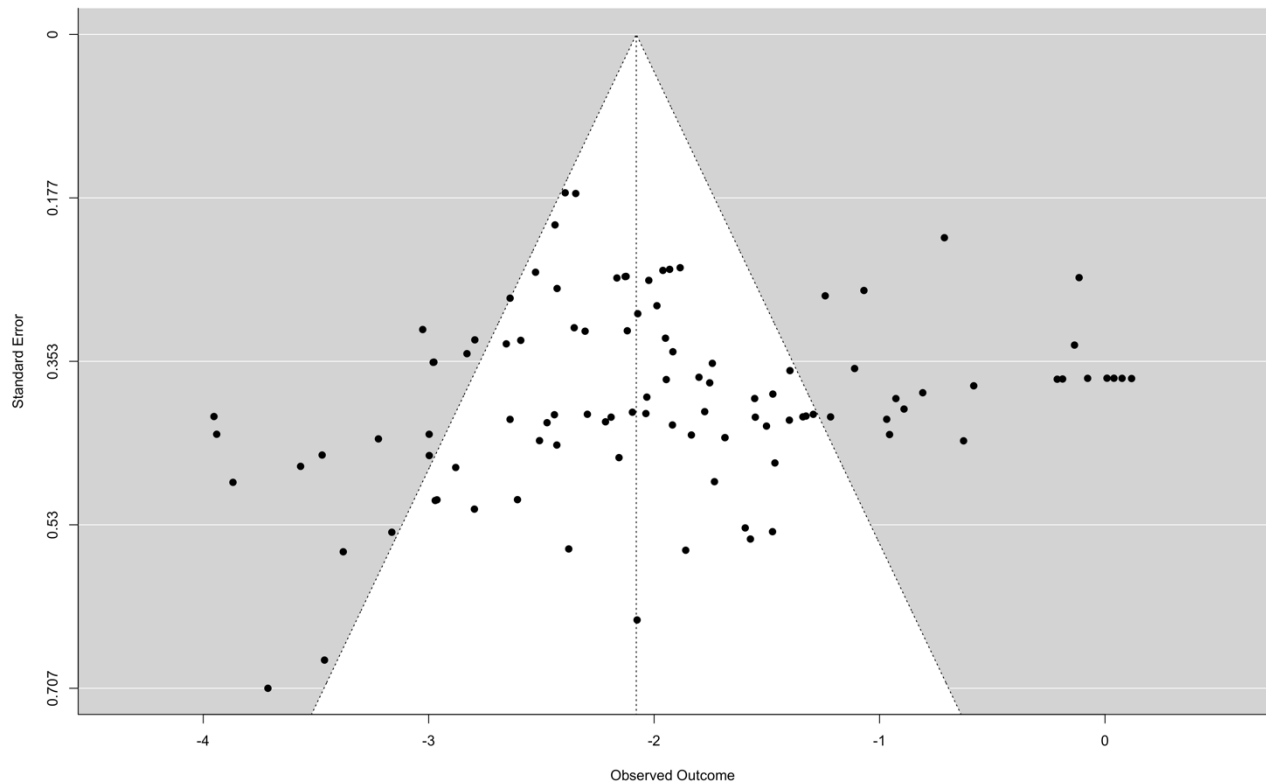
1413

1414

Appendix C

1415 Assessment of Publication Bias

1416

1417 **Figure C-1**1418 *Funnel plot of effects*

1419

1420 *Note.* As can be seen, there was some evidence of asymmetry. To

1421 detect potential outliers and data points contributing most to this asymmetry, Cook's

1422 distances were calculated for each data point. Studies with the highest Cook's distance were

1423 removed until the asymmetry was no longer evident. Through this analysis, four effect sizes

1424 were removed. The removal of these studies resulted in the funnel plot and Egger's

1425 regression test shown in Figure C2 and Table C2.

1426

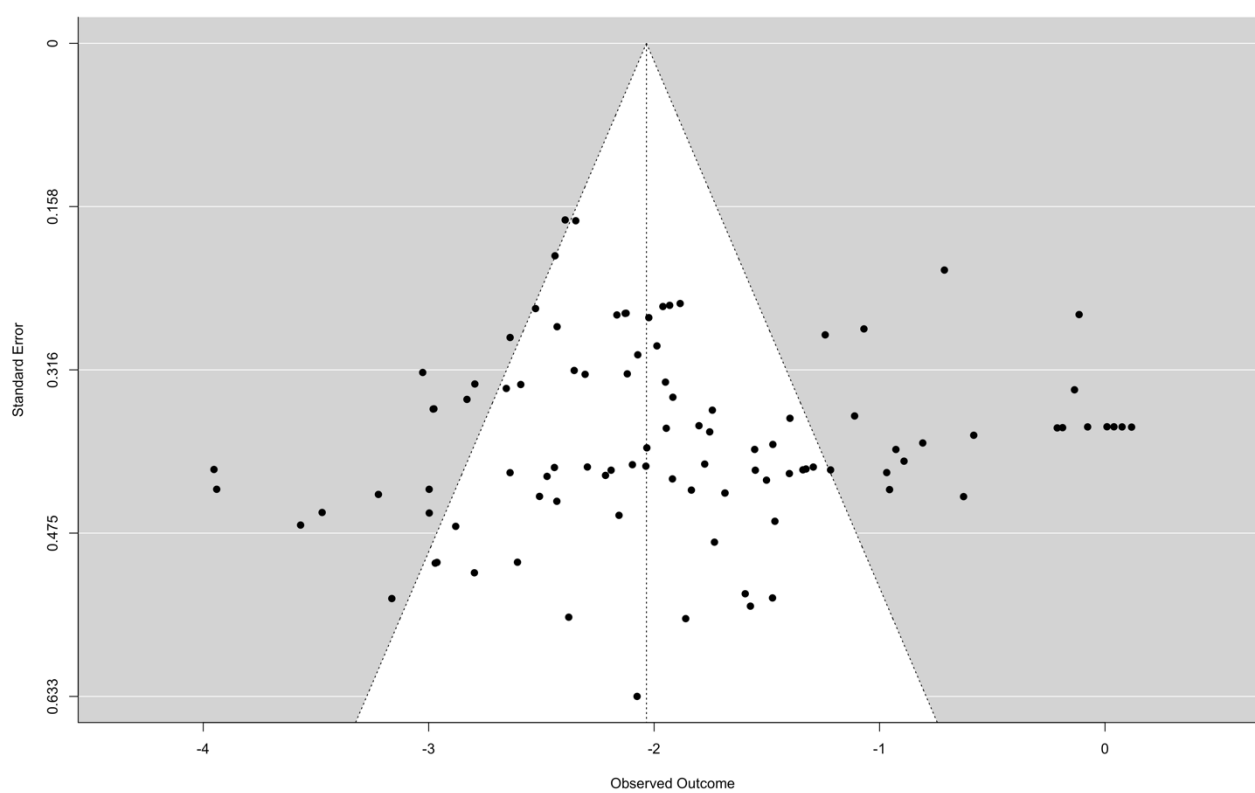
1427

1428

1429 **Table C-1**1430 *Results of Egger's regression test*

z	p
0.67	0.00122

1431

1432 **Figure C-2**1433 *Funnel plot of effects, after the removal of outliers*

1434

1435 **Table C-2**1436 *Results of Egger's regression test, after the removal of outliers*

1437

z	p
0.45	0.0532

1438

1439

1440 **Supplemental Material S1**

1441 This material shows the results of an exploratory analysis. As a result of a
1442 significant subgroup effect being found for type of matching of TD children (age- or
1443 language-matched) as part of RQ8g, this analysis involved the same analyses conducted for
1444 RQ8a-f being ran separately for effect sizes which just concerned age-matched groups. This
1445 was to assess whether any of the previous analyses were influenced by matching as a
1446 confounding factor.

1447 **RQ8 How does variability in study design and SR administration influence the effect**
1448 **size across the studies? More specifically does effect size vary as a function of the**
1449 **following factors:**

1450 Multiple subgroup analyses were conducted for age-matched DLD and TD groups
1451 only. Forest plots showing the overall results of these subgroup analyses are shown in
1452 Figures S3-1 and S3-2.

1453 **RQ8a. Task choice (standardised/norm references or unstandardised)** – The
1454 test for subgroup difference showed there was no significant subgroup effect (p
1455 $=.87$).

1456 **RQ8b. Stimuli presentation (pre-recorded or produced live)** – The test for
1457 subgroup difference showed no significant subgroup effect ($p =.32$).

1458 **RQ8c. Time of scoring (live or offline)** – The test for subgroup difference showed
1459 no significant subgroup effect ($p =.22$).

1460 **RQ8d. Type of scoring (sentence binary, sub-sentence binary, target binary or**
1461 **error scoring)** – There were two separate subgroup analyses run here. The first
1462 compared each of the four categorised types of scoring (sentence binary, sub-
1463 sentence binary, target scoring, and error scoring). This test for subgroup difference
1464 did not reveal a significant subgroup effect ($p =.42$). The second subgroup analysis
1465 compared sentence binary scoring to the three other types of scoring combined. This
1466 analysis also did not show a significant subgroup effect ($p =.37$).

1467 **RQ8e. Language of the task** – The test for subgroup difference showed no
1468 significant subgroup effect ($p = .14$). This is in comparison to the result of the main
1469 analysis which did find a significant subgroup effect of language. This may suggest
1470 that the significant effect found in the main analysis was confounded by the type of
1471 matching used in the studies.

1472 **RQ8f. DLD sample recruited (clinical or population)** – The test for subgroup
1473 difference showed there was no significant subgroup effect ($p = .31$).

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

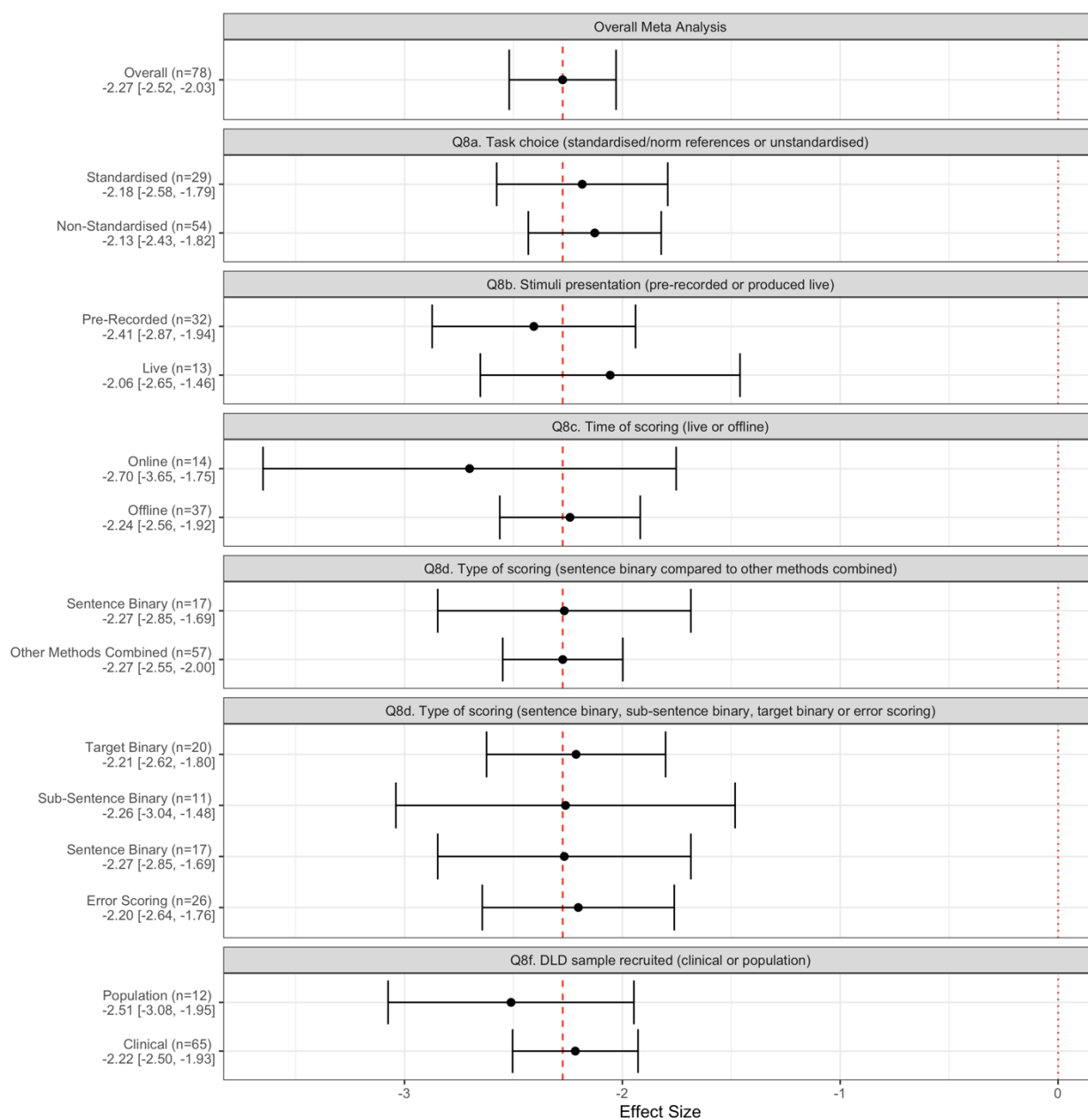
1490

1491

1492

1493

1494

1495 **Figure S3-1**1496 *Summary forest plot showing the pooled effect size for each subgroup analysis*

1497

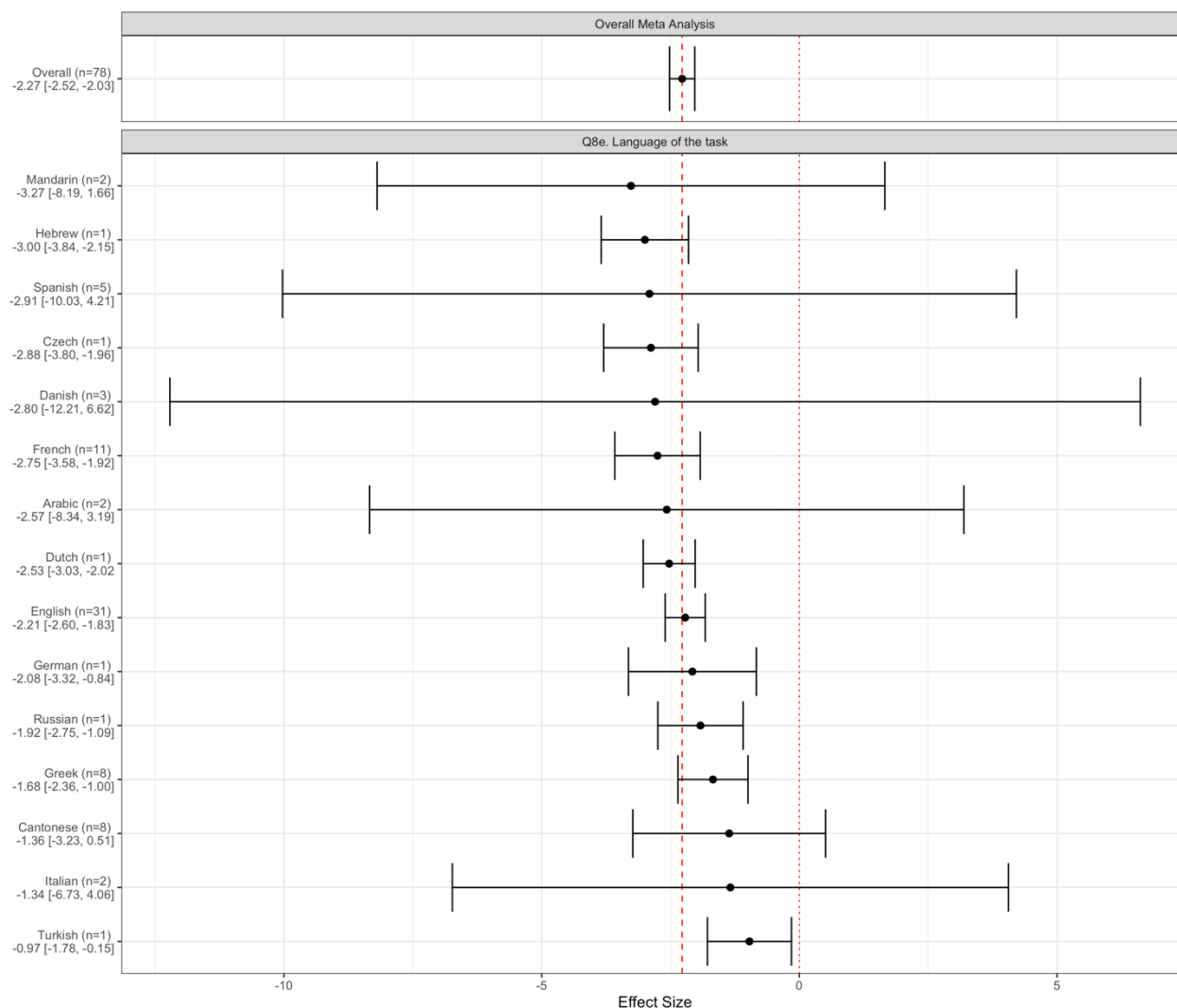
1498 *Note.* Points represent a calculation of the pooled estimate of effect size (Hedges' g) from a
 1499 multilevel meta-analysis for each defined subgroup (surrounded by 95% confidence
 1500 intervals). ' n ' refers to the number of datapoints included in each analysis (not number of
 1501 studies).

1502

1503

1504 **Figure S3-2**

1505 *Summary forest plot showing the pooled effect size for each subgroup as part of the*
 1506 *subgroup analysis ran for language of task*



1507

1508 *Note.* Datapoints are presented in order of effect

1509 Points represent a calculation of the pooled estimate of effect size (Hedges' *g*) from a

1510 multilevel meta-analysis for each defined subgroup (surrounded by 95% confidence

1511 intervals). '*n*' refers to the number of datapoints included in each analysis (not number of

1512 studies).

1513

1514

1515