frontiers in Pediatrics

SYSTEMATIC REVIEW
published: 23 February 2022
doi: 10.3389/fped.2022.801220

Systematic Review and Meta-Analysis of Screening Tools for Language Disorder

*Kevin K. H. So and Carol K. S. To**

Academic Unit of Human Communication, Development, and Information Sciences, Faculty of Education, The University of Hong Kong, Hong Kong, Hong Kong SAR, China

Language disorder is one of the most prevalent developmental disorders and is associated with long-term sequelae. However, routine screening is still controversial and is not universally part of early childhood health surveillance. Evidence concerning the detection accuracy, benefits, and harms of screening for language disorders remains inadequate, as shown in a previous review. In October 2020, a systematic review was conducted to investigate the accuracy of available screening tools and the potential sources of variability. A literature search was conducted using CINAHL Plus, ComDisCome, PsycInfo, PsycArticles, ERIC, PubMed, Web of Science, and Scopus. Studies describing, developing, or validating screening tools for language disorder under the age of 6 were included. QUADAS-2 was used to evaluate risk of bias in individual studies. Meta-analyses were performed on the reported accuracy of the screening tools examined. The performance of the screening tools was explored by plotting hierarchical summary receiver operating characteristic (HSROC) curves. The effects of the proxy used in defining language disorders, the test administrators, the screening-diagnosis interval and age of screening on screening accuracy were investigated by meta-regression. Of the 2,366 articles located, 47 studies involving 67 screening tools were included. About one-third of the tests (35.4%) achieved at least fair accuracy, while only a small proportion (13.8%) achieved good accuracy. HSROC curves revealed a remarkable variation in sensitivity and specificity for the three major types of screening, which used the child's actual language ability, clinical markers, and both as the proxy, respectively. None of these three types of screening tools achieved good accuracy. Meta-regression showed that tools using the child's actual language as the proxy demonstrated better sensitivity than that of clinical markers. Tools using long screening-diagnosis intervals had a lower sensitivity than those using short screening-diagnosis intervals. Parent report showed a level of accuracy comparable to that of those administered by trained examiners. Screening tools used under and above 4yo appeared to have similar sensitivity and specificity. In conclusion, there are still gaps between the available screening tools for language disorders and the adoption of these tools in population screening. Future tool development can focus on maximizing accuracy and identifying metrics that are sensitive to the dynamic nature of language development.

Systematic Review Registration: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=210505, PROSPERO: CRD42020210505.

Keywords: surveillance, screening, language disorder, PRISMA review, meta-analysis, summary receiver-operating characteristics, meta-regression

OPEN ACCESS

Edited by:
Daniel Holzinger, Hospitaller Brothers of Saint John of God Linz, Austria

Reviewed by:
Karin Wiefferink, Dutch Foundation for the Deaf and Hearing Impaired Child (NSDSK), Netherlands
Steffi Sachse, Heidelberg University of Education, Germany

****Correspondence:***
Carol K. S. To
tokitsum@hku.hk

Specialty section:
This article was submitted to Children and Health, a section of the journal Frontiers in Pediatrics

Received: *25 October 2021*
Accepted: *13 January 2022*
Published: *23 February 2022*

Citation:
So KKH and To CKS (2022) Systematic Review and Meta-Analysis of Screening Tools for Language Disorder. Front. Pediatr. 10:801220. doi: 10.3389/fped.2022.801220

INTRODUCTION

Language disorder refers to persistent language problems that can negatively affect social and educational aspects of an individual's life (1). It is prevalent and estimated to affect around 7.6% of the population (2). Children with language disorder may experience difficulties in comprehension and/or in the use of expressive languages (3). Persistent developmental language disorder not only has a negative impact on communication but is also associated with disturbance in various areas such as behavioral problems (4), socio-emotional problems (5), and academic underachievement (6).

Early identification of persistent language disorder is challenging. There are substantial variabilities in the trajectories of early language development (7, 8). Some children display consistently low language, some appear to resolve the language difficulties when they grow older, and some demonstrated apparently typical early development but develop late-emerging language disorder. This dynamic nature of early language development has introduced difficulties in the identification process in practice (9). Therefore, rather than a one-off assessment, late talkers under 2 years old are recommended to be reassessed later. Referral to evaluation may not be not based on positive results in universal screening, but mainly concerns from caregivers, the presence of extreme deviation in development, or the manifestation of behavioral or psychiatric disturbances under 5 years old (9). Those who have language problems in the absence of the above conditions are likely to be referred for evaluation after 5 years old. Only then will they usually receive diagnostic assessment.

Ideally, screening should identify at-risk children early enough to provide intervention and avoid or minimize adverse consequences for them, their families, and society, improving the well-being of the children and the health outcomes of the population at a reasonable cost. Despite the high prevalence and big impact of language disorder, universal screening for language disorder is not practiced in every child health surveillance. Screening in the early developmental stages is controversial (10). While early identification has been advocated to support early intervention, there are concerns about the net cost and benefits of these early screening exercises. For example, the US Preventive Task Force reviewed evidence concerning screening for speech and language delay and concluded that there was inadequate evidence regarding the accuracy, benefits, and harms of screening. The Task Force therefore did not support routine screening in asymptomatic children (11). This has raised concerns in the professional community who believe in the benefits of routine screening (12). However, it is undeniable that another contributing factor for the recommendation of the Task Force was that screening tools for language disorder vary greatly in design and construct resulting in the variability in identification accuracy.

Previous reviews of screening tools for early language disorders have shown that these tools make use of different proxies for defining language issues, including a child's actual language ability, clinical markers such as non-word repetition, or both (13). Screening tools have been developed for children at different ages [e.g., toddlers (14) and preschoolers (15)] given the higher stability of language status at a later time point (16, 17). Screening tools also differ in the format of administration. For example, some tools are in the form of a parent-report questionnaire while some have to be administered by trained examiners via direct assessment or observations. Besides the test design, methodological variations have also been noted in primary validation studies, such as the validation sample, the reference standards (i.e., the gold standard for language disorder), and the screening-diagnosis interval. These variations might eventually lead to different levels of screening accuracy, which has been pointed out in previous systematic reviews (10, 13).

These variations have been examined in terms of the screening accuracy (13). Parent-report instruments and trained-examiner screeners have been found to be comparable in screening accuracy. In longitudinal studies in which language disorder status has been validated at various time points, accuracy appears to be lower for longer-term prediction than for concurrent prediction. Although the reviews have provided a comprehensive overview regarding the variations in different language screening tools, the analyses have mainly been based on qualitative and descriptive data. In the current study we performed a systematic review of all currently available screening tools for early language disorders that have been validated against a reference standard. We report on the variations noted in terms of (1) the type of proxy used in defining language disorders, (2) the type of test administrators, (3) the screening-diagnosis intervals and (4) age of screening. Second, we conducted a meta-analysis of the diagnostic accuracy of the screening tools and examined the contributions of the above four factors to accuracy.

METHODS

The protocol for the current systematic review was registered at PROSPERO, an international prospective register of systematic reviews (Registration ID: CRD42020210505, record can be found on https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=210505). Due to COVID-19, the registration was published with basic automatic checks in eligibility by the PROSPERO team. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Diagnostic Test Accuracy (PRISMA-DTA) (18) checklist was used as a guide for the reporting of this review.

Search Strategy

A systematic search of the literature was conducted in 2020 October based on the following databases: CINAHL Plus, ComDisDome, PsycINFO, PsycArticles, ERIC, PubMed, Web of Science, and Scopus. The major search terms were as follows: Child* OR Preschool* AND "Language disorder" * OR "language impairment*" OR "language delay" AND Screening OR identif*. To be as exhaustive as possible, the earliest studies available in the databases and those up to October 2020 were retrieved and screened. **Appendix A Table A1** showed the detailed search strategies in each database. Articles from the previous reviews were also retrieved.

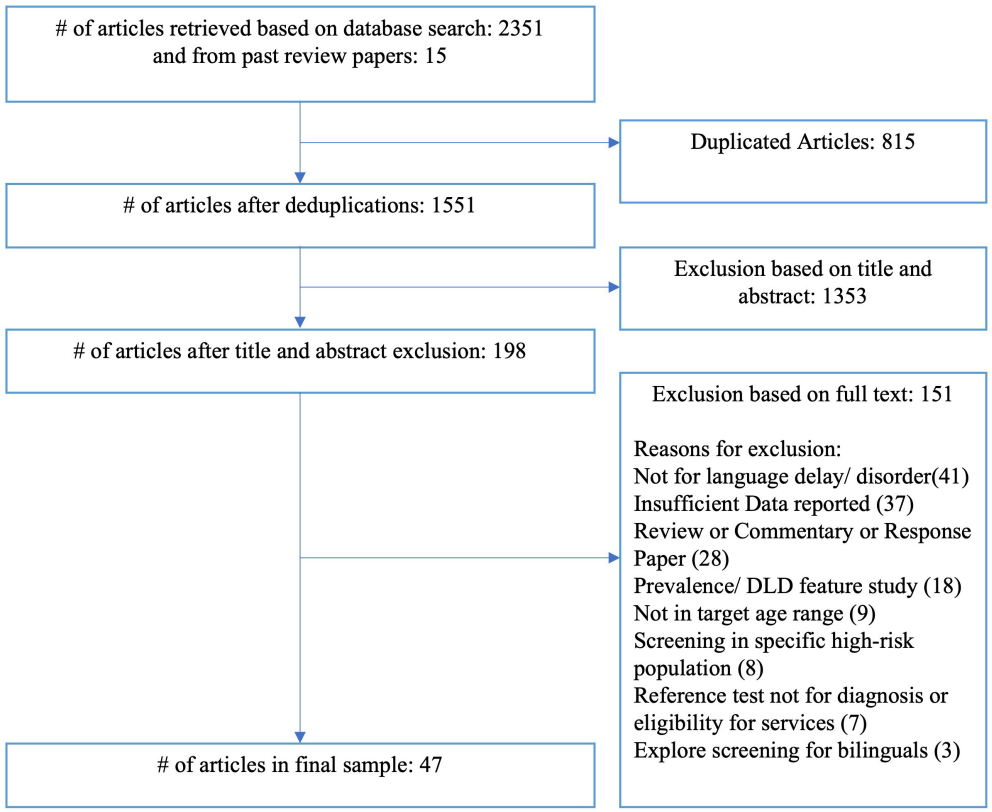

FIGURE 1 | Flow-chart for the inclusion and exclusion of articles in literature search.

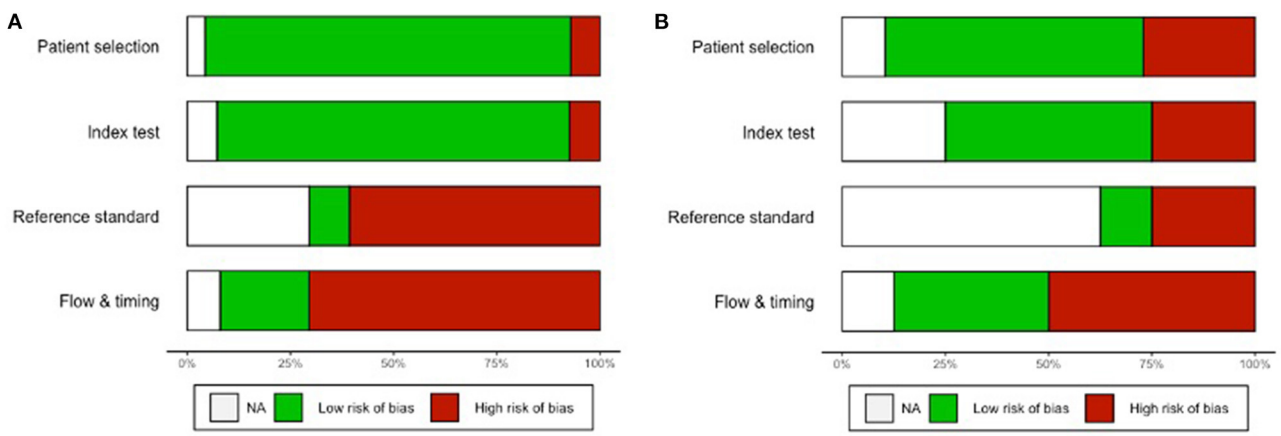

FIGURE 2 | (A) Weighted and **(B)** unweighted overall risk-of-bias as assessed using QUADAS-2.

Inclusion and Exclusion Criteria

The relevance of the titles, abstracts, and then the full texts were determined for eligibility. Cross-sectional or prospective studies validating screening tools or comparing different screening tools for language disorders were included in the review. The focus was on screening tools validated with children aged 6 or under from the general population or those with referral, regardless of the administration format of the tools, or how language disorder was defined in the studied. Studies that did not report adequate data on the screening results, and in which accuracy data cannot be deduced from the data reported, were excluded from the review (see **Appendix A Table A2** for details).

Data Extraction

Data was extracted by the first author using a standard data extraction form. The principal diagnostic accuracy measures

extracted were test sensitivity and specificity. The number of people being true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) was also extracted. Sensitivity and specificity were calculated based on 2 by 2 contingency tables in the event of discrepancy between the text description and the data reported. The data extraction process was repeated after the first extraction to improve accuracy. Screening tools with both sensitivity and specificity exceeding 0.90 were regarded as good and those with both measures exceeding 0.80 but below 0.90 were regarded as fair (19).

Quality Assessment

Quality assessment of included articles was conducted by the first author using QUADAS-2 by Whiting, Rutjes (20). QUADAS-2 can assist in assessing risk of bias (ROB) in diagnostic test accuracy studies with signaling questions concerning four major domains. The ROB in patient selection, patient flow, index tests, or the screening tools in the current review, and the reference standard tests were evaluated. Ratings of ROB for individual studies were illustrated using a traffic light plot. A summary ROB figure weighted with sample size was generated using the R package "robvis" (21). Due to the large discrepancy in the sample size across studies, an unweighted summary plot was also generated to show the ROB of the included studies.

Data Analysis

The overall accuracy of the tools was compared using descriptive statistics. Because sensitivity and specificity are correlated, increasing either one of them by varying the cut-off of test positivity would usually result in a decrease in the other. Therefore, a bivariate approach was used to jointly model sensitivity and specificity (22) in generating hierarchical summary receiver-operating characteristic (HSROC) curves to assess the overall accuracy of screening by proxy and by screening-diagnosis intervals. HSROC is a more robust method accounting for both within and between study variabilities (23).

Three factors that could be associated with screening accuracy, chosen *a priori*, were included in the meta-analysis: proxy used, test administrators, and screening-diagnosis interval. Effect of screening age on accuracy was also evaluated. The effect of each variable was evaluated using a separate regression model. The variables of proxy used were categorical, with the categories being "child's actual language," "performance in clinical markers," and "using both actual language and performance in clinical markers." Test administrator was also a categorical variable with the categories being "parent" and "trained-examiners." The variable of screening-diagnosis interval was dichotomously defined—intervals within 6 months were categorized as evaluating concurrent validity, whereas intervals of more than 6 months were categorized as evaluating predictive validity. The variable of screening age was also dichotomously defined with age 4 as the cut-off– those screened for children under the age of 4yo and those for children above 4yo. This categorization was primarily based on the age range of the sample, or the target screening age reported by the authors. Studies with age range that span across age 4 were excluded from the analysis. Considering the different thresholds used across studies and the correlated nature of sensitivity and specificity, meta-regression was conducted using a bivariate random effect model based on Reitsma et al. (22).

For studies examining multiple index tests and/or multiple cut-offs using the same population, only one screening test per category per study was included in the HSROC and meta-regression models. The test or cut-off with the highest Youden's index was included in the meta-analytical models. Youden's Index, *J*, was defined as

$$J = Sensitivity + Specificity - 1$$

All data analyses were conducted with RStudio Version 1.4.1106 using the package *mada* (24). Sensitivity analysis was carried out to exclude studies with a very high ROB (with 2 or more indicating a high risk in rating) to assess its influence on the results.

RESULTS

A total of 2351 articles, including 815 duplicates, were located using the search strategies, and an additional 15 articles were identified from previous review articles. After the inclusion and exclusion criteria were applied, a final sample of 47 studies were identified for inclusion in the review. **Figure 1** shows the number of articles included and excluded at each stage of the literature search.

Risk of Bias

The weighted overall ROB assessment for the 47 studies is shown in **Figure 2A**, and the individual rating for each study is shown in **Appendix B**. Overall, half of the data was exposed to a high ROB in the administration and interpretation of the reference standard test, while almost two-thirds of the data had a high ROB in the flow and timing of the study. As indicated by the unweighted overall ROB summary plot in **Figure 2B**, half of the 47 studies were unclear about whether the administration and interpretation of the reference standard test would introduce bias. This was mainly attributable to a lack of reporting of the reference standard test performance. About half of the studies had a high ROB in the flow and timing of the study. This usually arose from a highly variable or lengthy follow-up period.

Types and Characteristics of Current Screening Tools for Language Disorder

A total of 67 different index tests (or indices) were evaluated in the 47 included articles. The tests were either individual tests *per se* or part of a larger developmental test. The majority (50/67, 74.6%) of the screening tools examined children's actual language. Thirty of these index tests involved parents or caregivers as the main informants. Some of these screening tools were in the form of a questionnaire with Yes-No questions regarding children's prelinguistic skills, receptive language, or expressive language based on parent's observations. Some used a vocabulary checklist (e.g., CDI, LDS) in which parents checked off the vocabulary their child can was able to comprehend and/or produce. Some tools also asked parents to report

TABLE 1 | Studies involving tools based on a child's actual language ability.

References	Agent	Index test	Reference standard test(s)	Sc. age (months)[a]	*N*	SN	SP	Accuracy[b]	Included in meta-analysis
Allen and Bliss (25)	Trained personnel	The Northwestern Syntax Screening Test (26)	Sequenced inventory of communication development (27)	36–47	182	0.92	0.48	Below fair	✓
Blaxley et al. (28)	Trained personnel	Bankson Language Screening Test (29)	Developmental sentence scoring (30)	48–72	90	0.46	0.94	Below fair	✓
Burden et al. (31)	Parents/caregivers	The Parent Language Checklist and The Developmental Profile II (32)	Action Picture Test (33), Bus Story test (34), self-developed test on receptive and phonological ability	36–39	425	0.87	0.45	Below fair	✓
Carscadden et al. (35)	Parents/caregivers	Speech and Language Pathology Early Screening Instrument (35)	Receptive Expressive Emergent Language Test – 3rd Edition (36)	17–23	53	0.91	0.95	Good	✓
Chaffee et al. (37)	Parents/caregivers	Minnesota Child Development Inventory – Comprehension Conceptual Language	Reynell Developmental Language Scales – revised (38)	24–87 M = 49	152	0.76	0.63	Below fair	✓
		Minnesota Child Development Inventory – Expressive Language (39)				0.89	0.45	Below fair	×
Dias et al. (40)	Parents/caregivers	Screening Tool by ASHA (41)	ABFW test (42)	0–60	962	0.83	0.99	Fair	✓
Dixon et al. (43)	Trained personnel	The Hackney Early Language Screening Test (43)	Reynell Developmental Language Scales (44), Lowe and Costello Symbolic Play Test (45)	30	40	0.94	0.95	Good	✓
Gray et al. (46)	Trained personnel	Expressive One-word Picture Vocabulary Test (47)	Referred by speech-language pathologist	48–60	62	0.71	0.71	Below fair	×
		Peabody Picture Vocabulary Test – III (48)				0.74	0.71	Below fair	×
		Receptive One-word Picture Vocabulary Test (49)				0.77	0.77	Below fair	✓
		Expressive Vocabulary Test (50)				0.71	0.68	Below fair	×
Guiberson (14)	Parents/caregivers	Parent reported vocabulary	Bilingual early childhood assessment team identification, parent report of concern, Spanish Preschool Language Scale – 4th Edition (51)	24–35	62	0.86	0.88	Fair	✓

(Continued)

TABLE 1 | Continued

References	Agent	Index test	Reference standard test(s)	Sc. age (months)[a]	*N*	SN	SP	Accuracy[b]	Included in meta-analysis
		Parent report of mean length of child's three longest utterances				0.46	0.93	Below fair	×
Guiberson and Rodriguez (52)	Parents/caregivers	Pilot Inventories III, translated version of MacArthur- Bates Communicative Development Inventory-III (53)	Spanish Preschool Language Scale – 4th Edition (51)	36–62 M = 45.5	48	0.82	0.81	Fair	✓
		Ages and Stages Questionnaire – communication subscales (54)				0.59	0.92	Below fair	×
Guiberson et al. (55)	Parents/caregivers	Reported children's three longest utterances	Parent concern, enrollment in speech-language intervention services, Spanish Preschool Language Scale – 4th Edition (51)	24–35 M = 29.4	45	0.91	0.86	Fair	✓
		Ages and Stages Questionnaire – communication subscales (56)				0.56	0.95	Below fair	×
		The Inventarios del Desarrollo de Habilidades Communicatives Palabras u Enunciado (57)				0.87	0.86	Fair	×
Guiberson et al. (58)	Parents/caregivers	Vocabulary score	SLP assessment, parental concern, Spanish Preschool Language Scale – 4th Edition (51)	37–69 M = 53.7	82	0.79	0.77	Below fair	✓
		Language questions				0.74	0.69	Below fair	×
Heilmann et al. (59)	Parents/caregivers	MacArthur- Bates Communicative Development Inventory – Words and Sentences (60)	Preschool Language Scale – 3rd Edition (61), language sampling	24 M = 23.8	100	0.68	0.98	Below fair	✓
Klee et al. (62)	Parents/caregivers	The Language Development Survey (63)	Mullen Scales of Early Learning (64), language sampling, parent interview, direct observation	24–26 M=24.7	64	0.91	0.87	Fair	×[c]

(Continued)

TABLE 1 | Continued

References	Agent	Index test	Reference standard test(s)	Sc. age (months)[a]	*N*	SN	SP	Accuracy[b]	Included in meta-analysis
Klee et al. (65)	Parents/caregivers	The Language Development Survey (63)	Mullen Scales of Early Learning (64), language sampling, language sampling, parent interview, direct observation	24–26 M = 24.7	64	0.91	0.96	Good	✓
Laing et al. (66)	Trained Personnel	Structured Screening Test	Reynell Developmental Language Scales – III (67)	30–36 M=32	282	0.66	0.89	Below fair	✓
Law (68)	Trained personnel	Structured Screening Test	Reynell Developmental Language Scales (2nd revision) (44)	30	189	0.86	0.76	Below fair	✓
Levett and Muir (69)	Trained personnel	Levett-Muir Language Screening Test (69)	Reynell Developmental Language Scales (revised) (70), Goldman-Fristoe Test of Articulation (71), Language Assessment and Remediation Procedure (72)	34.9–39.6	42	1	1	Good	✓
Visser-Bochane et al. (73)	Parents/caregivers	Early Language Screen (73)	LLC (74), SLC (75), LLP (76), SWP, SSP (77), LS-CCS (78), CCC-PCS (79)	12–72	124	0.79	0.86	Below fair	✓
Visser-Bochane et al. (80)	Trained personnel	The Dutch well child language screening protocol (80)	SLC (75), SWP, SSP (77)	26	265	0.62	0.93	Below fair	✓
Mattsson et al. (81)	Parents/caregivers and trained personnel	Questionnaire and Direct Observation by nurse	Clinical Examination by SLP	28–32 M = 30	105	0.81	0.87	Fair	✓
McGinty (82)	Parents/caregivers and trained personnel	The Mayo Early Language Screening Test (83)	Reynell Developmental Language Scales (44), Edinburgh Articulation Test (84)	18–60	200	0.84	0.7	Below fair	✓
Nair et al. (85)	Trained personnel	The Language Evaluation Scale Trivandrum For 0–3 Years (85)	Receptive-Expressive Emergent Language Scale (86)	0–36	643	0.96	0.78	Below fair	✓
Nayeb et al. (87)	Trained personnel	Nurse screening	Clinical Examination by SLP	29–31	100	1	0.85	Fair	✓
Puglisi et al. (15)	Trained personnel	Screening for Identification of Oral Language Difficulties by Preschool Teachers (15)	Expressive Vocabulary Test (88), Test for Reception of Grammar Version 2 (89), The Brazilian Children's Test of Pseudoword Repetition (90),	51–65 M = 57	100	0.86	0.95	Fair	✓
Rescorla (63)	Parents/caregivers	The Language Development Survey (63)	Reynell Developmental Language Scales (38)	23.7–34.4 M = 25.9	81	0.76	0.89	Below fair	✓
Rescorla and Alley (91)	Parents/caregivers	The Language Development Survey (63)	Reynell Developmental Language Scales (44)	23.7–34.4 M = 25.9	66	0.89	0.77	Below fair	✓

(Continued)

TABLE 1 | Continued

References	Agent	Index test	Reference standard test(s)	Sc. age (months)[a]	*N*	SN	SP	Accuracy[b]	Included in meta-analysis
Sachse and Von Suchodoletz (92)	Parents/caregivers	German version of the CDI, Toddler Form-2 (93)	Language Test for 2-Year-Old Children (94)	24–26	117	0.93	0.87	Fair	✓
Stokes (95)	Trained personnel	Nurse screen	Language sampling, Reynell Developmental Language Scales (70)	34–40	366	0.77	0.97	Below fair	✓
	Parent/caregivers	Parent Questionnaire				0.75	0.95	Below fair	×
van Agt et al. (96)	Parents/caregivers	Van Wiechen (96)	Specialists' judgement	26–58 M = 39	8,877	0.71	0.89	Below fair	✓
		General Language Screen (97)				0.81	0.78	Below fair	×
		Language Screening Instrument – Parent Form (98)				0.86	0.73	Below fair	×
	Trained personnel	Language Screening Instrument – Child Test (98)				0.54	0.88	Below fair	×
Walker et al. (99)	Parents/caregivers	Early Language Milestone Scale (100)	Sequenced Inventory of Communication Development (27)	0–36	77	0.77	0.85	Below fair	✓
Wetherby et al. (101)	Parents/caregivers	Communication And Symbolic Behavior Scales – Developmental Profile, Infant-Toddler Checklist (102)	Behavior Sample	12–24 M = 14.5	151	0.89	0.74	Below fair	✓

For tests that were validated against multiple cut-offs, only the one with highest Youden's index was shown; Sc. Age, screening age; MA, Meta-analysis; ASHA, American Speech-Language and Hearing Association; ABFW, Andrade CRF, Befi-Lopes DM, Fernandes FDM, Wertzner HF. Teste de Language Infantil nas Áreas de Fonologia, Vocabulário, Fluência e Pragmática. 2nd ed. Barueri: Pró-Fono, 2011; LLC, Lexilist Comprehension; SLC, Schlichting test for Language Comprehension; LLP, Lexilist Production; SWP, Schlichting test for Word Production; SSP, Schlichting test for Sentence Production; LS-CCS – Language Standard – Communication Schlichting test for Language Composite Score; CCC-PCS, CCC-2-NL-Pragmatic Composite Score.

[a] *Age of screening is reported in range or mean in the form of X_1-X_2 and M=X_3; In case range or mean is not reported, the intended age for screening of the tool will be reports as X_4.*

[b] *Based on Plante and Vance (19), Fair = over 0.8 in both sensitivity and specificity; Good = over 0.9 in both sensitivity and specificity.*

[c] *Not included because the sample was identical to Klee et al. (65).*

their child's longest utterances according to their observation and generated indices. The other 20 index tests on language areas were administered by trained examiners such as nurses, pediatricians, health visitors or speech language pathologists (SLPs). These screening tools were constructed as checklists, observational evaluations, or direct assessments, tapping into children's developmental milestones, their word combinations and/or their comprehension, expression, and/or articulation. Some of these direct assessments involved the use of objects or pictures as testing stimuli for children.

A small proportion (3/67, 4.48%) of tests evaluated clinical markers performance including non-word repetitions and sentence repetitions rather than children's actual structural language skills or communication skills. About nine percent (6/67, 8.96%) screened for both language abilities and clinical markers. Both types of tests required trained examiners to administer them. The tests usually made use of a sentence repetition task and one test also included non-word repetition. Another nine percent (6/67, 8.96%) utilized indices from language sampling, such as percentage of grammatical utterances (PGU), mean length of utterances in words (MLU3-W), and number of different words (NDW) as proxies. These indices represented a child's syntactic, semantic, or morphological performance. The smallest proportion (2/67, 2.99%) of the tests elicited parental concerns about their children being screened for language disorder. One asked parents to rate their concern using a visual analog scale, while the other involved interviews with the parents by a trained examiner.

Sixty-five of the 67 screening tools had reported concurrent validity. **Tables 1–5** summarize the characteristics of these 65 studies by the proxy used. Nine studies investigated the predictive validity of screening tools. **Table 6** summarizes the studies. All the studies used child's actual language ability as the proxy.

Screening Accuracy

Two of the 67 screening tools only reported predictive validity. Of the 65 screening tools that reported concurrent validity, about one-third (23/65, 35.4%) achieved at least fair accuracy and a smaller proportion (9/65, 13.8%) achieved good accuracy. The nine tools which achieved good accuracy include (i) Non-word Repetition, (ii) Speech and Language Pathology Early Screening Instrument (35), (iii) The Hackney Early Language Screening Test (43), (iv) The Language Development Survey (63), (v) Levett-Muir Language Screening Test (69), (vi) The Grammar and Phonology Screening (GAPS) Test (105), (vii) Tamiz de Problemas de Lenguaje (113), (viii) The Screening Kit of Language Development (117) and (ix) Short Language Measures (120).

Screening Performance by Proxy and Screening-Diagnosis Interval

Screening tools based on children's actual language ability had a sensitivity ranging from 0.46 to 1 (median = 0.81) and a specificity of 0.45 to 1 (median = 0.86). About 30% of the studies showed that their tools achieved at least fair accuracy, while 8.89% achieved good accuracy. Screening tools using clinical markers had a sensitivity ranging from 0.3 to 1 (median =

TABLE 2 | Studies involving tools based on clinical marker.

References	Agent	Index test	Reference standard test(s)	Sc. age[a] (months)	*N*	SN	SP	Accuracy[b]	Included in meta-analysis
Guiberson et al. (58)	Trained personnel	Non-word Repetition	SLP assessment, parental concern, Spanish Preschool Language Scale – 4th Edition (51)	37–69 M = 53.7	82	0.74	0.75	Below fair	✓
Kapalkova et al. (103)	Trained personnel	Non-word repetition	Clinical judgment and qualitative assessment	51–66	32	0.94	1	Good	✓
Nash et al. (104)	Trained personnel	The Grammar and Phonology Screening (GAPS) Test (105)	Clinical Evaluation of Language Fundamentals – Preschool, 2nd Edition (106)	36–72 M = 62.3	106	0.3	0.91	Below fair	✓
Sturner et al. (107)	Trained personnel	The Sentence Repetition Screening Task (108)	Illinois Test of Psycholinguistic Abilities (109), Bankson Language Screening Test (29)	54–66 Med = 60	323	0.62	0.91	Below fair	✓
van der Lely et al. (110)	Trained personnel	The Grammar and Phonology Screening (GAPS) Test (105)	Assessment by SLP and educational psychologist	43–80	41	1	1	Good	✓

For tests that were validated against multiple cut-offs, only the one with highest Youden's index was shown; Sc. Age, screening age.

[a]*Age of screening is reported in range, mean or median in the form of X_1-X_2, M=X_3 and Med=X_4, respectively.*

[b]*Based on Plante and Vance (19), Fair = over 0.8 in both sensitivity and specificity; Good = over 0.9 in both sensitivity and specificity.*

TABLE 3 | Studies involving tools based on both language ability and clinical marker.

Study	Agent	Index test	Reference standard test(s)	Sc. age[a] (months)	*N*	SN	SP	Accuracy[b]	Included in meta-analysis
Allen and Bliss (25)	Trained personnel	The Fluharty Preschool Screening Test (111)	Sequenced Inventory of Communication Development (112)	36–47	182	0.6	0.81	Below fair	✓
Benavides et al. (113)	Trained personnel	Tamiz de Problemas de Lenguaje (113)	Clinical Evaluation of Language Fundamentals-5th edition, Spanish Version (114)	48–72	200	0.94	0.92	Good	✓
Blaxley et al. (28)	Trained personnel	The Fluharty Preschool Screening Test (115)	Developmental Sentence Scoring (116)	48–72	90	0.36	0.96	Below fair	✓
Bliss and Allen (117)	Trained personnel	The Screening Kit of Language Development (118)	Sequenced Inventory of Communication Development (112), clinical judgment by SLP	30–48	100	1	0.93	Good	✓
Lavesson et al. (119)	Trained personnel	Language tasks and non-word repetition (119)	SLP judgment based on test results	46–53 M = 48.5	328	0.84	0.96	Fair	✓
Matov et al. (120)	Trained personnel	Short Language Measures (121)	Clinical Evaluation of Language Fundamentals-4 (122)	63.6	126	0.94	0.93	Good	✓
Wright and Levin (123)	Trained personnel	Preschool Articulation and Language Screening (123)	SLP judgement based on test results	26–81	152	0.71	0.94	Below fair	✓

For tests that were validated against multiple cut-offs, only the one with highest Youden's index was shown; Sc. Age, screening age.

[a] *Age of screening is reported in range or mean in the form of X_1-X_2 and M=X_3; In case range or mean is not reported, the intended age for screening of the tool will be reports as X_4..*

[b] *Based on Plante and Vance (19), Fair = over 0.8 in both sensitivity and specificity; Good = over 0.9 in both sensitivity and specificity.*

TABLE 4 | Studies involving tools based on language sampling.

References	Agent	Index test	Reference standard test(s)	Sc. age[a] (months)	*N*	SN	SP	Accuracy[b]	Included in meta-analysis
Eisenberg and Guo (124)	Trained personnel	Percentage Grammatical Utterances	LI2: Previously diagnosed LI3: Parent rating, Structured Photographic Expressive Language Test – Preschool 2nd Edition (125)	36–47	34	1	0.88	Fair	✓
		Percentage Sentence Point			34	1	0.82	Fair	×
		Percentage Verb Tense Usage (126)			34	1	0.82	Fair	×
Guiberson et al. (58)	Trained personnel	Ungrammaticality Index	SLP assessment, parental concern, Spanish Preschool Language Scale – 4th Edition (51)	37–69 M = 53.7	82	0.59	0.67	Below fair	×
	Trained personnel	Mean Length of Utterances in Words				0.65	0.92	Below fair	✓
Guiberson (14)	Parents/caregivers	Number of Different Words	Bilingual early childhood assessment team identification, parent report of concern, Spanish Preschool Language Scale – 4th Edition (51)	24–35	62	0.73	0.83	Below fair	✓

For tests that were validated against multiple cut-offs, only the one with highest Youden's index was shown; Sc. Age, screening age; LI2, language impairment at age 2; LI3, language impairment at age 3.

[a] *Age of screening is reported in range or mean in the form of X_1-X_2 and M=X_3; In case range or mean is not reported, the intended age for screening of the tool will be reports as X_4.*

[b] *Based on Plante and Vance (19), Fair = over 0.8 in both sensitivity and specificity; Good = over 0.9 in both sensitivity and specificity.*

TABLE 5 | Studies involving tools based on parental concern.

References	Agent	Index test	Reference standard test(s)	Sc. age[a] (months)	*N*	SN	SP	Accuracy[b]	Included in meta-analysis
Laing et al. (66)	Parents/caregivers	Parent led method	Reynell Developmental Language Scales – III (67)	30–36 M = 32	176	0.79	0.74	Below fair	✓
van Agt et al. (96)	Parents/caregivers	Visual analog scale to evaluate child's language development	Specialists' judgement	26–58 M = 39	8,877	0.76	0.81	Below fair	✓

For tests that were validated against multiple cut-offs, only the one with highest Youden's index was shown; Sc. Age, screening age.

[a]*Age of screening is reported in range or mean in the form of X_1-X_2 and M=X_3; In case range or mean is not reported, the intended age for screening of the tool will be reports as X_4.*

[b]*Based on Plante and Vance (19), Fair = over 0.8 in both sensitivity and specificity; Good = over 0.9 in both sensitivity and specificity.*

0.71) and a specificity of 0.45 to 1 (median = 0.91). Two of the five studies[1] (40%) evaluating screening tools based on clinical markers showed their tools had good sensitivity and good specificity, but the other three studies showed a sensitivity and a specificity below fair. Concerning screening tools based on both actual language ability and clinical marker performance, the sensitivity ranged from 0.36 to 1 (median = 0.84), and the specificity ranged from 0.81 to 0.96 (median=0.93) and above half of these studies (4/7[2], 57.1%) achieved at least fair performance in both sensitivity and specificity, and 3 of the 7 studies achieved good performance. Screening tools based on indices from language sampling had sensitivity ranging from 0.59 to 1 (median = 0.865) and specificity ranging from 0.67 to 0.92 (median = 0.825). Half of these six screening tools achieved fair accuracy, but none achieved good accuracy. None of the two screening tools based on parental concern achieved at least fair screening accuracy.

Fifteen of the 65 studies also reported predictive validity, with a sensitivity ranging from 0.32 to 0.94 (median = 0.81) and a specificity ranging from 0.61 to 0.93 (median = 0.85). Three of the tools (20%) achieved at least fair accuracy in both sensitivity and specificity, but none of them were considered to have good accuracy.

Test Performance Based on HSROC

Three HSROC curves were generated for screening tools based on language ability, clinical markers, both language ability and clinical markers, and those assessing concurrent validity. Two HSROC curves were generated for screening tools administered by trained examiners and parents/ caregivers, respectively. Two HSROC curves were generated for screening under and above the age of 4, respectively. A separate HSROC curve was generated for screening tools assessing predictive validity. Screening based on indices from language sampling ($n = 3$) or parental concern ($n = 2$) were excluded from the HSROC analysis due to the small number of primary studies.

Figure 3 shows the overall performance of screening tools based on language ability, clinical markers and both. Visual inspection of the plotted points and confidence region revealed considerable variation in accuracy in all three major types of screening tools. The summary estimates and confidence regions indicated that the overall performance of screening tools based on language ability achieved fair specificity (<0.2 in false positive rate) but fair-to-poor sensitivity. Screening tools based on clinical markers showed considerable variation in both sensitivity and specificity in that both measures ranged from good-to-poor. Screening tools based on both language ability and clinical markers achieved good-to-fair specificity, but fair-to-poor sensitivity. **Figure 4** shows the overall performance of

[1]The total number here refers to the number of studies: there were five studies evaluating tools based on clinical markers, but there were in total three different tests; hence number is different from that in types and characteristics of current screening tools for language disorder.

[2]The total number here refers to the number of studies: there were seven studies evaluating tools based on both actual language and clinical markers, but there were in total six different tests; hence number is different from that in types and characteristics of current screening tools for language disorder.

TABLE 6 | Studies assessing predictive validity of screening tools.

References	Agent	Index test	Sc. age (months)	Sc-V int. (months)	F/U age (months)	Reference standard test(s)	*N*	SN	SP	Accuracy[a]	MA included
Bruce et al. (127)	Parents/caregivers and trained personnel	Direct assessment through play and parent questionnaire	18–22	NA	54	NELLI (128)[b], The Test for Reception of Grammar (129)	43	0.6	0.85	Below fair	✓
Frisk et al. (130)	Trained personnel	Early Screening Profiles (131)	54	NA	60	Preschool Language Scale – 4th Edition (132)	110	0.86	0.81	Fair	✓
	Parents/caregivers	Ages and Stages Questionnaire (54)				Bracken Basic Concepts Scale	110	0.84	0.66	Below fair	×
	Trained personnel	Battelle Developmental Inventory Screening Test (133)				Preschool (134) Language Scale – 4th Edition (132)	110	0.68	0.86	Below fair	×
	Trained personnel	Brigance Preschool Screen (135)				Preschool Language Scale – 4th Edition (132)	110	0.91	0.78	Below fair	×
Jessup et al. (136)	Trained personnel	Kindergarten Development Check (137)	48–54	8–12	NA	Clinical Evaluation of Language Fundamentals-4 (122)	286	0.5	0.93	Below fair	✓
Klee et al. (62)	Parents/caregivers	The Language Development Survey (63)	24	NA	36–40	Mullen Scales of Early Learning (64), language sampling, parent interview, direct observation	36	0.67	0.9	Below fair	✓
Pesco and O'Neill (138)	Parents/caregivers	Language Use Inventory (139)	24–47	14.54–54.76	NA	DELV- NR (140) CELF-2 (141), Children's Communication Checklist – 2nd Edition (142)	236	0.81	0.93	Fair	✓
Sachse and Von Suchodoletz (92)	Parent/caregivers	German Version of The CDI, Toddler Form-2 (93)	24–26	12	NA	Language Test For 3–5-Year-Old Children (94)	102	0.94	0.61	Below Fair	×
	Trained personnel	Language Test for 2-Year-Old Children (94)	24–26	12	NA	Language Test For 3–5-Year-Old Children (94)	102	0.94	0.64	Below Fair	✓
Visser-Bochane et al. (80)	Trained personnel	The Dutch well-child language screening protocol (80)	$M = 26$	12	NA	SLC (75), SWP, SSP (77)	123	0.82	0.74	Below Fair	✓
Westerlund et al. (143)	Parents/caregivers	The Swedish Communication Screening at 18 Months of Age (144, 145)	18	NA	36	LO-3 (146, 147)	891	0.5	0.9	Below Fair	✓
	Trained personnel	Traditional Methods	18	NA	36	LO-3 (146, 147)	1,189	0.32	0.91	Below Fair	×
Wetherby et al. (101)	Parents/caregivers	Communication And Symbolic Behavior Scales – Developmental Profile Infant-Toddler Checklist (102)	12–24	$M = 14.5$	NA	Mullen Scales of Early Learning (148), Preschool Language Scale – 3rd Edition (61)	246	0.81	0.79	Below Fair	×

(Continued)

TABLE 6 | Continued

References	Agent	Index test	Sc. age (months)	Sc-V int. (months)	F/U age (months)	Reference standard test(s)	*N*	SN	SP	Accuracy[a]	MA included
	Trained personnel	Behavioral Sample	12–24	*M* = 18.2	NA	Mullen Scales Of Early Learning (148), Preschool Language Scale – 3rd Edition (61)	90	0.84	0.85	Fair	✓

For tests that were validated against multiple cut-offs, only the one with highest Youden's index was shown; Sc. Age, screening age; Sc-V int., Screening-validation Interval; F/U age, age at follow-up; DELV-NR, Diagnostic Evaluation of Language Variation – Norm Referenced; CELF-2, Clinical Evaluation of Language Fundamentals – Preschool, 2nd Edition; LO-3, Language Observation at 3 years of age.

[a]*Based on Plante and Vance (19), Fair = over 0.8 in both sensitivity and specificity; Good = over 0.9 in both sensitivity and specificity.*

[b]*Sprakltg snabbscreening av forskolebarn 3–6 arunderlag for diagnostisering av art och grad av sprakstorning, Stora Fonemtestet. Pedagogisk, Grammatiktest. Pedagogisk.*

[c]*Based on* ***Table 5*** *in the paper, description in the discussion differed from the figures in the table.*

screening tools administered by parents/caregivers or trained examiners. Visual inspection revealed that both types of screening tools achieved fair-to-poor sensitivity and good-to-fair specificity. **Figure 5** shows the overall performance of screening for children under and above 4yo, respectively. Visual inspection revealed screening under 4yo achieved good-to-poor sensitivity and specificity, while screening above 4yo achieved good-to-poor sensitivity and good-to-fair specificity. **Figure 6** shows the performance of the screening tools evaluating predictive validity. These screening tools achieved fair-to-poor sensitivity and specificity.

Meta-Regression Investigating Effects of Screening Proxy, Test Administrator, Screening-Diagnosis Interval, and Age of Screening

The effects of screening proxy, test administrator, screening-diagnosis interval and age of screening on screening accuracy were investigated using bivariate meta-regression. **Table 7** summarizes the results. Screening tools with <6-month screening-diagnosis interval (i.e., concurrent validity) were associated with higher sensitivity when compared to those with longer than a 6-month interval (i.e., predictive validity). Tools using language ability as the proxy showed a marginally significantly higher sensitivity than those based on clinical markers. Screening tools based on language ability and those based on both language ability and clinical markers appeared to show a similar degree of sensitivity. For tools assessing concurrent validity, screening under the age of 4 had a higher sensitivity with marginal statistical significance but showed similar specificity with screening above the 4yo. As for tools assessing predictive validity, screening under and above 4yo appeared to show similar sensitivity and specificity. Similarly, screening tools relying on parent report and those conducted by trained examiners appeared to show a similar sensitivity. Despite the large variability in specificity, none of the factors in the meta-regression model explained this variability.

Results of sensitivity analysis after excluding studies with high ROB are illustrated in **Table 8**. The observed higher sensitivity for screening tools using actual language as proxy compared with those using clinical markers became statistically significant. The difference in sensitivity between screening tools assessing concurrent validity and those assessing predictive validity appeared to be larger than before the removal of the high ROB studies. However, the observed marginal difference between screening under and above 4yo became non-significant after the exclusion of high-risk studies. Similar to the results without excluding studies with high ROB, none of the included factors in sensitivity analysis explained variation in specificity.

DISCUSSION

The present review shows that currently available screening tools for language disorders during preschool years varies widely in their design and screening performance. Large variability in screening accuracy across different tools was a major issue

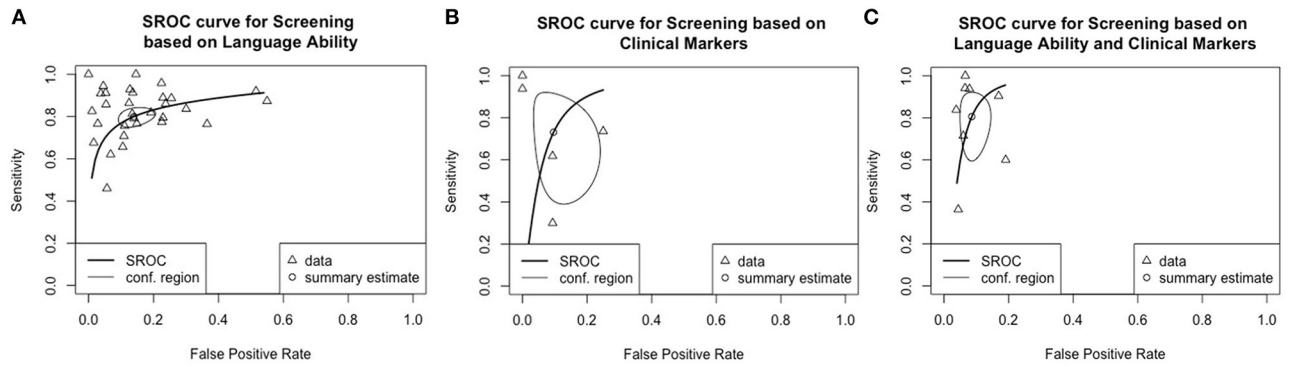

FIGURE 3 | Summary receiver operating characteristics curves for screening tools based on **(A)**, language ability, **(B)** clinical markers, and **(C)** language & clinical markers.

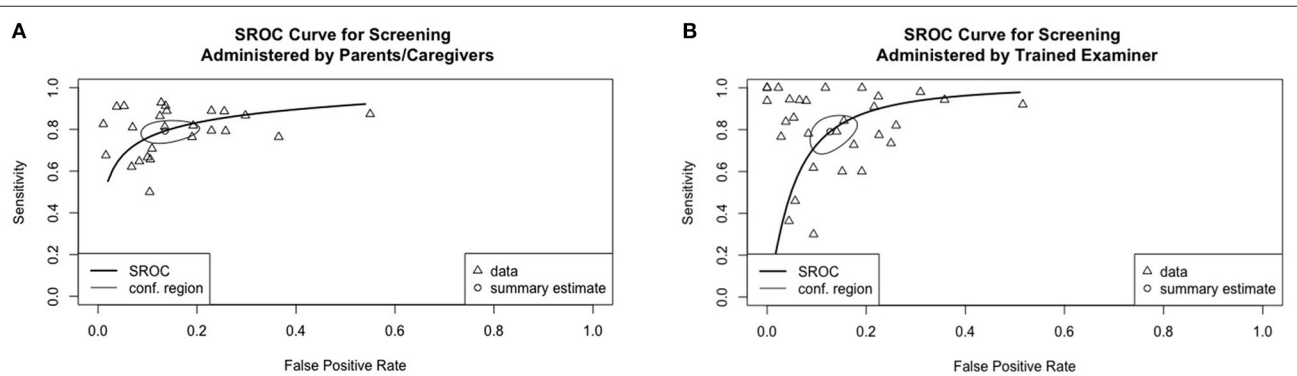

FIGURE 4 | Summary receiver operating characteristics curves for screening tools administered by **(A)** parents/caregivers and **(B)** trained examiners.

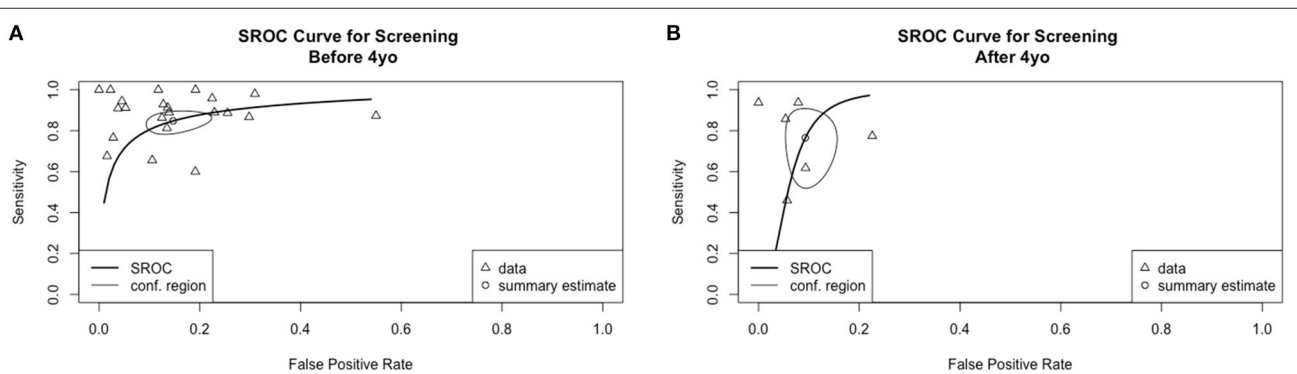

FIGURE 5 | Summary receiver operating characteristics curves for screening **(A)** under 4-year-old and **(B)** above 4-year-old.

in screening for language disorder. The present review also revealed that the variations arose from the choices of proxy and screening-diagnosis interval.

Screening tools based on children's actual language ability were shown to have higher sensitivity than tools based on clinical markers. The fact that screening tools based on clinical markers did not prove to be sensitive may be related to the mixed findings from primary studies. Notably, one of the primary studies using non-word repetition and sentence repetition tasks showed perfect accuracy in classifying all children with and without language disorder (110). The findings, however, could not be replicated in another study, using exactly the same test, which identified only 3 of the 10 children with language disorder (104). The difference highlighted the large variability in the performance of non-word and sentence repetition even among children with language disorders, in addition to the inconsistent difference found between children with and without language disorder (149). Another plausible explanation for the relatively

higher sensitivity of using child's actual language skills lies in the resemblance between the items used for screening based on the child's actual language and the diagnostic tests used as the reference standard. Differences in task design and test item selection across studies may have further increased the inconsistencies (149). Therefore, in future tool development or refinement, great care should be taken in the choice of screening proxy. More systematic studies directly comparing how different proxies and factors affect screening accuracy are warranted.

There was no evidence that other factors related to tool design, such as the test administrators of the screening tools, explained variability in accuracy. In line with a previous review (13), parent-report screening appeared to perform similarly to screening administered by trained examiners. This seemingly comparable accuracy supports parent-report instruments as a viable tool for screening, in addition to their apparent advantage of lower cost of administration. Primary studies directly comparing both types of screening in the same population may provide stronger evidence concerning the choice of administrators.

As predicted, long term prediction was harder to achieve than estimating concurrent status. Meta-analysis revealed that screening tools reporting predictive validity showed a significantly lower sensitivity than that of tools reporting concurrent validity, which was also speculated in the previous review (13). One possible explanation lies in the diverse developmental trajectories of language development in the preschool years. Some of the children who perform poorly in early screening may recover spontaneously at a later time point, while some who appeared to be on the right track at the beginning may develop language difficulties later on (7). Current screening tools might not be able to capture this dynamic change in language development in the preschool years, resulting in lower predictive validity than expected. Hence, language disorder screening should concentrate on identifying or introducing new proxies or metrics that are sensitive to the dynamic nature of language development. Vocabulary *growth* estimates, for example, might be more sensitive to long-term outcomes than a single point estimation (150). Although the current review has shown that different proxies has been used in screening language disorder, there is a limited number of studies examining how proxies other than children's actual language ability perform in terms of predictive validity. It would be useful to investigate the interaction between the proxy used and the screening-diagnosis interval in future studies.

Age of screening was expected to be affected by the varying developmental trajectories. Screening at an earlier age might have lower accuracy than screening at a later age when language development becomes more stable. This expected difference was not found in the current meta-analysis. However, it is worth noting that screening tools used at different ages not only differed in the age of screening, but also other domains. In the meta-analysis, over half (55%, 16/29) of the screening under 4 relied on parent reports and used tools such as vocabulary checklists and reported utterances while none of the screening above 4 (0/8) were based on parent reports. Inquiry about the effect of screening age on screening accuracy is crucial as it has direct implication on the optimal time of screening. Future studies that compare the screening accuracy at different ages with the method of assessment being kept constant (e.g., using the same screening tool) may reveal a clearer picture.

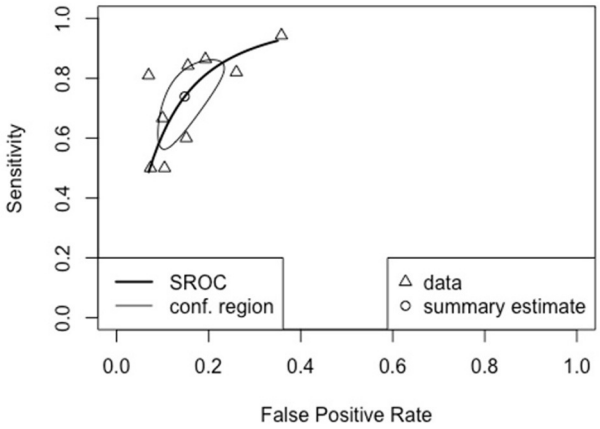

FIGURE 6 | Summary receiver operating characteristic curve for screening tools reporting predictive validity.

Overall, only a small proportion of all the available screening tools achieved good accuracy in identifying both children with and without language disorder. Yet, there is still insufficient evidence to recommend any screening tool, especially given the presence of ROB in some studies. Besides, the limited number of valid tools may explain partly why screening for language disorder has not yet been adopted as a routine surveillance exercise in primary care, in that the use of any one type of screening tools may result in a considerable amount of over-identification and missing cases, which can lead to long term social consequences (19). As shown in the current review, in the future development of screening tools, the screening proxy should be carefully chosen in order to maximize test sensitivity. However, as tools that have good accuracy are limited, there remains room for discussion on whether future test development should aim at maximizing sensitivity even at the expense of specificity. The cost of over-identifying a false-positive child for a more in-depth assessment might be less than that of under-identifying a true-positive child and depriving the child of further follow-ups (104). If this is the case, the cut-off for test positivity can be adjusted. The more stringent the criteria used in screening, the higher the sensitivity the test yield but with the trade-off of a decrease in specificity. However, the decision should be made by fully acknowledging the harms and benefits, which has not been addressed in the current review. While an increase in sensitivity by adjusting the cut-off might lead to the benefit of better follow-ups, the accompanying increase in false positive rate might lead to the harms of stigmatization and unnecessary procedures. Given the highly variable developmental trajectories in asymptomatic children, another direction for future studies could be to evaluate

TABLE 7 | Bivariate meta-regression on studies-related factors on sensitivity and false-positive rate.

Factor	Transformed sensitivity				Transformed false positive rate			
	Coeff.	95% CI		*p*-value	Coeff.	95% CI		*p*-value
		LL	UL			LL	UL	
Types (L vs. Cm)	0.657	−0.055	1.370	0.070#	0.325	−0.774	1.423	0.562
Types (L vs. Mx)	−0.300	−0.855	0.255	0.290	0.435	−0.330	1.201	0.265
Types (Mx vs. Cm)	0.885	−0.244	2.015	0.124	−0.094	−0.958	0.770	0.832
Time (P vs. C)	−0.528	−1.018	−0.037	0.035*	−0.016	−0.726	0.695	0.965
Sc. AgeC (<4yo vs. ≥ 4*yo*)	1.676	−0.115	1.467	0.094#	0.560	−0.292	1.412	0.198
Sc. AgeP (<4yo vs. ≥ 4*yo*)	1.061	−1.115	3.238	0.339	0.663	−0.737	2.064	0.353
Informant (TP∧ vs. Pa)	−0.003	−0.525	0.519	0.992	−0.031	−0.836	0.773	0.939

First group in the bracket as the reference; L, language only; Cm, clinical markers; Mx, both language and clinical markers; P, predictive validity; C, concurrent validity; Pa, parent; TE, trained personnel; ScAgeC, Screening Age (for studies evaluating concurrent validity); ScAgeC, Screening Age (for studies evaluating predictive validity).
*#p < 0.1; *p < 0.05.*

TABLE 8 | Bivariate meta-regression of study-related factors on sensitivity and false-positive rate excluding high ROB studies.

Factor	Transformed sensitivity				Transformed false positive rate			
	Coeff.	95% CI		*p*-value	Coeff.	95% CI		*p*-value
		LL	UL			LL	UL	
Types (L∧ vs. Cm)	0.960	0.291	1.629	0.005**	−0.020	−1.295	1.256	0.976
Types (L∧ vs. Mx)	−0.173	−0.784	0.439	0.580	0.157	−0.753	1.067	0.735
Types (Mx∧ vs. Cm)[a]	-	-	-	-	-	-	-	-
Time (P∧ vs. C)	−0.819	−1.377	−0.262	0.004*	−0.104	−1.009	0.801	0.822
Sc. Age C(<4yo vs. ≥ 4*yo*)	0.234	−0.926	1.394	0.692	0.520	−0.388	1.428	0.262
Sc. Age P(<4yo vs. ≥ 4*yo*)[a]	-	-	-	-	-	-	-	-
Informant (TE∧ vs. Pa)	0.149	−0.514	0.812	0.660	0.160	−0.870	1.189	0.761

First group in the bracket as the reference; L, language only; Cm, clinical markers; Mx, both language and clinical markers; P, predictive validity; C, concurrent validity; Pa, parent; TE, trained examiner; ScAgeC, Screening Age (for studies evaluating concurrent validity); ScAgeC, Screening Age (for studies evaluating predictive validity).
[a]Too few studies after exclusion for a valid analysis.
*#p < 0.1. *p < 0.05.*

the viability of targeted screening in a higher-risk population and compare it with universal screening.

This is the first study to use meta-analytical techniques specifically to evaluate the heterogeneity in screening accuracy of tools for identifying children with language disorder. Nonetheless, there were several limitations of the study. One limitation was related to the variability and validity of the gold standard in that the reference standard tests. Different countries or regions use different localized standardized or non-standardized tools and criteria to define language disorder. There is no one consensual or true gold standard. More importantly, the significance and sensitivity and specificity of the procedures used to identify children with language disorders in those reference tests were not examined. Some reference tests may employ arbitrary cut-offs (e.g.,−1.25 SD) to define language disorders while some researchers advocate children's well-being as the outcome, such that when children's lives are negatively impacted by their language skills, they are considered as having language disorders (151). This lack of consensus might further explain the diverse results or lack of agreement in replication studies.

Another limitation of the study was that nearly all the included studies had at least some ROB. This was mainly due to many unreported aspects in the studies. It is suggested that future validation studies on screening tools should follow reporting guidelines such as STARD (152). A third limitation was that the rating of ROB only involved one rater, and more raters may minimize potential bias. Lastly, not all included screening tools were analyzed in the meta-analysis. Some studies evaluated multiple screening tools at a number of cut-offs or times of assessment. Only one data point per study was included in the meta-analysis and the data used in meta-analysis were chosen based on Youden's index. This selection would inevitably inflate the accuracy shown in the meta-analysis. With the emergence of new methods for meta-analysis for diagnostic studies, more sophisticated methods for handling this complexity of data structure may be employed in future reviews.

This review shows that current screening tools for developmental language disorder vary largely in accuracy, with only some achieving good accuracy. Meta-analytical data identified some sources for heterogeneity. Future development

of screening tools should aim at improving overall screening accuracy by carefully choosing the proxy or designing items for screening. More importantly, metrics that are more sensitive to persistent language disorder should be sought. To fully inform surveillance for early language development, future research in the field can also consider broader aspects, such as the harms and benefits of screening as there is still a dearth of evidence in this respect.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Reference lists of the article.

AUTHOR CONTRIBUTIONS

KS and CT conceived and designed the study, wrote the paper, conducted the format and tables, reviewed, and edited the manuscript. KS performed the statistical analysis. All authors have approved the final manuscript for submission.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fped.2022.801220/full#supplementary-material

REFERENCES

1. Bishop DV, Snowling MJ, Thompson PA, Greenhalgh T, Consortium C, Adams C, et al. Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *J Child Psychol Psychiatry.* (2017) 58:1068–80. doi: 10.1111/jcpp.12721
2. Norbury CF, Gooch D, Wray C, Baird G, Charman T, Simonoff E, et al. The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. *J Child Psychol Psychiatry.* (2016) 57:1247–57. doi: 10.1111/jcpp.12573
3. American Speech-Language-Hearing Association. *Preschool Language Disorders*. Available online at: https://www.asha.org/public/speech/disorders/preschool-languagedisorders/
4. Beitchman JH, Wilson B, Brownlie E, Walters H, Inglis A, Lancee W. Long-term consistency in speech/language profiles: II. Behavioral, emotional, and social outcomes. *J Am Acad Child Adolesc Psychiatry.* (1996) 35:815–25. doi: 10.1097/00004583-199606000-00022
5. Brownlie E, Bao, Bao L, Beitchman J. Childhood language disorder and social anxiety in early adulthood. *J Abnormal Child Psychol.* (2016) 44:1061–70. doi: 10.1007/s10802-015-0097-5
6. Beitchman JH, Wilson B, Brownlie EB, Walters H, Lancee W. Long-term consistency in speech/language profiles: I. Developmental and academic outcomes. *J Am Acad Child Adolesc Psychiatry.* (1996) 35:804–14. doi: 10.1097/00004583-199606000-00021
7. Zambrana IM, Pons F, Eadie P, Ystrom E. Trajectories of language delay from age 3 to 5: Persistence, recovery and late onset. *Int J Lang Commun Disord.* (2014) 49:304–16. doi: 10.1111/1460-6984.12073
8. Armstrong R, Scott JG, Whitehouse AJ, Copland DA, Mcmahon KL, Arnott W. Late talkers and later language outcomes: Predicting the different language trajectories. *Int J Speech-Lang Pathol.* (2017) 19:237–50. doi: 10.1080/17549507.2017.1296191
9. Bishop DV, Snowling MJ, Thompson PA, Greenhalgh T, Consortium C. CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLoS ONE.* (2016) 11:e0158753. doi: 10.1371/journal.pone.0158753
10. Sim F, Thompson L, Marryat L, Ramparsad N, Wilson P. Predictive validity of preschool screening tools for language and behavioural difficulties: A PRISMA systematic review. *PLoS ONE.* (2019) 14:e0211409. doi: 10.1371/journal.pone.0211409
11. Siu AL. Screening for speech and language delay and disorders in children aged 5 years or younger: US Preventive Services Task Force recommendation statement. *Pediatrics.* (2015) 136:e474–81. doi: 10.1542/peds.2015-1711
12. Wallace IF. *Universal Screening of Young Children for Developmental Disorders: Unpacking the Controversies*. Occasional Paper. RTI Press Publication OP-0048-1802. RTI International (2018). doi: 10.3768/rtipress.2018.op.0048.1802
13. Wallace IF, Berkman ND, Watson LR, Coyne-Beasley T, Wood CT, Cullen K, et al. Screening for speech and language delay in children 5 years old and younger: a systematic review. *Pediatrics.* (2015) 136:e448–62. doi: 10.1542/peds.2014-3889
14. Guiberson M. Telehealth measures screening for developmental language disorders in Spanish-speaking toddlers. *Telemed E-Health.* (2016) 22:739–45. doi: 10.1089/tmj.2015.0247
15. Puglisi ML, Blasi HF, Snowling MJ. Screening for the identification of oral language difficulties in Brazilian preschoolers: a validation study. *Lang Speech Hear Serv Schools.* (2020) 51:852–65. doi: 10.1044/2020_LSHSS-19-00083
16. Bornstein MH, Hahn CS, Putnick DL, Suwalsky JT. Stability of core language skill from early childhood to adolescence: A latent variable approach. *Child Dev.* (2014) 85:1346–56. doi: 10.1111/cdev.12192
17. Tomblin JB, Zhang X, Buckwalter P, O'Brien M. The stability of primary language disorder. *J Speech Lang Hear Res.* (2003) 46:1283–96. doi: 10.1044/1092-4388(2003/100)
18. Salameh J-P, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *BMJ.* (2020) 370:m2632. doi: 10.1136/bmj.m2632
19. Plante E, Vance R. Selection of preschool language tests: A data-based approach. *Lang Speech Hear Serv Schools.* (1994) 25:15–24. doi: 10.1044/0161-1461.2501.15
20. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Internal Med.* (2011) 155:529–36. doi: 10.7326/0003-4819-155-8-201110180-00009
21. McGuinness LA, Higgins JP. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. *Res Synthesis Methods.* (2021) 12:55–61. doi: 10.1002/jrsm.1411
22. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol.* (2005) 58:982–90. doi: 10.1016/j.jclinepi.2005.02.022
23. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. London: The Cochrane Collaboration (2010).
24. Doebler P, Holling H. *mada: Meta-Analysis of Diagnostic Accuracy*. R package version 0.5.10. (2020). Available online at: https://CRAN.R-project.org/package=mada
25. Allen DV, Bliss LS. Concurrent validity of two language screening tests. *J Commun Disord.* (1987) 20:305–17. doi: 10.1016/0021-9924(87)90012-8
26. Lee LL. *Northwestern Syntax Screening Test (NSST)*. Press Evanston, IL: Northwestern University Press Evanston (1971).
27. Hedrick DL, Prather EM, Tobin AR. *Sequenced Inventory of Communication Development*. Washington, DC: University of Washington Press Seattle (1984).
28. Blaxley L, et al. Two language screening tests compared with developmental sentence scoring. *Lang Speech Hear Serv Schools.* (1983) 14:38–46. doi: 10.1044/0161-1461.1401.38

29. Bankson NW. *Bankson Language Screening Test*. Baltimore, MD: University Park Press (1977).
30. Lee LL. *Developmental Sentence Analysis: A Grammatical Assessment Procedure for Speech and Language Clinicians*. Evanston, IL: Northwestern University Press (1974).
31. Burden V, Stott CM, Forge J, Goodyer I, Burden V, Stott CM, et al. The Cambridge Language and Speech Project (CLASP). I. Detection of language difficulties at 36 to 39 months. *Dev Med Child Neurol.* (1996) 38:613–31. doi: 10.1111/j.1469-8749.1996.tb12126.x
32. Alpern G, Boll T, Shearer M. Developmental profile II. *J Read.* (1980) 18:287–91.
33. Renfrew CE. Persistence of the open syllable in defective articulation. *J Speech Hear Disord.* (1966) 31:370–3. doi: 10.1044/jshd.3104.370
34. Renfrew C. *The Bus Story: A Test of Continuous Speech*. Oxford: England (1969).
35. Carscadden J, Corsiatto P, Ericson L, Illchuk R, Esopenko C, Sterner E, et al. A pilot study to evaluate a new early screening instrument for speech and language delays. *Canad J Speech-Lang Pathol Audiol.* (2010) 34:87–95.
36. Bzoch KR, League R, Brown V. *Receptive-Expressive Emergent Language Test*. Austin, TX: Pro-Ed (2003).
37. Chaffee CA, Cunningham CE, Secord-Gilbert M, Elbard H, Richards J. Screening effectiveness of the Minnesota child development inventory expressive and receptive language scales: sensitivity, specificity, and predictive value. *Psychol Assessment.* (1990) 2:80–5. doi: 10.1037/1040-3590.2.1.80
38. Reynell J. *Reynell Developmental Language Scales, Revised (Windsor: NFER)*. Windsor: NFER Publishing Company (1977).
39. Ireton HR, Thwing EJ. *Minnesota Child Development Inventory*. Minneapolis, MN: Behavior Science Systems (1972).
40. Dias DC, Rondon-Melo S, Molini-Avejonas DR. Sensitivity and specificity of a low-cost screening protocol for identifying children at risk for language disorders. *Clinics.* (2020) 75:e1426. doi: 10.6061/clinics/2020/e1426
41. American Speech-Language and Hearing Association. How does your child hear and talk? (2006). Available online at: http://www.asha.org/public/speech/development/chart/
42. Andrade CRF, Befi-Lopes DM, Fernandes FDM, Wertzner HF. *Teste de Language Infantil nas Áreas de Fonologia, Vocabulário, Fluência e Pragmática*. 2nd ed. Barueri: Pró-Fono (2011).
43. Dixon J, Kot A, Law J. Early language screening in City and Hackney: work in progress. *Child Care Health Dev.* (1988) 14:213–29. doi: 10.1111/j.1365-2214.1988.tb00576.x
44. Reynell J, Huntley M. *Reynell Developmental Language Scales, 2nd revision*. Windsor: NFER-Nelson (1985).
45. Lowe M, Costello A. *The Symbolic Play Test*. Windsor: NFER-Nelson (1976).
46. Gray S, Plante E, Vance R, Henrichsen M. The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Lang Speech Hear Serv Schools.* (1999) 30:196–206. doi: 10.1044/0161-1461.3002.196
47. Gardner MF. *EOWPVT-R: Expressive One-Word Picture Vocabulary Test, Revised*. Novato, CA: Academic Therapy Publications (1990).
48. Dunn LM, Dunn LM. *Peabody Picture Vocabulary Test: PPVT-IIIB*. Circle Pines, MN: American Guidance Service Circle Pines (1997). doi: 10.1037/t15145-000
49. Gardner MF. *Receptive One-Word Picture Vocabulary Test*. Novato, CA: Academic Therapy Publications (1985).
50. Williams K. *Expressive Vocabulary Test (EVT)*. Circle Pines, MN: American Guidance Service Inc. (1997).
51. Zimmerman IL, Steiner VG. *Preschool Language Scale 4th edition, Spanish*. Antonio, TX: The Psychological Corporation (2004).
52. Guiberson M, Rodriguez BL. Measurement properties and classification accuracy of two Spanish parent surveys of language development for preschool-age children. *Am J Speech-Lang Pathol.* (2010) 19:225–37. doi: 10.1044/1058-0360(2010/09-0058)
53. Guiberson M. Concurrent validity of a parent survey measuring communication skills of Spanish speaking preschoolers with and without delayed language. *Perspect Commun Disord Sci.* (2008) 15:73–81. doi: 10.1044/cds15.3.73
54. Squires J, Potter l, Bricker D. *The ASQ User's Guide (2nd ed.)*. Baltimore, MD: Paul H. Brookes (1999).
55. Guiberson M, Rodríguez BL, Dale PS. Classification accuracy of brief parent report measures of language development in Spanish-speaking toddlers. *Lang Speech Hear Serv Sch.* (2011) 42:536–49. doi: 10.1044/0161-1461(2011/10-0076)
56. Squires J, Bricker D, Mounts L, Potter L, Nickel R, Twombly E, et al. *Ages and Stages Questionnaire*. Baltimore, MD: Paul H. Baltimore, MD: Brookes (1999).
57. Jackson-Maldonado D, Thal DJ, Fenson L. MacArthur *Inventarios del Desarrollo de Habilidades Comunicativas: User's Guide and Technical Manual*. Baltimore, MD: Brookes Pub (2003).
58. Guiberson M, Rodriguez BL, Zajacova A. Accuracy of telehealth-administered measures to screen language in Spanish-speaking preschoolers. *Telemed E-Health.* (2015) 21:714–20. doi: 10.1089/tmj.2014.0190
59. Heilmann J, Weismer SE, Evans J, Hollar C. Utility of the MacArthur-bates communicative development inventory in identifying language abilities of late-talking and typically developing toddlers. *Am J Speech-Lang Pathol.* (2005) 14:40–51. doi: 10.1044/1058-0360(2005/006)
60. Board CA. *The MacArthur-Bates Communicative Development Inventory: Words and Sentences*. Baltimore, MD: Paul H Brookes (1992).
61. Zimmerman I, Steiner V, Pond R. *The Preschool Language Scale-3*. Columbus: Merrill (1992).
62. Klee T, Carson DK, Gavin WJ, Hall L, Kent A, Reece S. Concurrent and predictive validity of an early language screening program. *J Speech Lang Hear Res.* (1998) 41:627–41. doi: 10.1044/jslhr.4103.627
63. Rescorla L. The language development survey: A screening tool for delayed language in toddlers. *J Speech Hear Disord.* (1989) 54:587–99. doi: 10.1044/jshd.5404.587
64. Mullen EM. *Infant MSEL Manual: Infant Mullen Scales of Early Learning*. Cranston, RI: TOTAL Child (1989).
65. Klee T, Pearce K, Carson DK. Improving the positive predictive value of screening for developmental language disorder. *J Speech Lang Hear Res.* (2000) 43:821–33. doi: 10.1044/jslhr.4304.821
66. Laing GJ, Law J, Levin A, Logan S. Evaluation of a structured test and a parent led method for screening for speech and language problems: prospective population based study. *BMJ.* (2002) 325:1152. doi: 10.1136/bmj.325.7373.1152
67. Edwards S, Fletcher P, Garman M, Hughes A, Letts C, Sinka I. *The Reynell Developmental Language Scales III, NFER*. NELSON Publishing, The University of Reading Edition (1997).
68. Law J. Early language screening in City and Hackney: The concurrent validity of a measure designed for use with 2½-year-olds. *Child.* (1994) 20:295–308. doi: 10.1111/j.1365-2214.1994.tb00392.x
69. Levett L, Muir J. Which three year olds need speech therapy? Uses of the Levett-Muir language screening test. *Health Visitor.* (1983) 56:454–6.
70. Reynell J, Reynell Huntley M. *Reynell Developmental Language Scales (Revised)*. Windsor: NFER Publishing Company Ltd. (1977).
71. Goldman R. *Goldman-Fristoe Test of Articulation*. Circle Pines, MI: American Guidance Service. Inc. (1969).
72. Crystal D, Fletcher P, Garman M. *The Grammatical Analysis of Language Disability: A Procedure for Assessment and Remediation*. Hodder Education (1976).
73. Visser-Bochane MI, van der Schans CP, Krijnen WP, Reijneveld SA, Luinge MR. Validation of the early language scale. *Eur J Pediatrics.* (2020) 180:63–71. doi: 10.1007/s00431-020-03702-8
74. Schlichting JEPT, Spelberg L. *Lexilijst Begrip [Lexilist Comprehension]*. Amsterdam: Harcourt Test Publishers (2007).
75. Schlichting L, Spelberg HL. *Schlichting test voor taalbegrip [Schlichting test for language comprehension]*. Houten: Bohn Stafleu van Loghum (2010).
76. Schlichting JEPT. Lexilijst Nederlands. (2002). doi: 10.1240/sav_gbm_2002_h_000262
77. Schlichting JEPT, Spelberg HC. *Schlichting Test voor Taalproductie-II: voor Nederland en Vlaanderen*. Houten: Bohn Stafleu van Loghum (2012). doi: 10.1007/978-90-313-9842-3
78. Slofstra-Bremer C, van der Meulen S, Lutje Spelberg H. *De Taalstandaard [The Language standard]*. Amsterdam: Pearson (2006).

79. Geurts H. *Handleiding CCC-2-NL*. Amsterdam: Harcourt Test Publishers (2004).
80. Visser-Bochane M, Luinge M, Dieleman L, van der Schans C, Reijneveld S. The Dutch well child language screening protocol for 2-year-old children was valid for detecting current and later language problems. *Acta Paediatrica.* (2020) 110:2825–32. doi: 10.1111/apa.15447
81. Mattsson CM, Mårild S, Pehrsson NG. Evaluation of a language-screening programme for 2.5-year-olds at Child Health Centres in Sweden. *Acta Paediatr.* (2001) 90:339–44. doi: 10.1080/080352501300067776
82. McGinty C. An investigation into aspects of the Mayo early language screening test. *Child Care Health Dev.* (2000) 26:111–28. doi: 10.1046/j.1365-2214.2000.00176.x
83. Garvey-Cecchetti B, Heslin C, Laundon O, McGinty C, O'Malley L, Dowd P, et al. *The Mayo Early Language Screening Test*. Western Health Board: Mayo Speech and Language Therapy Department (1993).
84. Anthony A, Bogle D, Ingram T, McIsaac M. *1971: Edinburgh Articulation Test*. Edinburgh: Churchill Livingstone (1971).
85. Nair MKC, Harikumaran Nair GS, Mini AO, Indulekha S, Letha S, Russell PS. Development and validation of language evaluation scale Trivandrum for children aged 0-3 years - LEST (0-3). *Indian Pediatrics.* (2013) 50:463–7. doi: 10.1007/s13312-013-0154-5
86. Bzoch K, League R. *Receptive-Expressive Emergent Language Scale*. Gainesville, FL: Tree of Life Press. Inc. (1971).
87. Nayeb L, Lagerbere D, Westerlund M, Sarkadi A, Lucas S, Eriksson M. Modifying a language screening tool for three-year-old children identified severe language disorders six months earlier. *Acta Paediatrica.* (2019) 108:1642–8. doi: 10.1111/apa.14790
88. Andrade Cd, Befi-Lopes DM, Fernandes FDM, Wertzner HF. *ABFW: teste de linguagem infantil nas áreas de fonologia, vocabulário, fluência e pragmática*. São Paulo: Pró-Fono. (2004).
89. Bischop D. *The Test for Reception of Grammar, Version 2 (TROG-2)*. Oxford: Pearson (2009).
90. Santos FH, Bueno OFA. Validation of the Brazilian Children's Test of Pseudoword Repetition in Portuguese speakers aged 4 to 10 years. *Brazil J Med Biol Res.* (2003) 36:1533–47. doi: 10.1590/S0100-879X2003001100012
91. Rescorla L, Alley A. Validation of the language development survey (LDS): A parent report tool for identifying language delay in toddlers. *J Speech Lang Hear Res.* (2001) 44:434–45. doi: 10.1044/1092-4388(2001/035)
92. Sachse S, Von Suchodoletz W. Early identification of language delay by direct language assessment or parent report? *J Dev Behav Pediatrics.* (2008) 29:34–41. doi: 10.1097/DBP.0b013e318146902a
93. Grimm H, Doil H. *Elternfragebögen für die Früherkennung von Risikokindern*. ELFRA: Hogrefe, Verlag für Psychologie (2000).
94. Grimm H. *Sprachentwicklungstest für sweijährige slindes-SETK-2 und für dreibis fünfjährige Kinder-SETK 3-5*. Göttingen: Hogrefe (2000).
95. Stokes SF. Secondary prevention of paediatric language disability: a comparison of parents and nurses as screening agents. *Eur J Disord Commun.* (1997) 32:139–58. doi: 10.1111/j.1460-6984.1997.tb01628.x
96. van Agt HM, van der Stege HA, de Ridder-Sluiter JG, de Koning HJ. Detecting language problems: accuracy of five language screening instruments in preschool children. *Dev Med Child Neurol.* (2007) 49:117–22. doi: 10.1111/j.1469-8749.2007.00117.x
97. Stott CM, Merricks MJ, Bolton PF, Goodyer IM. Screening for speech and language disorders: The reliability, validity and accuracy of the General Language Screen. *Int J Lang Commun Disord.* (2002) 37:133–51. doi: 10.1080/13682820110116785
98. Brouwers-de Jong E, Burgmeijer R, Laurent de Angulo M. Ontwikkelingsonderzoek op het consultatiebureau: handboek bij het vernieuwde Van Wiechenonderzoek. (1996).
99. Walker D, Gugenheim S, Downs MP, Northern JL. Early Language Milestone Scale and language screening of young children. *Pediatrics.* (1989) 83:284–8. doi: 10.1542/peds.83.2.284
100. Coplan J. *ELM Scale: The Early Language Milestone Scale*. Pro-Ed (1983).
101. Wetherby AM, Goldstein H, Cleary J, Allen L, Kublin K. Early identification of children with communication disorders: Concurrent and predictive validity of the CSBS Developmental Profile. *Infants Young Child.* (2003) 16:161–74. doi: 10.1097/00001163-200304000-00008
102. Wetherby AM, Prizant BM. *Communication and Symbolic Behavior Scales: Developmental Profile*. Baltimore, MD: Paul H Brookes Publishing Co. (2002). doi: 10.1037/t11529-000
103. Kapalkova S, Polisenska K, Vicenova Z. Non-word repetition performance in Slovak-speaking children with and without SLI: novel scoring methods. *Int J Lang Commun Disord.* (2013) 48:78–89. doi: 10.1111/j.1460-6984.2012.00189.x
104. Nash H, Leavett R, Childs H. Evaluating the GAPS test as a screener for language impairment in young children. *Int J Lang Commun Disord.* (2011) 46:675–85. doi: 10.1111/j.1460-6984.2011.00038.x
105. Van der Lely H, Gardner H, McClelland AGR, Froud KE. *Grammar and Phonology Screening Test: (GAPS)* London: DLDCN (2007).
106. Wiig E, Secord W, Semel E. *Clinical Evaluation of Language Fundamentals-Preschool*. Second *UK edition*. London: Harcourt Assessment (2006).
107. Sturner RA, Funk SG, Green JA. Preschool speech and language screening: further validation of the sentence repetition screening test. *J Dev Behav Pediatr.* (1996) 17:405–13. doi: 10.1097/00004703-199612000-00006
108. Sturner R, Kunze L, Funk S, Green J. Elicited imitation: its effectiveness for speech and language screening. *Dev Med Child Neurol.* (1993) 35:715–26. doi: 10.1111/j.1469-8749.1993.tb11717.x
109. Kirk SA, Kirk WD, McCarthy JJ. *Illinois Test of Psycholinguistic Abilities*. Champaign, IL: University of Illinois press (1968).
110. van der Lely HKJ, Payne E, McClelland A. An investigation to validate the grammar and phonology screening (GAPS) test to identify children with specific language impairment. *PLoS ONE.* (2011) 6:e022432. doi: 10.1371/journal.pone.0022432
111. Fluharty NB. The design and standardization of a speech and language screening test for use with preschool children. *J Speech Hear Disord.* (1974) 39:75–88. doi: 10.1044/jshd.3901.75
112. Hedrick DL, Prather EM, Tobin AR. *Sequenced Inventory of Communication Development*. Seattle, WA: University of Washington Press (1975).
113. Benavides AA, Kapantzoglou M, Murata C. Two grammatical tasks for screening language abilities in Spanish-speaking children. *Am J Speech-Lang Pathol.* (2018) 27:690–705. doi: 10.1044/2017_AJSLP-17-0052
114. Zimmerman IL, Steiner VG, Pond RE. *Preschool Language Scale-Fifth Edition Spanish Screening Test (PLS-5 Spanish Screening Test). [Measurement assessment]*. San Antonio, TX: Pearson (2011) doi: 10.1037/t15141-000
115. Fluharty NB. *Fluharty Preschool Speech and Language Screening Test: Teaching Resources*. Austin, TX: Pro-Ed (1978).
116. Lee LL, Koenigsknecht RA, Mulhern ST. *Developmental Sentence Scoring*. Evanston, IL: North Western University (1974).
117. Bliss LS, Allen DV. *Screening Kit of Language Development*. Baltimore, MD: University Park Press (1983).
118. Bliss LS, Allen DV. Screening kit of language development: A preschool language screening instrument. *J Commun Disord.* (1984) 17:133–41. doi: 10.1016/0021-9924(84)90019-4
119. Lavesson A, Lovden M, Hansson K. Development of a language screening instrument for Swedish 4-year-olds. *Int J Lang Commun Disord.* (2018) 53:605–14. doi: 10.1111/1460-6984.12374
120. Matov J, Mensah F, Cook F, Reilly S. Investigation of the language tasks to include in a short-language measure for children in the early school years. *Int J Lang Commun Disord.* (2018) 53:735–47. doi: 10.1111/1460-6984.12378
121. Matov J, Mensah F, Cook F, Reilly S, Dowell R. The development and validation of the Short Language Measure (SLaM): A brief measure of general language ability for children in their first year at school. *Int J Lang Commun Disord.* (2020) 55:345–58. doi: 10.1111/1460-6984.12522
122. Semel E, Wiig EH, Secord W. *Clinical Evaluation of Language Fundamentals-Fourth Edition, Australian Standardised Edition*. Sydney, NSW: PsychCorp (2006).
123. Wright R, Levin B. A preschool articulation and language screening for the identification of speech disorders. *Final Rep.* (1971).
124. Eisenberg SL, Guo L-Y. Differentiating children with and without language impairment based on grammaticality. *Lang Speech Hear Serv Schools.* (2013) 44:20–31. doi: 10.1044/0161-1461(2012/11-0089)

125. Dawson JI, Stout C, Eyer J, Tattersall PJ, Fonkalsrud J, Croley K. *SPELT-P 2: Structured Photographic Expressive Language Test*. Preschool: Janelle Publications (2005).
126. Eisenberg SL, Guo L, Germezi M. How grammatical are three-year-olds? *Lang Speech Hear Serv Schools.* (2012) 43:36–52. doi: 10.1044/0161-1461(2011/10-0093)
127. Bruce B, Kornfalt R, Radeborg K, Hansson K, Nettelbladt U. Identifying children at risk for language impairment: screening of communication at 18 months. *Acta Paediatrica.* (2003) 92:1090–5. doi: 10.1080/08035250310004414
128. Holmberg E, Nelli SB. *Neurolingvistisk undersökningsmodell fr språkstörda barn.* Utbildningsproduktion AB (kommer inom kort att ges ut i ny, något omarbetad upplaga av Pedagogisk Design) (1986).
129. Bishop D. *TROG svensk manual [svensk översättning och bearbetning: Eva Holmberg och Eva Lundälv].* Göteborg: SIH Läromedel (Originalarbete publicerat 1983). (1998).
130. Frisk V, Montgomery L, Boychyn E, Young R, Vanryn E, McLachlan D, et al. Why screening canadian preschoolers for language delays is more difficult than it should be. *Infants Young Child.* (2009) 22:290–308. doi: 10.1097/IYC.0b013e3181bc4db6
131. Harrison PL. *AGS Early Screening Profiles*. Circle Pines, MN: American Guidance Service (1990).
132. Zimmerman I, Steiner V, Pond R. *Preschool Language Scale-Fourth Edition (PLS-4).* San Antonio, TX: The Psychological Corporation (2002) doi: 10.1037/t15140-000
133. Newborg J, Stock J, Wnek L, Guidabaldi J, Svinicki J. *Battelle Developmental Inventory*. Itasca: Riverside (1988).
134. Bracken BA. *Bracken Basic Concept Scale-Revised*. San Antonio, TX: Psychological Corporation (1998).
135. Gascoe FP. *Technical Report for Brigance Screens*. Cheltenham, VIC: Hawker Brownlow Education (1998).
136. Jessup B, Ward E, Cahill L, Keating D. Teacher identification of speech and language impairment in kindergarten students using the Kindergarten Development Check. *Int J Speech-Lang Pathol.* (2008) 10:449–59. doi: 10.1080/17549500802056151
137. Office for Educational Review. *Revised Kindergarten Development Check*. Hobart: Office for Educational Review (2003).
138. Pesco D, O'Neill DK. Predicting later language outcomes from the language use inventory. *J Speech Lang Hear Res.* (2012) 55:421–34. doi: 10.1044/1092-4388(2011/10-0273)
139. O'Neill D. *Language Use Inventory: An Assessment of Young Children's Pragmatic Language Development for 18-to 47-Month-Old Children*. Waterloo, ON: Knowledge in Development (2009).
140. Seymour HN, Roeper T, De Villiers JG, De Villiers PA. *Diagnostic Evaluation of Language Variation, Norm Referenced*. Pearson: DELV (2005).
141. Wiig E, Secord W, Semel E. *Clinical Evaluation of Language Fundamentals (Preschool 2nd edition ed.)*. San Antonio, TX: The Psychological Corporation: Harcourt Assessment Inc. (2004).
142. Bishop DV. *CCC-2: Children's Communication Checklist-2*. Pearson (2006).
143. Westerlund M, Berglund E, Eriksson M. Can severely language delayed 3-year-olds be identified at 18 months? Evaluation of a screening version of the MacArthur-Bates communicative development inventories. *J Speech Lang Hear Res.* (2006) 49:237–47. doi: 10.1044/1092-4388(2006/020)
144. Berglund E, Eriksson M. Communicative development in Swedish children 16-28 months old: The Swedish early communicative development inventory-words and sentences. *Scand J Psychol.* (2000) 41:133–44. doi: 10.1111/1467-9450.00181
145. Eriksson M, Berglund E. Swedish early communicative development inventories: Words and gestures. First *Lang.* (1999) 19:55–90. doi: 10.1177/014272379901905503
146. Westerlund M, Sundelin C. Can severe language disability be identified in three-year-olds? Evaluation of a routine screening procedure. *Acta Paediatrica.* (2000) 89:94–100. doi: 10.1111/j.1651-2227.2000.tb01195.x
147. Westerlund M, Sundelin C. Screening for developmental language disability in 3-year-old children. Experiences from a field study in a Swedish municipality. *Child.* (2000) 26:91–110. doi: 10.1046/j.1365-2214.2000.00171.x
148. Mullen EM. *Mullen Scales of Early Learning*. Circle Pines, MN: AGS (1995).
149. Ahufinger N, Berglund-Barraza A, Cruz-Santos A, Ferinu L, Andreu L, Sanz-Torrent M, et al. Consistency of a nonword repetition task to discriminate children with and without developmental language disorder in Catalan-Spanish and European Portuguese speaking children. *Children.* (2021) 8:85. doi: 10.3390/children8020085
150. Rowe ML, Raudenbush SW, Goldin-Meadow S. The pace of vocabulary growth helps predict later vocabulary skill. *Child Dev.* (2012) 83:508–25. doi: 10.1111/j.1467-8624.2011.01710.x
151. Nippold MA, Tomblin JB. *Understanding Individual Differences in Language Development Across the School Years*. New York, NY: Psychology Press (2014). doi: 10.4324/9781315796987
152. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clin Chem.* (2015) 61:1446–52. doi: 10.1373/clinchem.2015.246280

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.