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

Leah Ward[1], Kamila Polišenská[1, 2] and Colin Bannard[3]

[1] Division of Psychology, Communication and Human Neuroscience,

The University of Manchester

[2] Department of Language and Communication Science,

City, University of London

[3] Department of Linguistics and English Language,

The University of Manchester

Author Note

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

Kamila Polišenská https://orcid.org/0000-0001-7405-6689

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

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

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

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

Abstract

Purpose: This 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.

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.

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

Measures of diagnostic accuracy are a useful tool in assessing the diagnostic utility of a task. The most commonly used measures of diagnostic accuracy are sensitivity and specificity (for an introduction see Chu, 1999). Sensitivity is a measure of a task's ability to identify disorder, here the proportion of children with DLD correctly identified by the SR task. Specificity is a measure of a task's ability to reject the presence of disorder, being the proportion of those without DLD correctly identified by the task. Determining the most effective cut-off point (see Yang & Berdine, 2017) for a SR task helps to reduce the two types of classification error caused by low sensitivity (under-diagnosis from false negatives) or low specificity (over-diagnosis from false positives), both of which are harmful in clinical contexts. Other measures of diagnostic accuracy involve likelihood ratios (LR) which directly link to the pre-test and post-test probability of the disorder (Deeks & Altman, 2004). The LR for positive test results (LR+) refers to how likely the positive result (identification of DLD) is to occur in those with the disorder (children who have DLD) compared to without (TD children). In contrast, LR for negative test results (LR-) refers to how likely the negative result is to occur (identification of not having DLD) in those with the disorder (children who have DLD) compared to without (TD children).

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

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

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

Objectives

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

Systematic review

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

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

3) What methods are used to score children's productions on the task?
4) What levels of reliability does the task achieve?
5) What languages are the tasks administered in?
6) How has DLD and TD been defined in the sample?

Meta-analysis

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

7) Do SR tasks reveal significant performance differences between groups of TD children and children with DLD? What is the main effect size of the studies?
8) How does variability in study design and SR administration influence the effect size across the studies? More specifically does effect size vary as a function of the following factors:
    a. Task choice (standardised/norm references or unstandardised)
    b. Stimuli presentation (pre-recorded or produced live)
    c. Time of scoring (live or offline)
    d. Type of scoring (sentence binary, sub-sentence binary, target binary or error scoring)
    e. Language of the task
    f. DLD sample recruited (clinical or population)
    g. Matching of TD children (age- or language-matched)

Methods

This review was conducted and reported in line with the Preferred Reporting Items for Systematic Review (PRISMA) guidelines (Page et al., 2021). The review is registered with the PROSPERO international prospective register of systematic reviews and meta-analyses, accessible at https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022303100.

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

Eligibility Criteria

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

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

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

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

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

Information Sources

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

Search Strategy

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

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

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

Search 3 – 1 AND 2

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

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

Selection Process

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

Data Collection Process

Following the selection process and in answering the research questions, information was obtained from the included studies using a data extraction form created by the authors (see https://osf.io/usw2k/). Data was collected into this spreadsheet by the first author, with the second author independently extracting data from 10% of included articles using the same form. The third author was responsible for assessing agreement, individually comparing each input across forms, and judging whether the information recorded reflected the same level of information. The outcome of this agreement is included in the results.

Data Items

The information that was obtained from the studies included:

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

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

Type of Scoring

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

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

Full details of these categories are provided in the data extraction guide [https://osf.io/usw2k/].

Risk of Bias Assessment

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

Effect Measures

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

Effect Size Calculations

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

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

Synthesis Methods

Systematic Review

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

Meta-analysis

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

Heterogeneity was assessed with Cochran´s Q statistic (Cochran, 1954) and the I^2 Index (Higgins & Thompson, 2002). Because heterogeneity was expected between studies, random-effects modelling was used.

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

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

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

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

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

Q8d. Type of scoring – Studies were grouped according to how they scored and measured accuracy in SR productions. This involved the four grouping labels described previously: sentence binary, sub-sentence binary, target binary, and error scoring. Two separate subgroup analyses were run using this information. The first comparing each of these four categories. The second compares just sentence binary scoring against each other type of scoring as a single subgroup (a collapsed group containing parts binary scoring, target binary scoring and error scoring).

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

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

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

Reporting Bias Assessment

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

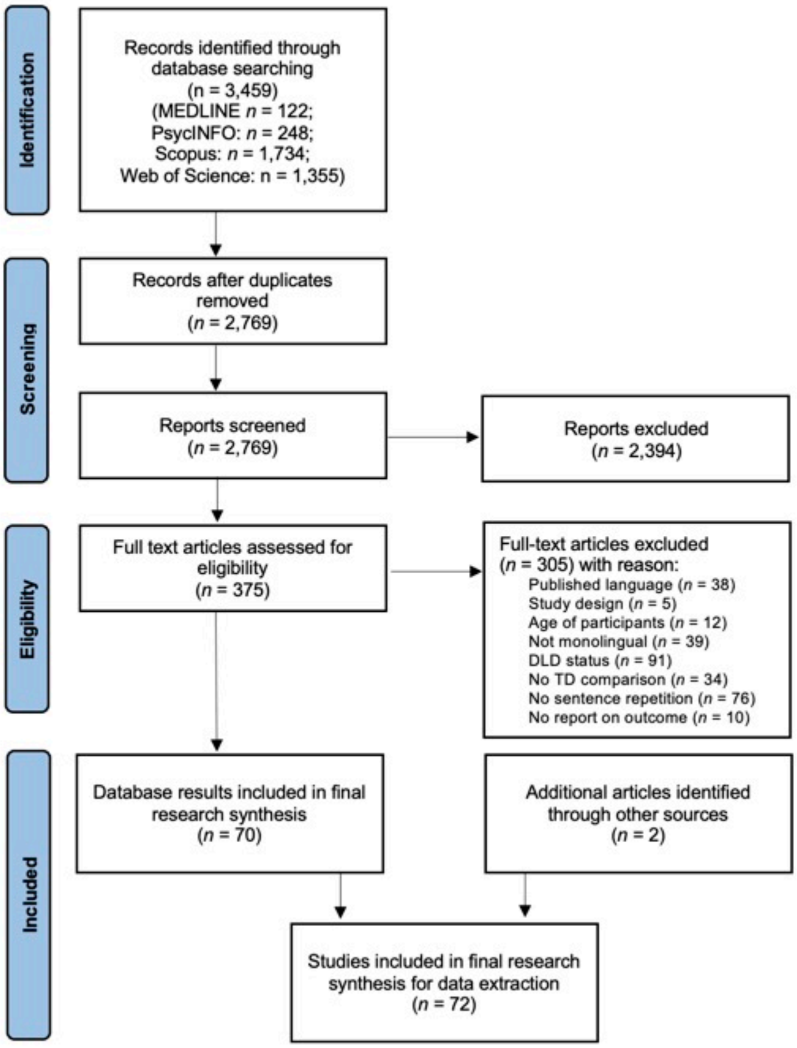
Results

Study Selection

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

Figure 1

PRISMA Flow Diagram Outlining the Selection Process



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

In both the abstract screening and full text screening stages, 10% of articles (277 and 38 randomly selected articles respectively) were independently screened by the second author. Interrater agreement was 93.86% for the abstract screening stage, with agreement on the relevance (include or exclude) of 260 out of the 277 articles. Of the 17 disagreed upon, a separate screening of their full texts identified that none met the eligibility criteria for inclusion, if hypothetically, they had all made it to the next stage. Following on from this, in the full text screening phase, interrater agreement was 97.37%, with agreement on 37 of the 38 articles. The second author also independently extracted data from 10% of the articles (seven studies) included in the final review. The overall percentage of interrater agreement (as determined by the third author) for this was 80.26%.

Study Characteristics

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

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

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

Risk of Bias Assessment

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

Table 1

Summary of Study Characteristics

Study	Language	DLD		TD		Matching	Task
		n	Age	n	Age		
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 subtest, 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 subtest
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)
Petruccelli 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
Tsimpli et al. (2016)*	Greek	21	5;5–11;6 (9;3)	21	5;2–11;5 (9;0)	Age	DVIQ (Stavrakaki & Tsimpli, 2000), Sentence Repetition Subtest
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

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

Systematic Review

RQ1 What is the diagnostic accuracy of SR in distinguishing between children with DLD and typically developing children (TD)?

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

Table 2

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

Note. Only studies that provided values based on a calculated cutoff point of performance are included. Sen = sensitivity; Spe = specificity; LR+ = positive likelihood ratio, calculated using the following formula: sensitivity/(1 – specificity); LR− = negative likelihood ratio, calculated using the following formula: (1 – sensitivity)/specificity. '—' indicates that LR+ could not be calculated.

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

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

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

RQ2 What kinds of SR tasks have been used?

Looking back to Table 1, a variety of different SR tasks were utilised by the studies. Table 3 provides a summary of the specific tasks used. As can be seen, 18 tasks were original and created by their authors. The remaining 41 tasks seen involved using or adapting a pre-existing task or principles. Of these, 22 tasks were standardised/norm referenced tasks.

Table 3

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

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

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

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

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

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

In terms of specific methods of administration, seven were presented using PowerPoint slides (Armon-Lotem & Meir, 2016; De Almeida et al., 2021; Fleckstein et al., 2018; Oetting et al., 2016; Pham & Ebert, 2020; Theodorou et al., 2017; Wang et al., 2022), six were presented in a task involving a puppet producing the sentences to repeat (Christensen & Hansson, 2012; Frizelle & Fletcher, 2014a, 2014b; Riches, 2012, 2015, 2017), four were presented with accompanying figures or pictures (Caselli et al., 2008; Garraffa et al., 2015; Orsolini et al., 2001; Stokes & Fletcher, 2003), and three involved sentences embedded within stories (Leroy et al., 2013; Thordardottir & Brandeker, 2013; Thordardottir et al., 2011).

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

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

Table 4

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

Note. Full details of these categories are provided in the data extraction guide (https://osf.io/usw2k/).

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

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

RQ4 What levels of reliability does the task achieve?

Only 51.52% of studies reported that the transcriptions or scorings of children's productions were verified, and the reliability assessed. The depth of this ranged from assessing a subsample of 5% of the sample to looking to the whole sample.

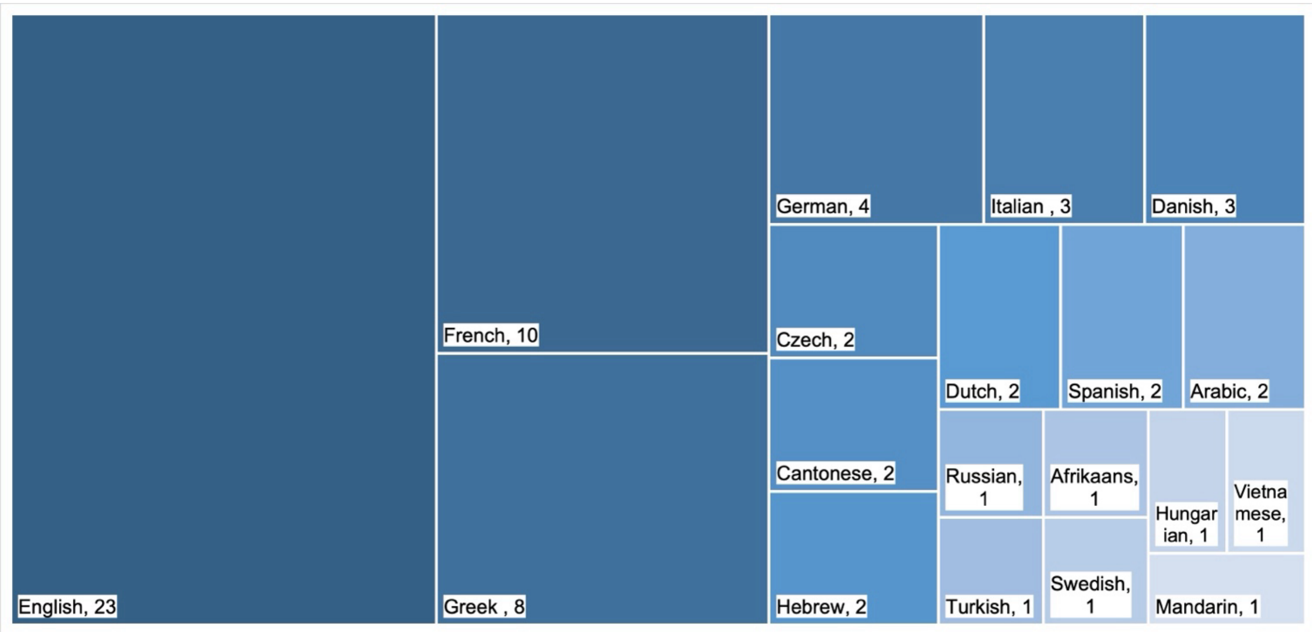
For studies which provided specific measures of reliability, levels were generally high. Inter-transcriber agreement was reported in eight studies (12.12%) and ranged from 92.5% to 99.6%. Inter-scorer agreement was reported in 27 studies (40.91%) and ranged from 86.5% to 100%.

RQ5 What languages are the tasks administered in?

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

Figure 2

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



Note. Areas of the tree map are in proportion to the frequencies of studies seen in the systematic review.

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

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

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

Meta-Analysis

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

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

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

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

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

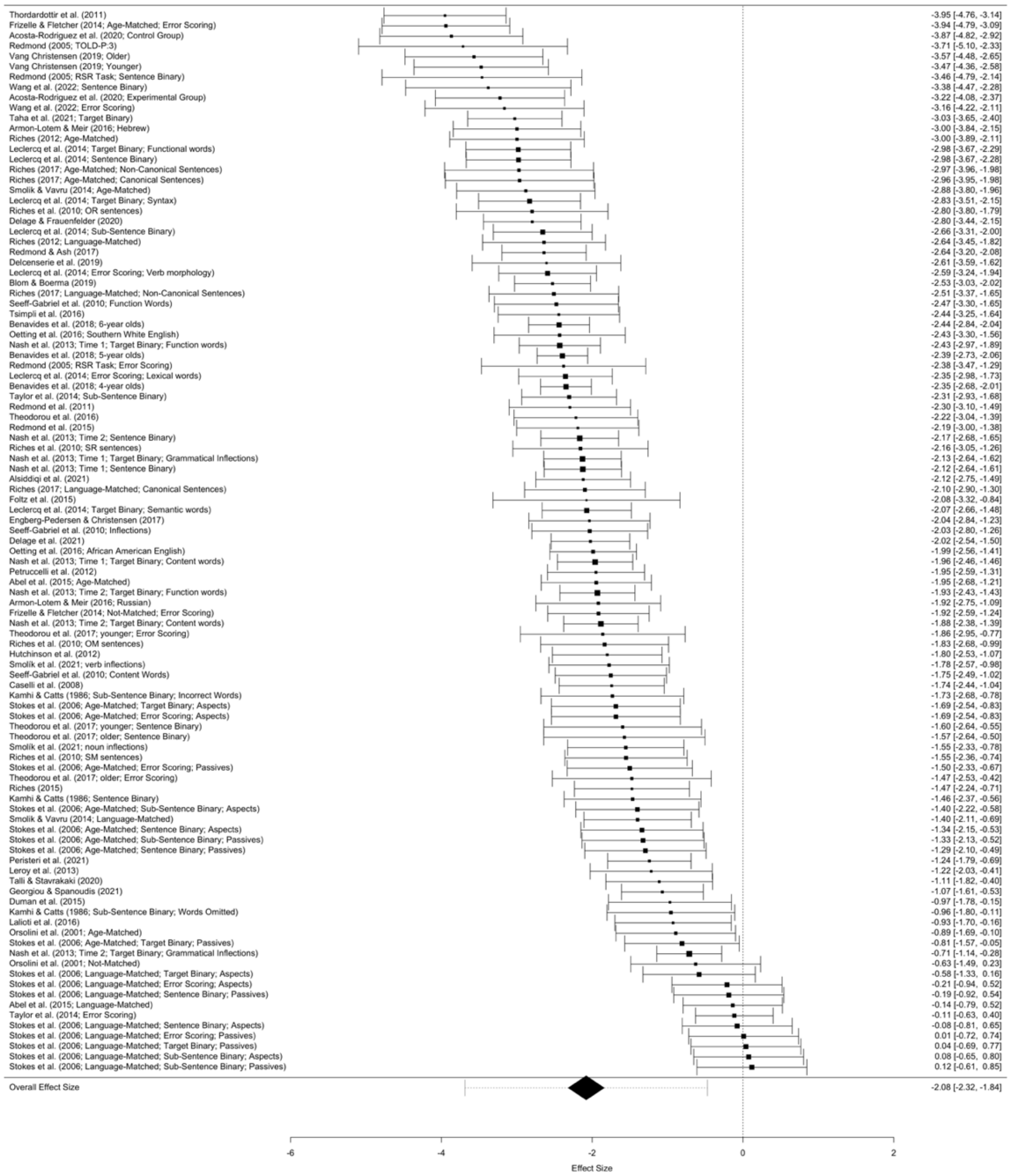
In exploring the sources of heterogeneity and what specific factors influence the power of SR tasks in discriminating groups of TD children and children with DLD, multiple subgroup analyses were conducted. Forest plots showing the overall results of these subgroup analyses are shown in Figures 4 and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

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

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

Figure 3

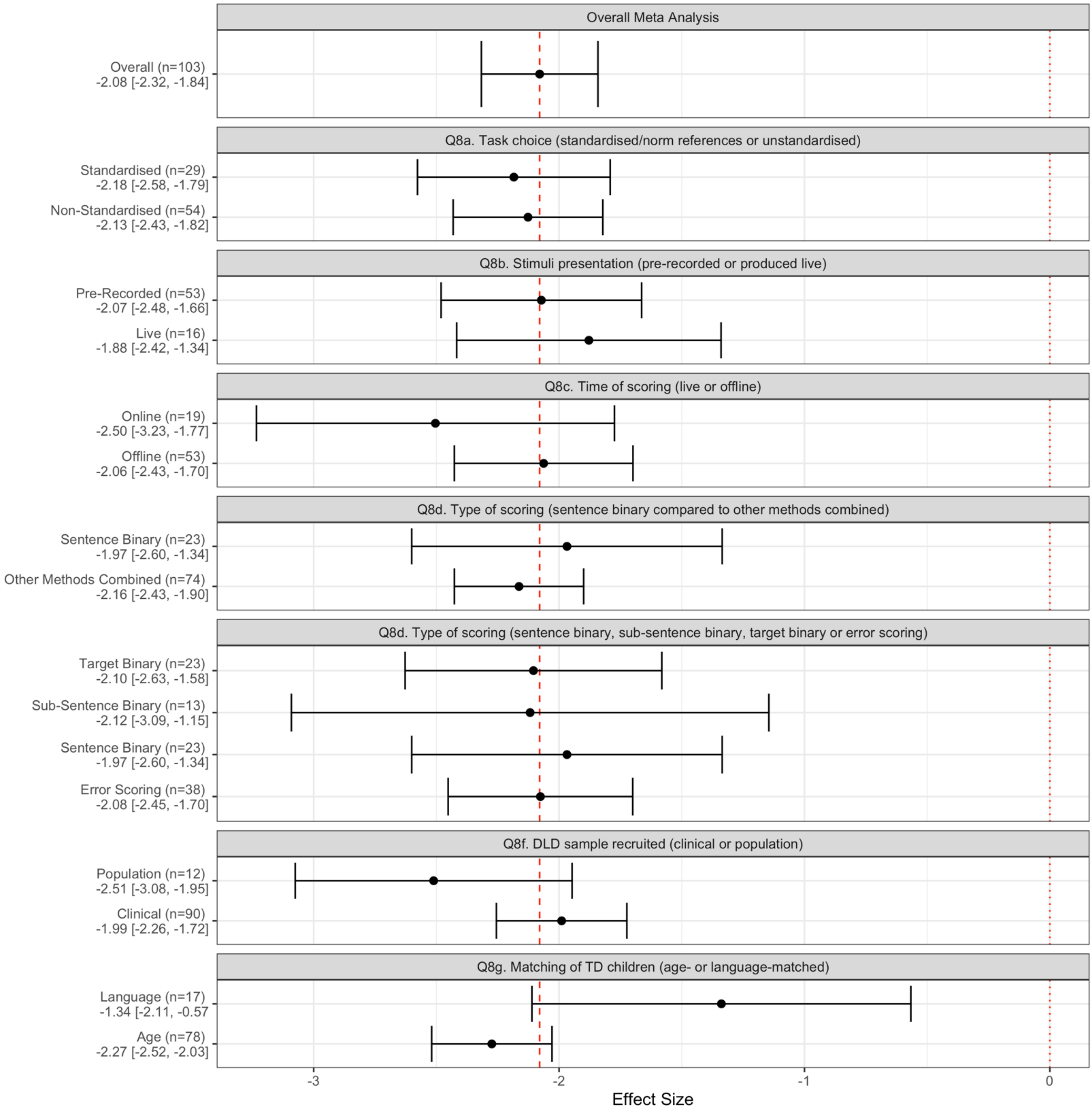
Forest plot showing the effect size of each included study and calculated pooled effect size



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

Figure 4

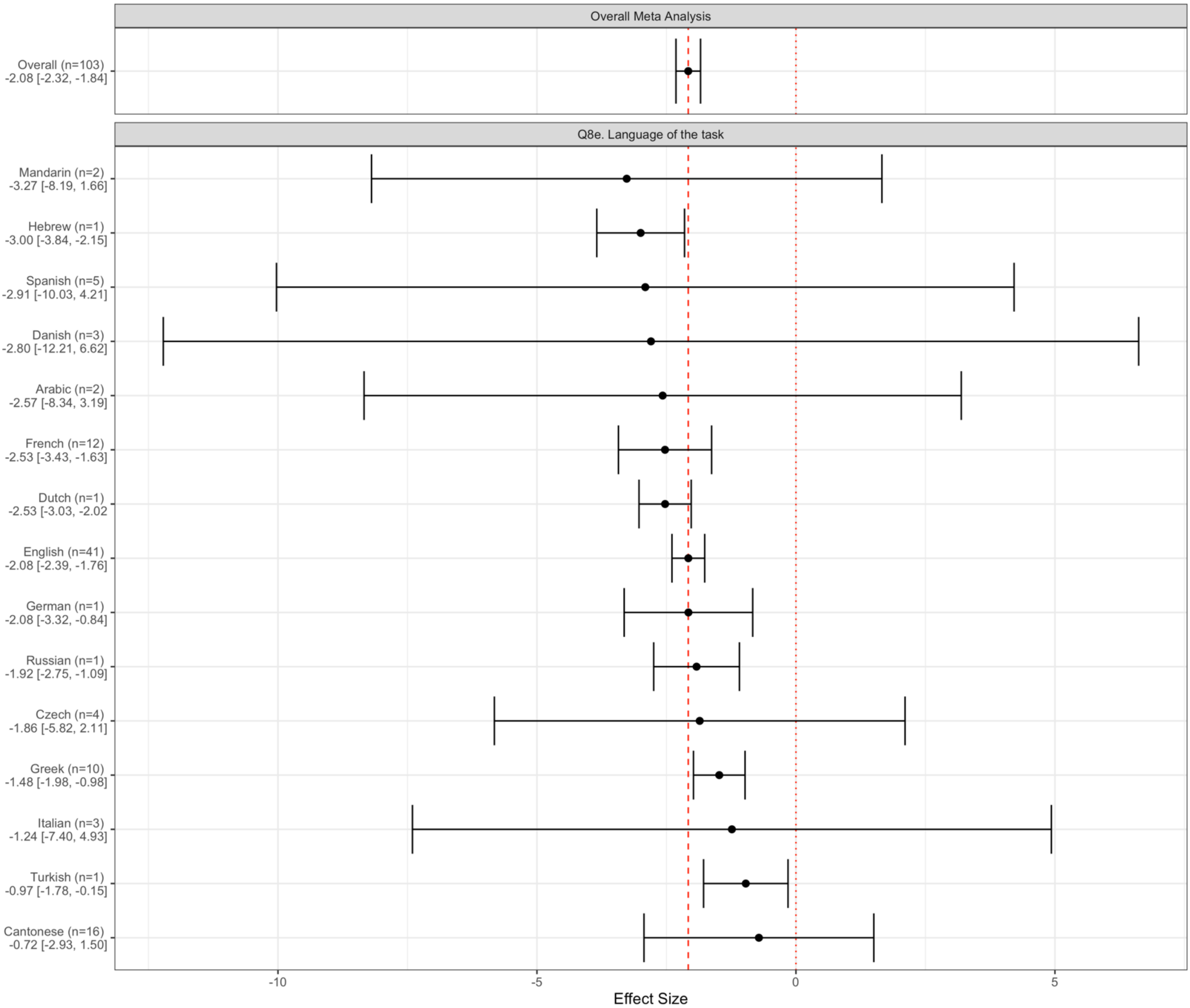
Summary forest plot showing the pooled effect size for each subgroup analysis



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

Figure 5

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



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

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

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

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

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

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

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

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

Reporting Bias Assessment

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

Table 5

Summary of study characteristics for studies included in the meta-analysis which involved 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

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

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

Discussion

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

As McGregor (2020) points out, to be of clinical use a tool must be able to detect cases of disorder (sensitivity) and its absence (specificity). Diagnostic accuracy metrics were reported in 18 of the studies. Of the values provided, the majority (75.86%) indicated acceptable levels of sensitivity and specificity (above 80% for both values) and can be viewed as having fair (to good) diagnostic accuracy when following the recommendations set out by Plante and Vance (1994). On the other hand, this meant that 24.14% of the values reported in the studies included show poor sensitivity and specificity (under 80% on at least one value). If applied in a clinical context this could cause harm by either misdiagnosing a child with DLD (false positive) or missing a diagnosis (false negative). The authors of a particularly low specificity (57%) study (Pham & Ebert, 2020) suggest that their SR task (with binary scoring) could present a quick and effective screening tool to identify those in need of further testing, rather than acting as a diagnostic test.

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

Our multilevel meta-analysis found a very large effect size of g = -2.08 (95% CI [-2.32, -1.84]) for the difference in performance between groups of children with DLD and groups of TD children. This is a larger effect size than reported for meta-analyses looking at other methods of identifying children with DLD when compared with subgroups of TD children. A meta-analysis by Winters et al. (2022) looked at narrative performance (g = -0.82 (95% CI [-0.99, -0.66])), and there have been two meta-analyses to date looking specifically at nonword repetition (Schwob et al., 2021 and Estes et al., 2007; g = 1.57 (95% CI [1.37, 1.72]) and d = 1.27 (95% CI [1.15, 1.39]) respectively). Note however, that Winters et al. (2022), and Schwob et al. (2021) did not exclude studies and results from bilingual populations, whereas our review and meta-analysis did. From the available evidence, SR appears to be the best available means of discriminating children with DLD from typically developing children. SR provides a test of lexical phonology and morphosyntax (Polišenská et al., 2015), with each repetition requiring short-term memory and prior language knowledge to process, store and regenerate the sentences. This overall reflection of language ability is likely what sets SR apart from alternative methods, as it targets areas in which those with DLD are impaired.

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

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

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

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

A significant subgroup effect was found for how DLD and TD groups were matched – the size of the difference between groups was significantly larger when TD children were matched to those with DLD by age, by comparison to when TD children were matched to those with DLD by language ability. However, while the effect was smaller, the overall effect size across studies which compared DLD performance to language-matched TD groups remained large (g = -1.31 (95% CI [-2.0360, -0.5918], with children with DLD showing less accurate SR performance in comparison to younger, language-matched children. In a clinical context children would be compared to those of a similar age, with standardised SR tasks such as the CELF (Wiig et al., 2013) having norm-referenced comparisons for age. However, this remains an important finding because a task distinguishing age-matched DLD and TD children may just target general language properties that a child with other impairments including language delay would perform poorly on when compared to children of the same age (Van der Lely & Howard, 1993). In distinguishing between those with DLD and language-matched, younger, TD children, a SR task is likely to be targeting the specific structures which cause low performance in DLD specifically, leading to more crude individual differences. Riches (2012) concluded that their finding that children with DLD perform significantly worse than language-matched controls is indeed strong validating evidence of the use of SR as a clinical marker.

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

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

The language of the SR task also was found to significantly impact the size of the effect in a subgroup analysis. However, when this analysis was rerun with just studies who used age-matched TD groups, this effect disappears, suggesting that there was a confounding effect of how TD children were matched. It can therefore be tentatively concluded that SR tasks reliably result in a difference in performance between DLD and TD groups across different languages (in monolingual children), even those with a vastly different morphosyntactic structure to English, such as Arabic (Alsiddiqi et al., 2021; Taha et al., 2021). This may be in part due to the standardising influence of COST Action IS0804 "Language Impairment in a Multilingual Setting: Linguistic Pattern and the Road to Assessment", and the LITMUS-SRep task (Marinis & Armon-Lotem, 2015) that was developed as part of the project. LITMUS was a collaborative effort to develop methods of language assessment (including a SR task) which can identify DLD within a bilingual setting. The most frequent task seen was those developed following LITMUS-SRep principles (and used here in a monolingual setting) which propose that the sentences used should differ in the grammatical structures known to be difficult to those with DLD across languages (e.g., relative clauses) as well as the language specific to the task. Global collaborations such as this may be important for the development of SR tasks in the future and to promote a more standardised use across clinical and research contexts.

Limitations

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

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

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

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

Conclusions and Clinical Implications

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

Acknowledgements

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

Data Availability Statement

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

References

Abel, A. D., Rice, M. L., & Bontempo, D. E. (2015). Effects of verb familiarity on finiteness marking in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 58*(2), 360–372. https://doi.org/10.1044/2015_jslhr-l-14-0003

Acosta-Rodríguez, V. M., Ramírez-Santana, G. M., Hernández Expósito, S., & Axpe Caballero, Á. (2020). Intervention in syntactic skills in pupils with developmental language disorder. *Psicothema, 32*(4), 541–548. https://doi.org/10.7334/psicothema2020.160

Alsiddiqi, Z. A., Stojanovik, V., & Pagnamenta, E. (2021). Emergent literacy skills of Saudi Arabic speaking children with and without developmental language disorder. *Clinical Linguistics & Phonetics, 36*(4–5), 301–318. https://doi.org/10.1080/02699206.2021.1955299

Archibald, L. M. D., & Joanisse, M. F. (2009). On the Sensitivity and Specificity of Nonword Repetition and Sentence Recall to Language and Memory Impairments in Children. *Journal of Speech, Language, and Hearing Research, 52*(4), 899-914. https://doi.org/10.1044/1092-4388

Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.). (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment* (Vol. 13). Multilingual matters.

Armon-Lotem, S., & Meir, N. (2016). Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: evidence from Russian and Hebrew. *International Journal of Language & Communication Disorders, 51*(6), 715-731. https://doi.org/10.1111/1460-6984.12242

Benavides, A. A., Kapantzoglou, M., & Murata, C. (2018). Two grammatical tasks for screening language abilities in Spanish-speaking children. *American Journal of Speech Language Pathology, 27*(2), 690–705. https://doi.org/10.1044/2017_AJSLP17-0052

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

Blom, E., & Boerma, T. (2019). Reciprocal relationships between lexical and syntactic skills of children with Developmental Language Disorder and the role of executive functions. *Autism & Developmental Language Impairments*, 4. https://doi.org/10.1177/2396941519863984

Calder, S. D., Brennan-Jones, C. G., Robinson, M., Whitehouse, A., & Hill, E. (2022). The prevalence of and potential risk factors for Developmental Language Disorder at 10 years in the Raine Study. *Journal of Paediatrics and Child Health, 58*(11), 2044-2050. https://doi.org/10.1111/jpc.16149

Caselli, M. C., Monaco, L., Trasciani, M., & Vicari, S. (2008). Language in Italian Children with Down Syndrome and with Specific Language Impairment. *Neuropsychology, 22(*1), 27-35. https://doi.org/10.1037/0894-4105.22.1.27

Chevrie-Muller, C., Simon, A. M., Fournier, S., & Brochet, M. O. (2010). *Batterie langage oral-langage écrit, mémoire-attention: L2ma* (2nd ed.). ECPA.

Christensen, R. V., & Hansson, K. (2012). The Use and Productivity of Past Tense Morphology in Specific Language Impairment: An Examination of Danish. *Journal of*

Speech Language and Hearing Research, 55(6), 1671-1689. https://doi.org/10.1044/1092-4388(2012/10-0350)

Christensen, R. V., Jensen, S. T. & Nielsen, I. I. (2012). *Sætningsgentagelsestesten [The Sentence Repetition Test]*. Institut for Nordiske Studier ogSprogvidenska [Department of Nordic Studies and Linguistics], KøbenhavnsUniversitet [University of Copenhagen].

Chu, K. (1999). An introduction to sensitivity, specificity, predictive values and likelihood ratios. *Emergency Medicine, 11*(3), 175-181. https://doi.org/10.1046/j.14422026.1999.00041.x

Coady, J. A., Evans, J. L., & Kluender, K. R. (2010). The Role of Phonotactic Frequency in Sentence Repetition by Children With Specific Language Impairment. *Journal of Speech Language and Hearing Research, 53(*5), 1401-1415. https://doi.org/10.1044/1092-4388(2010/07-0264)

Cochran, W. G. (1954). Some methods for strengthening the common χ 2 tests. *Biometrics, 10*(4), 417-451. https://doi.org/10.2307/3001616

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry, 42*(6), 741-748. https://doi.org/10.1111/1469-7610.00770

Cook, R. D., & Weisberg, S. (1982). Criticism and Influence Analysis in Regression. *Sociological Methodology, 13*, 313. https://doi.org/10.2307/270724

de Almeida, L. de, Ferré, S., Morin, E., Prévost, P., Santos, C. dos, Tuller, L., Zebib, R., & Barthez, M. A. (2017). Identification of bilingual children with Specific Language Impairment in France. *Language Impairment in Bilingual Children, 7*(3-4), 331-358. https://doi.org/10.1075/lab.15019.alm

Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *Bmj, 329*(7458), 168-169. https://doi.org/10.1136/bmj.329.7458.168

Delage, H., & Frauenfelder, U. (2012). Développement de la mémoire de travail et traitement des phrases complexes: Quelle relation? *SHS Web of Conferences, 1*, 1555-1573. https://doi.org/10.1051/shsconf/20120100141

Delage, H., & Frauenfelder, U. H. (2020). Relationship between working memory and complex syntax in children with developmental language disorder. *Journal of Child Language, 47*(3), 600–632. https://doi.org/10.1017/s0305000919000722

Delage, H., Stanford, E., & Durrleman, S. (2021). Working memory training enhances complex syntax in children with Developmental Language Disorder. *Applied Psycholinguistics, 42(*5), 1341-1375, http://doi.org/ 10.1017/S0142716421000369

Delcenserie, A., Genesee, F., Trudeau, N., & Champoux, F. (2019). A multi-group approach to examining language development in at-risk learners. *Journal of Child Language, 46*(1), 51–79. https://doi.org/10.1017/S030500091800034X

Devescovi, A., & Caselli, M. C. (2001). Una prova di ripetizione di frasi per la valutazione del primo sviluppo grammaticale. *Psicologia clinica dello sviluppo*, (3), 341-364.

Devescovi, A., Caselli, M. C. and Ossella, T. (1992). Rilevazione delle prime fasi dello sviluppo morfosintattico attraverso una prova di ripetizione. *Rassegna di Psicologia, 2*, 25–42.

Dosi, I., & Koutsipetsidou, E.-C. (2019). Measuring linguistic and cognitive abilities by means of a sentence repetition task in children with developmental dyslexia and developmental language disorder. *European Journal of Research in Social Sciences, 7*(4), 10–19.

Duman, T. Y., Blom, E., & Topbas, S. (2015). At the Intersection of Cognition and Grammar: Deficits Comprehending Counterfactuals in Turkish Children With Specific Language Impairment. *Journal of Speech Language and Hearing Research, 58*(2), 410-421. https://doi.org/10.1044/2015_jslhr-l-14-0054

Eadie, P. A., Fey, M. E., Douglas, J. M., & Parsons, C. L. (2002). Profiles of grammatical morphology and sentence imitation in children with specific language impairment and

Down syndrome. *Journal of Speech, Language, and Hearing Research, 45*(4), 720–732. https://doi.org/10.1044/1092-4388(2002/058)

Ebert, K. D., Rak, D., Slawny, C. M., & Fogg, L. (2019). Attention in bilingual children with developmental language disorder. *Journal of Speech, Language, and Hearing Research, 62*(4), 979-992. https://doi.org/10.1044/2018_JSLHR-L-18-0221

Engberg-Pedersen, E., & Christensen, R. V. (2017). Mental states and activities in Danish narratives: Children with autism and children with language impairment. *Journal of Child Language, 44*(5), 1192-1217. https://doi.org/10.1017/S0305000916000507

Estes, K. G., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 50*(1), 177-195. https://doi.org/10.1044/1092-4388(2007/015)

Fleckstein, A., Prevost, P., Tuller, L., Sizaret, E., & Zebib, R. (2018). How to identify SLI in bilingual children: A study on sentence repetition in French. *Language Acquisition, 25(*1), 85-101. https://doi.org/10.1080/10489223.2016.1192635

Foltz, A., Thiele, K., Kahsnitz, D., & Stenneken, P. (2015). Children's syntactic-priming magnitude: Lexical factors and participant characteristics. *Journal of Child Language, 42(*4), 932-945. https://doi.org/https://dx.doi.org/10.1017/S0305000914000488

Frizelle, P., & Fletcher, P. (2014a). Relative clause constructions in children with specific language impairment. *International Journal of Language & Communication Disorders, 49*(2), 255-264. https://doi.org/10.1111/1460-6984.12070

Frizelle, P., & Fletcher, P. (2014b). Profiling relative clause constructions in children with specific language impairment. *Clinical Linguistics & Phonetics, 28*(6), 437-449. https://doi.org/10.3109/02699206.2014.882991

Frizelle, P., O'Neill, C., & Bishop, D. V. (2017). Assessing understanding of relative clauses: A comparison of multiple-choice comprehension versus sentence repetition*. Journal of Child Language, 44*(6), 1435-1457. https://doi.org/10.1017/S0305000916000635

Gagiano, S., & Southwood, F. (2015). The use of digit and sentence repetition in the identification of language impairment: The case of child speakers of Afrikaans and South African English. *Stellenbosch Papers in Linguistics, 44*, 37–60. https://doi.org/10.5774/44-0-187

Garraffa, M., Coco, M. I., & Branigan, H. P. (2015). Effects of Immediate and Cumulative Syntactic Experience in Language Impairment: Evidence from Priming of Subject Relatives in Children with SLI. *Language Learning and Development, 11*(1), 18-40. https://doi.org/10.1080/15475441.2013.876277

Georgiou, N., & Spanoudis, G. (2021). Developmental language disorder and autism: Commonalities and differences on language. *Brain Sciences, 11*(5), Article 589. https://doi.org/10.3390/brainsci11050589

Grimm, H., & Schöler, H. (1977). *Heidelberger Sprachentwicklungstest (HSET).* Hogrefe.

Grimm, H., & Schöler, H. (1991). *Heidelberger Sprachentwicklungstest (H-S-E-T)*. Hans Huber Verlag.

Håkansson, G., & Hansson, K. (2000). Comprehension and production of relative clauses: A comparison between Swedish impaired and unimpaired children. *Journal of Child Language, 27*(2), 313–333. https://doi.org/10.1017/s0305000900004128

Hamann, C., & Abed Ibrahim, L. (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Frontiers in Communication, 2*, Article 16. https://doi.org/10.3389/fcomm.2017.00016

Hannus, S., Kauppila, T., & Launonen, K. (2009). Increasing prevalence of specific language impairment (SLI) in primary healthcare of a Finnish town, 1989–99. *International journal of language & communication disorders, 44*(1), 79-97. https://doi.org/10.1080/13682820801903310

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128. https://doi.org/10.3102/10769986006002107

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine, 21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Hutchinson, E., Bavin, E., Efron, D., & Sciberras, E. (2012). A comparison of working memory profiles in school-aged children with specific language impairment, attention deficit/hyperactivity disorder, comorbid SLI and ADHD and their typically developing peers. *Child Neuropsychology, 18*(2), 190–207. https://doi.org/10.1080/09297049.2011.601288

Kamhi, A. G., & Catts, H. W. (1986). Toward an understanding of developmental language and reading disorders. *The Journal of speech and hearing disorders, 51*(4), 337-347. https://doi.org/10.1044/jshd.5104.337

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

Kmet, L. M., Lee, R. C., & Cook, L. S. (2004) Standard quality assessment criteria for evaluating primary research papers from a variety of fields. *Alberta Heritage Foundation for Medical Research, 13*, 1–22.

Komeili, M., & Marshall, C. R. (2013). Sentence repetition as a measure of morphosyntax in monolingual and bilingual children. *Clinical Linguistics & Phonetics, 27*(2), 152-162. https://doi.org/10.3109/02699206.2012.751625

Korkman, M., Kirk, U., & Kemp, S. L. (2007). *NEPSY-II: A developmental neuropsychological assessment*. The Psychological Corporation.

Kueser, J. B., & Leonard, L. B. (2020). The Effects of Frequency and Predictability on Repetition in Children With Developmental Language Disorder. *Journal of Speech Language and Hearing Research, 63*(4), 1165-1180. https://doi.org/10.1044/2019_jslhr-19-00155

Lalioti, M., Stavrakaki, S., Manouilidou, C., & Talli, I. (2016). Subject–verb agreement and verbal short-term memory: A perspective from Greek children with specific language impairment. *First Language, 36*(3), 279–294. https://doi.org/10.1177/0142723716648844

Leclercq, A. L., Quemart, P., Magis, D., & Maillart, C. (2014). The sentence repetition task: A powerful diagnostic tool for French children with specific language impairment. *Research in Developmental Disabilities, 35*(12), 3423-3430. https://doi.org/https://dx.doi.org/10.1016/j.ridd.2014.08.026

Leroy, S., Parisse, C., & Maillart, C. (2013). The influence of the frequency of functional markers on repetitive imitation of syntactic constructions in children with specific language impairment, from their own language productions. *Clinical Linguistics & Phonetics, 27*(6-7), 508-520. https://doi.org/10.3109/02699206.2013.787546

Lukacs, A., Kas, B., & Leonard, L. B. (2013). Case marking in Hungarian children with specific language impairment. *First Language, 33*(4), 331-353. https://doi.org/10.1177/0142723713490601

Marinis, T., & Armon-Lotem, S. (2015). Sentence Repetition. In Armon-Lotem, S., de Jong, J. & Meir, N. (Eds.). *Methods for assessing multilingual children: disentangling bilingualism from Language Impairment*. Multilingual Matters.

McGregor, K. K. (2020). How we fail children with developmental language disorder. *Language, speech, and hearing services in schools, 51*(4), 981-992. https://doi.org/10.1044/2020_LSHSS-20-00003

Nag, S., Snowling, M. J., & Mirković, J. (2018). The role of language production mechanisms in children's sentence repetition: Evidence from an inflectionally rich language. *Applied Psycholinguistics, 39*(2), 303-325. https://doi.org/10.1017/S0142716417000200

Nash, H. M., Hulme, C., Gooch, D., & Snowling, M. J. (2013). Preschool language profiles of children at family risk of dyslexia: Continuities with specific language impairment. The *Journal of Child Psychology and Psychiatry, 54*(9), 958–968. https://doi.org/10.1111/jcpp.12091

Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development-Primary* (3rd edition). Pro-Ed.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America, 95*(2), 1085-1099. https://doi.org/10.1121/1.408469

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of child psychology and psychiatry, 57*(11), 1247-1257. https://doi.org/10.1111/jcpp.12573

Nudel, R., Christensen, R. V., Kalnak, N., Schwinn, M., Banasik, K., Dinh, K. M., ... & DBDS Genomic Consortium. (2023). Developmental language disorder–a comprehensive study of more than 46,000 individuals. *Psychiatry Research, 323*, 115171. https://doi.org/10.1016/j.psychres.2023.115171Oetting, J. B., McDonald, J. L., Seidel, C. M., & Hegarty, M. (2016). Sentence Recall by Children With SLI Across Two 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

Orsolini, M., Sechi, E., Maronato, C., Bonvino, E., & Corcelli, A. (2001). Nature of phonological delay in children with specific language impairment. *International Journal of Language & Communication Disorders, 36(*1), 63-90. https://doi.org/10.1080/13682820150217572

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology, 134*, 178-189. https://doi.org/10.1016/j.jclinepi.2021.03.001

Pawłowska, M. (2014). Evaluation of three proposed markers for language impairment in English: A meta-analysis of diagnostic accuracy studies. *Journal of Speech,* *Language, and Hearing Research, 57*(6), 2261-2273. https://doi.org/10.1044/2014_JSLHR-L-13-0189

Peristeri, E., Andreou, M., Tsimpli, I. M., & Durrleman, S. (2021). Bilingualism effects in the narrative comprehension of children with Developmental Language Disorder and L2-Greek. In U. Bohnacker & N. Gagarina (Eds.), *Studies in bilingualism* (Vol. 61, pp. 297–330). John Benjamins.

Petruccelli, N., Bavin, E. L., & Bretherton, L. (2012). Children with specific language impairment and resolved late talkers: Working memory profiles at 5 years. *Journal of Speech, Language, and Hearing Research, 55*(6), 1690–1703. https://doi.org/10.1044/1092-4388(2012/11-0288)

Pham, G., & Ebert, K. D. (2020). Diagnostic Accuracy of Sentence Repetition and Nonword Repetition for Developmental Language Disorder in Vietnamese. *Journal of Speech Language and Hearing Research, 63*(5), 1521-1536. https://doi.org/10.1044/2020_jslhr-19-00366

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15-24. https://doi.org/10.1044/0161-1461.2501.15

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language & Communication Disorders, 50*(1), 106-118. https://doi.org/10.1111/1460-6984.12126

Poll, G. H., Miller, C. A., & Van Hell, J. G. (2016). Sentence repetition accuracy in adults with developmental language impairment: Interactions of participant capacities and sentence structures. *Journal of Speech, Language, and Hearing Research, 59*(2), 302-316. https://doi.org/10.1044/2015_JSLHR-L-15-0020

Potter, M. C. (2012). Conceptual short term memory in perception and thought. *Frontiers in Psychology, 3*, 113. https://doi.org/10.3389/fpsyg.2012.00113

Redmond, S. M. (2005). Differentiating SLI from ADHD using children's sentence recall and production of past tense morphology. *Clinical Linguistics & Phonetics, 19*(2), 109–127. https://doi.org/10.1080/02699200410001669870

Redmond, S. M., & Ash, A. C. (2017). Associations between the 2D:4D proxy biomarker for prenatal hormone exposures and symptoms of developmental language disorder. *Journal of Speech, Language, and Hearing Research, 60*(11), 3226–3236. https://doi.org/10.1044/2017_jslhr-l-17-0143

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

Redmond, S. M., Ash, A. C., & Hogan, T. P. (2015). Consequences of co-occurring attention-deficit/hyperactivity disorder on children's language impairments. *Language, Speech, and Hearing Services in Schools, 46*(2), 68–80. https://doi.org/10.1044/2014_lshss-14-0045

Redmond, S. M., Thompson, H. L., & Goldstein, S. (2011). Psycholinguistic profiling differentiates specific language impairment from typical development and from attention deficit/hyperactivity disorder. *Journal of Speech, Language, and Hearing Research, 54*(1), 99–117. https://doi.org/10.1044/1092-4388(2010/10-0010)

Riches, N. (2015). Past tense -ed omissions by children with specific language impairment: The role of sonority and phonotactics. *Clinical Linguistics & Phonetics, 29*(6), 482-497. https://doi.org/10.3109/02699206.2015.1027832

Riches, N. G. (2012). Sentence repetition in children with specific language impairment: an investigation of underlying mechanisms. *International Journal of Language & Communication Disorders, 47*(5), 499-510. https://doi.org/10.1111/j.1460-6984.2012.00158.x

Riches, N. G. (2017). Complex sentence profiles in children with Specific Language Impairment: Are they really atypical? *Journal of Child Language, 44*(2), 269-296. https://doi.org/10.1017/s0305000915000847

Riches, N. G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2010). Sentence repetition in adolescents with specific language impairments and autism: An investigation of complex syntax. *International Journal of Language & Communication Disorders, 45*(1), 47–60. https://doi.org/10.3109/13682820802647676

Royle, P., & Thordardottir, E. (2003). *Le grand déménagement [French adaptation of the Recalling Sentences in Context subtest of the CELF–P]*. Unpublished research tool, McGill University, Montreal, Quebec, Canada.

Rujas, I., Mariscal, S., Murillo, E., & Lázaro, M. (2021). Sentence repetition tasks to detect and prevent language difficulties: A scoping review. *Children, 8*(7), 578. https://doi.org/10.3390/children8070578

Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa, K. (2021). Using nonword repetition to identify developmental language disorder in monolingual and bilingual children: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research, 64*(9), 3578-3593. https://doi.org/10.1044/2021_JSLHR-20-00552

Seeff-Gabriel, B., Chiat, S., & Dodd, B. (2010). Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe speech difficulties. *International Journal of Language & Communication Disorders, 45*(6), 691–702. https://doi.org/10.3109/13682820903509432

Seeff-Gabriel, B., Chiat, S., & Roy, P. (2008). *The early repetition battery*. Pearson Assessment.

Semel, E., Wiig, E., & Secord, W. (1994). *Clinical Evaluation of Language Fundamentals Revised*. The Psychological Corporation.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF–4)*. The Psychological Corporation.

Semel, E., Wiig, H., Secord, W., & Langdon, W. (2006). *CELF 4: Clinical evaluation of language fundamentals 4: Spanish edition*. Psychological Corporation.

Semel, E., Wiig, H., Secord, W., & Sabers, D. (1987). *CELF-R: Clinical Evaluation of Language Fundamentals – Revised (technical manual).* Psychological Corporation.

Smolík, F., Matiasovitsová, K., & Camarata, S. M. (2021). Sentence imitation with masked morphemes in Czech: Memory, morpheme frequency, and morphological richness. *Journal of Speech, Language, and Hearing Research, 64*(1), 105–120. https://doi.org/10.1044/2020_JSLHR-20-00370

Smolík, F., & Vávrů, P. (2014). Sentence Imitation as a Marker of SLI in Czech: Disproportionate Impairment of Verbs and Clitics. *Journal of Speech Language and Hearing Research, 57*(3), 837-849. https://doi.org/10.1044/2014

Spanoudis, G., & Pahiti, J. (2014). *Expressive and Receptive Language Evaluation: 5–12 Years of Age*. Department of Psychology, University of Cyprus

Stavrakaki, S., & Tsimpli, I. M. (2000). Diagnostic verbal IQ test for Greek preschool and school age children: Standardization, statistical analysis, psychometric properties. In *Proceedings of the 8th Symposium of the Panhellenic Association of Logopedists* (pp.95-106). Ellinika Grammata.

Stoel-Gammon, C. (2001). Transcribing the speech of young children. *Topics in language disorders, 21*(4), 12-21.

Stokes, S., & Fletcher, P. (2003). Aspectual forms in Cantonese children with specific language impairment*. Linguistics, 41(*2), 381-405. https://doi.org/10.1515/ling.2003.013

Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research, 49*(2), 219–236. https://doi.org/10.1044/1092-4388(2006/019)

Taha, J., Stojanovik, V., & Pagnamenta, E. (2021). Sentence Repetition as a Clinical Marker of Developmental Language Disorder: Evidence From Arabic. *Journal of Speech*

Language and Hearing Research, 64(12), 4876-4899. https://doi.org/10.1044/2021_jslhr-21-00244

Talli, I., & Stavrakaki, S. (2020). Short-term memory, working memory and linguistic abilities in bilingual children with Developmental Language Disorder. *First Language, 40*(4), 437–460. https://doi.org/10.1177/0142723719886954

Taylor, L. J., Maybery, M. T., Grayndler, L., & Whitehouse, A. J. (2014). Evidence for distinct cognitive profiles in autism spectrum disorders and specific language impairment. *Journal of Autism and Developmental Disorders, 44*(1), 19-30. https://doi.org/https://dx.doi.org/10.1007/s10803-013-1847-2

Theodorou, E., Kambanaros, M., & Grohmann, K. K. (2016). Diagnosing bilectal children with SLI: Determination of identification accuracy. *Clinical Linguistics & Phonetics, 30*(12), 925–943. https://doi.org/10.1080/02699206.2016.1182591

Theodorou, E., Kambanaros, M., & Grohmann, K. K. (2017). Sentence Repetition as a Tool for Screening Morphosyntactic Abilities of Bilectal Children with SLI. *Frontiers in Psychology, 8*, 2104. https://doi.org/10.3389/fpsyg.2017.02104

Thordardottir, E., & Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders, 46*(1), 1–16. https://doi.org/10.1016/j.jcomdis.2012.08.002

Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., Trudeau, N., & Chilingaryan, G. (2011). Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. *Journal of Speech, Language, and Hearing Research, 54*(2), 580–597. https://doi.org/10.10 44/1092-4388(2010/09-0196)

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of speech, language, and hearing research, 40*(6), 1245-1260. https://doi.org/10.1044/jslhr.4006.1245

Tsimpli, I. M., Peristeri, E., & Andreou, M. (2016). Narrative production in monolingual and bilingual children with specific language impairment. *Applied Psycholinguistics, 37*(1), 195–216. https://doi.org/10.1017/S0142716415000478

Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prevost, P., dos Santos, C., Abed Ibrahim, L., & Zebib, R. (2018). Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language & Communication Disorders, 53*(4), 888–904. https://doi.org/10.1111/1460-6984.12397

Van der Lely, H. K., & Howard, D. (1993). Children With Specific Language Impairment: Linguistic Impairment or Short-Term Memory Deficit? *Journal of Speech, Language, and Hearing Research, 36*(6), 1193-1207. https://doi.org/10.1044/jshr.3606.1193

Van Der Meulen, S., Janssen, P., & Os, E. D. (1997). Prosodic abilities in children with specific language impairment. *Journal of Communication Disorders, 30*(3), 155–170. https://doi.org/10.1016/S0021-9924(96)00059-7

Vang Christensen, R. (2019). Sentence repetition: A clinical marker for developmental language disorder in Danish. *Journal of Speech, Language, and Hearing Research, 62*(12), 4450–4463. https://doi.org/10.1044/2019_JSLHR-L-18-0327

Verhoeven, L., & Vermeer, A. (2001). *Taaltoets Alle Kinderen (TAK).* Cito.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Wang, D., Zheng, L., Lin, Y., Zhang, Y., & Sheng, L. (2022). Sentence Repetition as a Clinical Marker for Mandarin-Speaking Preschoolers with Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research, 65*(4), 1543-1560. https://doi.org/10.1044/2021_JSLHR-21-00401

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence-Revised.* The Psychological Corporation.

Wiig E. H., Semel E., Secord W. A. (2013). *Clinical Evaluation of Language Fundamentals–Fifth Edition (CELF-5)*. Pearson.

Winters, K. L., Jasso, J., Pustejovsky, J. E., & Byrd, C. T. (2022). Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research, 65*(10), 3908-3929. https://doi.org/10.1044/2022_JSLHR-22-00017

Wu, S., Zhao, J., de Villiers, J., Liu, X. L., Rolfhus, E., Sun, X., ... & Jiang, F. (2023). Prevalence, co-occurring difficulties, and risk factors of developmental language disorder: first evidence for Mandarin-speaking children in a population-based study. *The Lancet Regional Health Western Pacific, 34*, 1-11. http://doi.org/10.1016/j.lanwpc.2023.100713

Yang, S., & Berdine, G. (2017). The receiver operating characteristic (ROC) curve. *The Southwest Respiratory and Critical Care Chronicles, 5*(19), 34-36.

Ziethe, A., Eysholdt, U., & Doellinger, M. (2013). Sentence repetition and digit span: Potential markers of bilingual children with suspected SLI? *Logopedics Phoniatrics Vocology, 38*(1), 1–10. https://doi.org/10.3109/14015439.2012.664652

Appendix A

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).

Appendix B

Table B

Table showing likelihood ratio tests comparing model fit

Model	AIC	*pval*
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

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

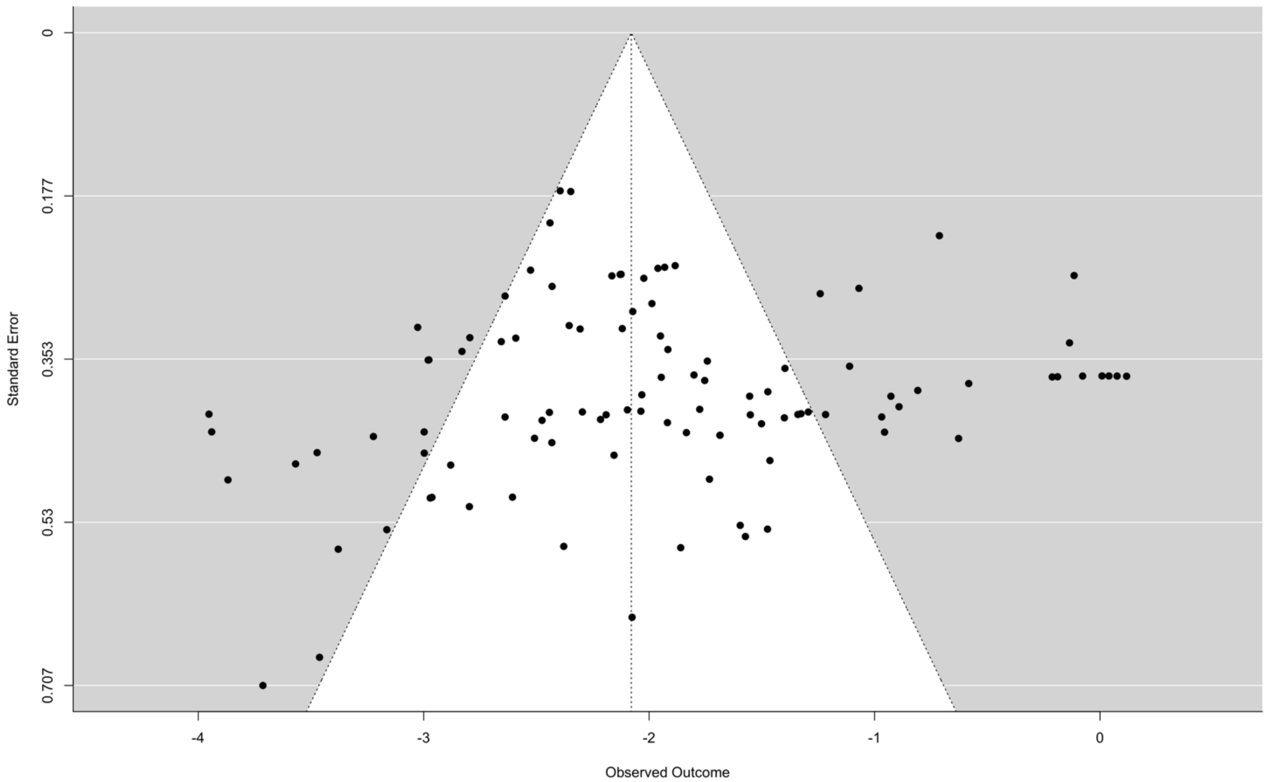
[a] The multilevel meta-analysis model where effect sizes are nested by study is the model chosen for the final analysis.

Appendix C

Assessment of Publication Bias

Figure C-1

Funnel plot of effects



Note. As can be seen, there was some evidence of asymmetry. To detect potential outliers and data points contributing most to this asymmetry, Cook's distances were calculated for each data point. Studies with the highest Cook's distance were removed until the asymmetry was no longer evident. Through this analysis, four effect sizes were removed. The removal of these studies resulted in the funnel plot and Egger's regression test shown in Figure C2 and Table C2.

Table C-1

Results of Egger's regression test

z	p
0.67	0.00122

Figure C-2

Funnel plot of effects, after the removal of outliers

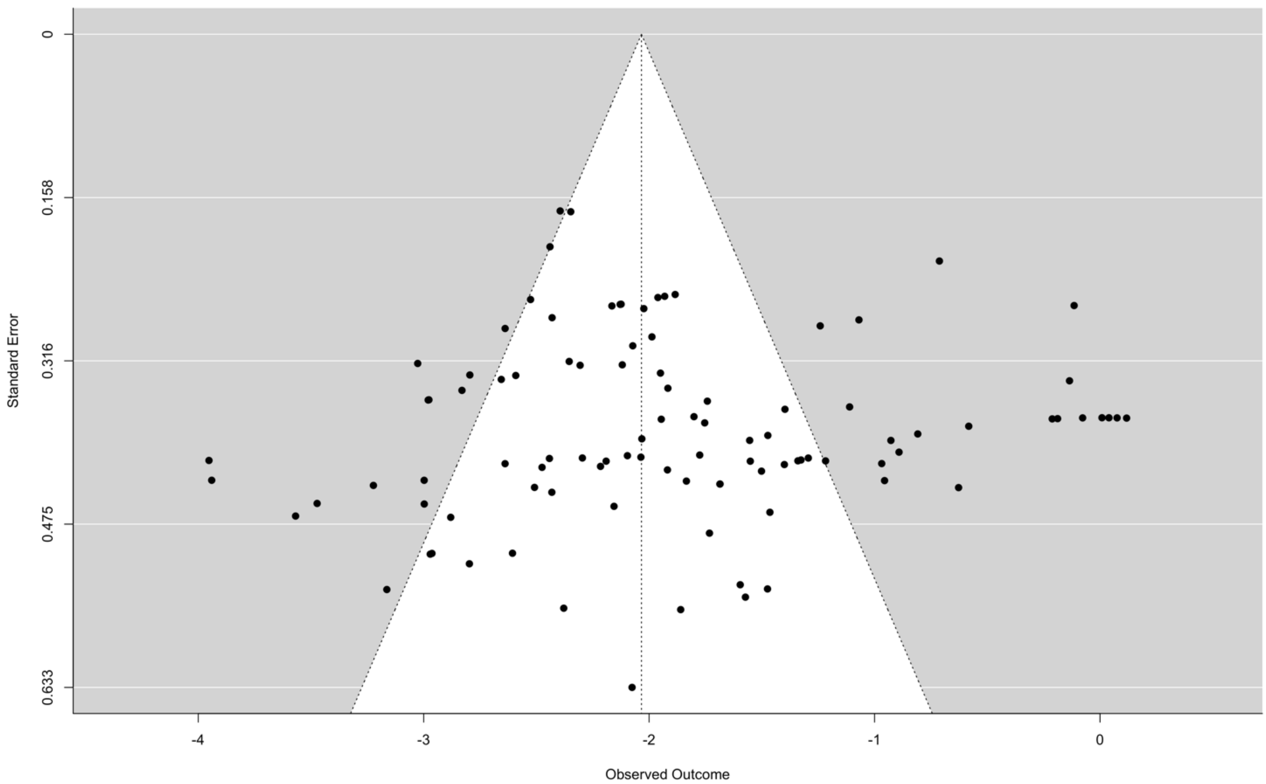


Table C-2

Results of Egger's regression test, after the removal of outliers

z	p
0.45	0.0532

Supplemental Material S1

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

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

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

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

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

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

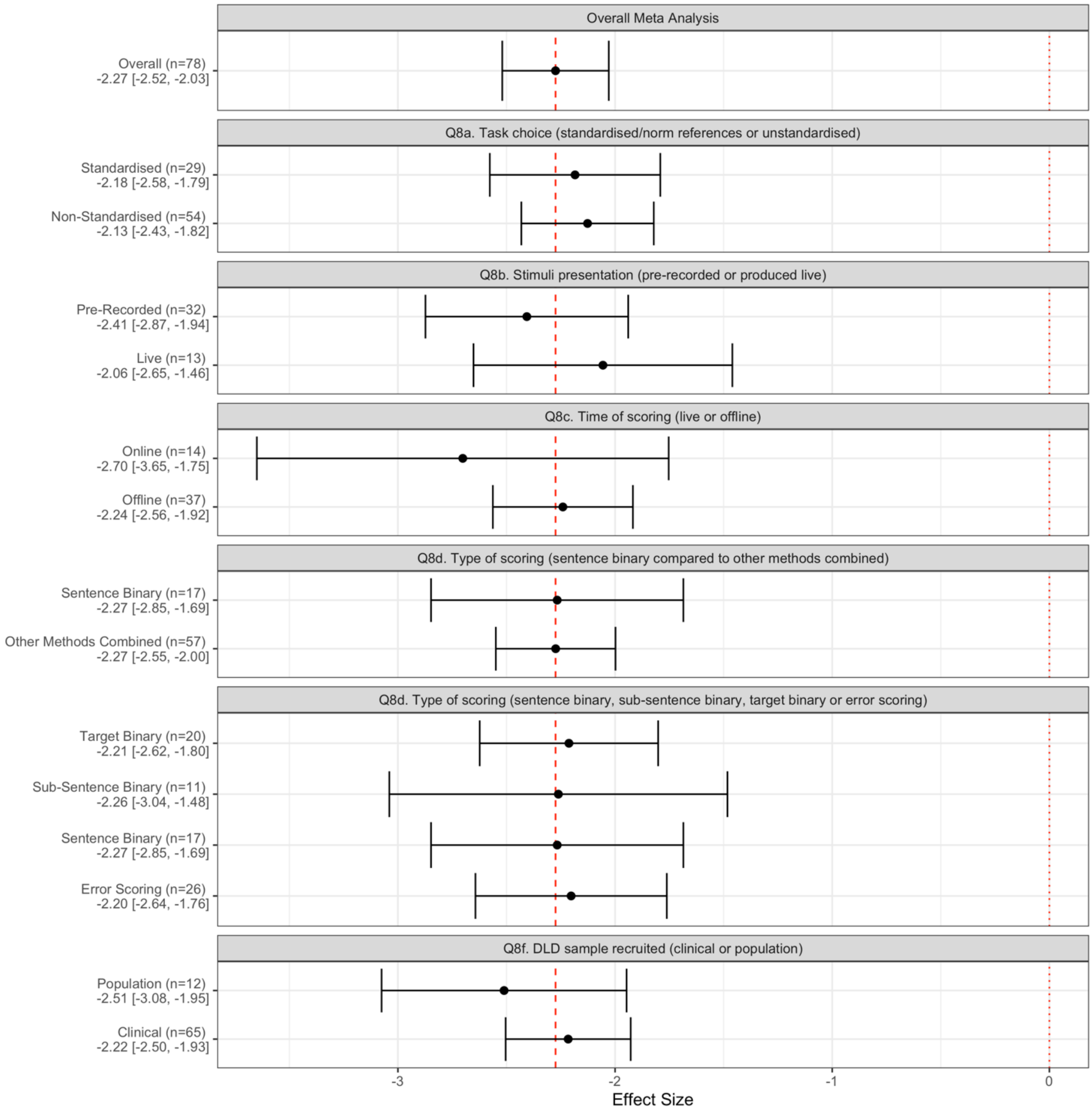
RQ8d. Type of scoring (sentence binary, sub-sentence binary, target binary or error scoring) – There were two separate subgroup analyses run here. The first compared each of the four categorised types of scoring (sentence binary, sub-sentence binary, target scoring, and error scoring). This test for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

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

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

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

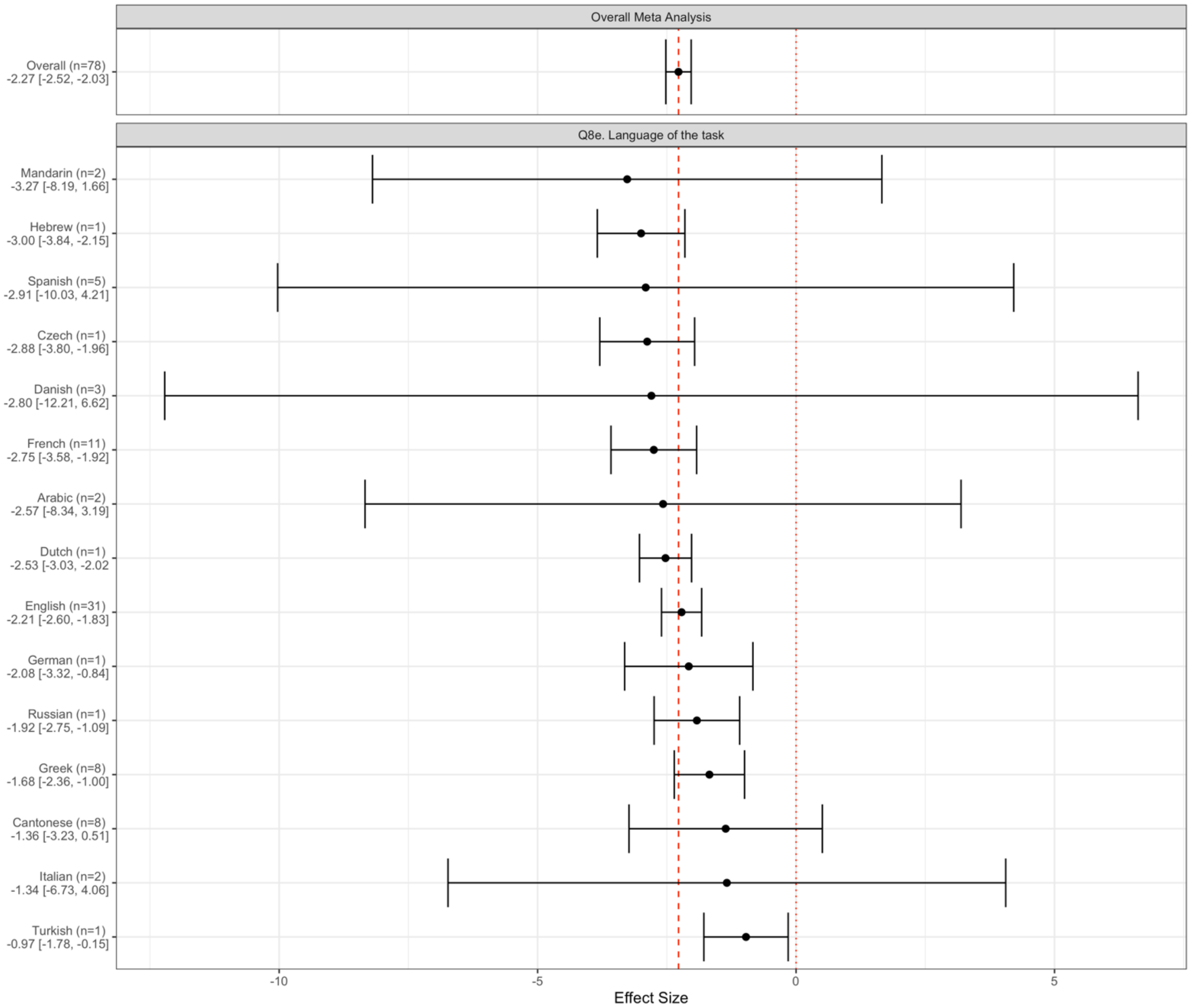
1501 studies).

1502

1503

Figure S3-2

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



Note. Datapoints are presented in order of effect

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